Université de Montréal

# Extraction d'information à partir de transcriptions de conversations téléphoniques spécialisées

par

Narjès Boufaden

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Ph.D.
en informatique

Décembre 2004

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

# Extraction d'information à partir de transcriptions de conversations téléphoniques spécialisées

présentée par

Narjès Boufaden

a été évaluée par un jury composé des personnes suivantes :

Michel Boyer
président-rapporteur

Guy Lapalme
directeur de recherche

Yoshua Bengio
co-directeur de recherche

Douglas Eck
membre du jury

Benoît Habert
examinateur externe

Thèse acceptée le :

. . . . . .

*À mes parents et mes enfants*

# Résumé

Le but de l'extraction d'information (EI) est de structurer des informations pertinentes à un domaine particulier. Il s'agit de repérer des instances d'une classe d'événements ou de relations et d'en extraire ses attributs qui sont des faits individuels.

Les approches standard d'EI reposent essentiellement sur l'analyse syntaxique du contexte, c'est-à-dire la phrase, pour définir une relation sujet-verbe-objet qui établit les rôles thématiques attribués à l'information pertinente, tandis que l'analyse sémantique de l'information pertinente est restreinte à l'étiquetage des noms propres faisant référence à des personnes, des organisations et des lieux. L'application de ces approches à des textes conversationnels spécialisés se heurte à deux problèmes liés à la rhétorique et à la spécialisation de ces types de textes.

La fragmentation de l'information et les irrégularités langagières de l'oral telles que les répétitions et omissions sont des exemples de difficultés en EI à partir des textes conversationnels. Les paires question-réponse modifient la structure de l'information et le caractère spontané des conversations diminue la densité d'information dans un énoncé et augmente leur nombre. L'information est communiquée sur plusieurs énoncés ancrés au thème par le biais d'anaphores pronominales. De plus, les irrégularités langagières de l'oral modifient la

structure syntaxique de l'énoncé en insérant des éléments superflus ou en supprimant un ou plusieurs éléments constituant ainsi un obstacle pour l'analyse syntaxique complète d'un énoncé. Par ailleurs, les textes spécialisés se caractérisent par un vocabulaire composé de mots et termes spécialisés définissant les concepts pertinents du domaine et qui sont souvent différents des entités nommées. L'extraction de ces informations nécessite une étape d'étiquetage sémantique qui va au-delà de l'extraction des entités nommées.

Dans cette thèse, nous montrons que les approches standard d'extraction d'information sont inappropriées pour les textes conversationnels spécialisés. Nous proposons une nouvelle approche d'EI qui repose davantage sur le contenu des mots que sur leur contexte, pour repérer les informations pertinentes.

Le coeur de notre approche est une étape d'étiquetage sémantique robuste des informations pertinentes au domaine avec des concepts issus d'une ontologie du domaine pour l'apprentissage des relations prédicat-arguments pertinentes au domaine. En plus de l'étiquetage sémantique, notre approche est composée de quatre autres étapes. La segmentation linguistique définit des unités linguistiques adéquates pour l'apprentissage des relations prédicat-arguments. La segmentation thématique découpe les conversations en unités thématiquement cohérentes facilitant l'étape de résolution des anaphores pronominales. Cette dernière fait émerger une partie des relations pertinentes masquées par la pronominalisation du thème. L'étape d'apprentissage des patrons d'extraction utilise des modèles de Markov intégrant des états *joker* et des transitions nulles pour gérer les bruits introduits par les irrégularités langagières de l'oral.

Notre approche testée sur des transcriptions manuelles de conversations téléphoniques dans le domaine de la recherche et sauvetage maritime a permis d'extraire les faits individuels avec un F-score de 78,9 %.

Mots clé : extraction d'information, analyse de conversations, étiquetage sémantique, apprentissage statistique, segmentation de textes, traitement automatique du langage naturel

# Abstract

Information extraction (IE) is about seeking instances of event classes and relations and extracting their arguments from text within a particular domain. Standard approaches rely heavily on syntactic processing to determine the thematic roles of relevant information, while little semantic processing is done and is restricted to named entity extraction. Applying standard approaches to specialized conversational texts faces two problems that have to do with the text's rhetoric structure and with specialization.

Scattered information and disfluencies such as repetitions and omissions are examples of difficulties in IE from conversational texts. Question-answer pairs, widely used in conversations, are examples where bits of information are conveyed through successive utterances. In addition, the spontaneous character of composition decreases information density conveyed in utterances while increasing their number. Information is often conveyed through several utterances by the means of pronominal anaphora.

Edited words, omissions and interruptions are examples of disfluencies that alter the utterance structure causing a significant decrease of performance in part-of-speech tagging and parsing. Furthermore, altering the syntactic structure of utterances makes syntactic-driven learning of extraction patterns difficult if not impossible.

On the other hand, specialized texts are characterized by a sub-language including a

specialized vocabulary referring to relevant domain concepts that require a semantic tagging process that goes beyond named entity extraction.

In this thesis, we show that standard approaches are unsuitable for specialized conversational texts. We propose a new IE approach that takes into account the characteristics of specialized conversational texts. Our main claim is that a word-meaning based approach is more suitable than a context based approach for these texts. The core component of our approach is a robust semantic tagger based on a statistical model that labels relevant bits of information with concepts drawn from a domain ontology. Word labels are used to learn predicate-arguments relations relevant to the domain.

In addition to the robust semantic tagger, our five step approach includes a linguistic segmentation stage that defines units suitable for the pattern learning process. Topic segmentation identifies topically coherent units that help anaphora resolution. The latter helps to identify more relevant relations hidden by the pronominalization of the topic. This stage precedes the pattern learning stage, which is based on Markov models that include *wild card* states designed to handle edited words and null transitions to handle omissions.

We tested our approach on manually transcribed telephone conversations in the domain of maritime search and rescue, and succeeded in extracting individual facts with an F-score of 78.9%.

Key words : information extraction, conversation analysis, semantic parsing, statistical learning, text segmentation, natural language processing

# Remerciements

Je tiens à remercier mon directeur de thèse Guy Lapalme pour m'avoir accueilli au laboratoire de Recherche Appliquée en Linguistique Informatique (RALI) et pour avoir dirigé ma thèse. L'enthousiasme et la ténacité dont il a fait preuve, ainsi que la confiance et la liberté qu'il m'a accordées au cours de cette thèse, m'ont permis d'entreprendre de nombreuses expériences et ont grandement contribué à la richesse de ce travail. Également, je lui suis reconnaissante de son humanité qui m'a permis de concilier vie de mère et d'étudiante.

Je souhaite aussi remercier Yoshua Bengio membre du Laboratoire Informatique de Systèmes Adaptatifs (LISA) pour avoir co-dirigé ce travail de thèse. Sa disponibilité, sa patience et son expertise m'ont permis d'étendre mes connaissances à l'apprentissage automatique.

Je tiens également à remercier Benoît Habert, professeur à l'Université Paris X-Nanterre effectuant sa recherche au LIMSI (CNRS), Douglas Eck et Michel Boyer, professeurs à l'Université de Montréal, pour l'intérêt qu'ils ont bien voulu porter à ce travail et pour avoir accepté de faire partie du jury.

Je remercie particulièrement Graham Russell, chercheur au RALI, pour son implication dans mon travail, pour ses suggestions pertinentes tout au long de mon parcours de doctorat et pour avoir accepté de lire cette thèse.

J'exprime ma sincère reconnaissance à mon collègue Luc Plamondon, avec qui j'ai eu le

plaisir de partager le bureau, pour son soutien technique et moral et pour son aide tout au long de la rédaction de cette thèse et ce jusqu'à la toute fin.

Je souhaite remercier Elliot Macklovitch, responsable du laboratoire RALI, pour sa disponibilité, ses conseils et pour avoir rendu mon passage au RALI des plus enrichissants sur le plan intellectuel et social.

J'aimerais aussi remercier Leila Kosseim, professeure à l'Université Concordia, pour son aide au début de ma thèse, Philippe Langlais, professeur à l'Université de Montréal et George Foster, actuellement chercheur à l'Institut de Technologie Langagière, pour leur disponibilité aux moments où j'en avais besoin et pour les outils informatique qu'ils ont bien voulu me fournir. Je n'oublie pas les gens du support technique pour leurs dépannages rapides et efficaces et je pense en particulier à Jean-Louis, Bernard et Mohammed.

Je souhaite aussi remercier Robert Parks de l'organisation Wordsmyth pour m'avoir permis d'utiliser une version électronique de Wordsmyth. Également, je remercie le Centre Recherche et Développement pour la Défense Canada à Valcartier pour avoir rendu disponibles les transcriptions des conversations utilisées pour cette thèse, le Secrétariat National Recherche et Sauvetage pour les manuels de Recherche et Sauvetage aisni que Luc Lamontagne, actuellement professeur à l'Université Laval, pour son initiation au jargon du domaine.

Aussi, je tiens à remercier mes amis, Leila Arras, Simona Gandrabur et Horacio Saggion, avec qui j'ai passé des moments très agréables et qui m'ont offert leur soutien moral quand j'en avais le plus besoin.

Je remercie mon mari Mourad pour sa patience, sa compréhension et son amour et bien sûr mes deux bouts de chou Nawel et Emir pour tous les merveilleux moments courts mais intenses qu'ils m'ont permis de passer avec eux.

Enfin, à mes très chers parents envers qui je serai éternellement reconnaissante, je vous dis merci pour votre soutien continuel et inconditionnel. Cette thèse est le fruit de vos encouragements.

# TABLE DES MATIÈRES

# LISTE DES TABLEAUX

# TABLE DES FIGURES

# Chapitre 1

# Introduction

## 1.1 Introduction

Converser est un acte spontané grâce auquel nous communiquons, échangeons nos idées et socialisons. Au-delà de la socialisation, c'est un outil efficace en milieu de travail comme en témoigne l'expansion des centres d'appels.

La téléphonie est devenue une industrie à part entière qui touche différents secteurs de l'économie. De nos jours, les transactions bancaires, les réservations de billets d'avion et la vente de produits se font par téléphone. Parallèlement, motivées par le désir de compétitivité, de plus en plus d'entreprises prennent conscience de l'importance du rôle des technologies de l'information dans l'amélioration des services et l'accroissement de la rentabilité. D'ores et déjà, des centres d'appels enregistrent les conversations téléphoniques dans le but d'être analysées pour le contrôle de la qualité.

Par ailleurs, la mise en marché de systèmes de reconnaissance de la parole de plus en plus performants fait de cette conjoncture un indice révélateur du besoin d'outils pour le traitement automatique des transcriptions de conversations téléphoniques.

Notre travail s'inscrit dans le cadre innovateur de la compréhension automatique des

transcriptions manuelles de conversations téléphoniques (que nous appelons textes conversationnels) dans le but d'extraire des informations utiles pour le peuplement d'une base de données.

La compréhension automatique du langage naturel est un des enjeux principaux des applications du traitement automatique des langues (TAL). Cette tâche difficile et ambitieuse a donné lieu à différents systèmes de TAL moins exigeants, en termes de représentation du contenu, mais plus accessibles en termes de réalisation. Parmi ceux-ci, les systèmes de génération automatique de résumés[1] et d'extraction d'information (EI) sont des technologies relativement maîtrisées du TAL [Chinchor et Dungca, 1995].

Dans cette thèse, nous abordons la problématique de la compréhension des conversations téléphoniques d'un point de vue de l'EI. Cela consiste à fournir une représentation structurée des informations pertinentes contenues dans un texte conversationnel pour un domaine d'application donné.

Depuis les années 80, différents types de textes ont été étudiés en EI. Ils peuvent être classés selon deux critères : (1) la rhétorique (texte structuré[2] et non bruité[3], par opposition à non structuré et bruité) et (2) la spécialisation (domaine général, par opposition à domaine spécialisé). Sur les vingt années de travaux dans ce domaine, nous avons observé un engouement pour les textes écrits non spécialisés de type dépêche journalistique, notamment à cause des campagnes d'évaluation Message Understanding Conferences (MUC). Plus récemment, depuis les années 2000, l'intérêt pour l'EI s'est porté vers des textes plus spécialisés[4]. Plusieurs *workshops* ont été organisés dans ce domaine et les travaux en sont à l'expérimentation de ces applications. Un élément prépondérant pour ces deux domaines d'intérêt en EI a été la disponibilité des corpora pour ces types de textes : le corpus British

[1] `http://duc.nist.gov/`

[2] La notion de texte structuré en extraction d'information fait généralement référence à des textes struturés de manière automatique avec un format XML, par exemple [Soderland, 1998]. Un exemple de ces textes sont les rapports météorologiques sur internet. Dans le cadre de cette thèse nous associons la notion de structuration à celle de grammaticalité.

[3] Les textes bruités présentent des irrégularités langagières de l'oral qui sont les répétitions, les ommissions et les reprises. Dans la section 2.2.1, nous donnons des exemples de ces irrégularités.

[4] Les textes spécialisés sont des textes qui sont rédigés dans un sous-langage défini par un vocabulaire spécialisé, des relations sémantiques et des constructions syntaxiques particulières. Les rapports météorologiques, les manuels de réparation d'avions et les articles scientifiques en pharmacologie en sont des exemples.

National Corpus (BNC)[5] pour les dépêches journalistiques et Medline[6] pour les textes en biomédecine.

À la différence des travaux en EI à partir de textes écrits, peu de recherches ont porté sur les textes conversationnels, notamment à cause du manque de corpora de transcriptions de la parole. Des travaux ont été consacrés à l'EI à partir de transcriptions de bulletins du journal télévisé dans le cadre des conférences HUB [Robinson *et al.*, 1999], toutefois, ceux-ci portaient sur la parole préparée non conversationnelle (par opposition à la parole spontannée conversationelle), et uniquement sur la tâche d'extraction des entités nommées.

Cependant, depuis 1997, des campagnes d'évaluation de systèmes de reconnaissance de la parole sont organisées annuellement par le National Institute of Standards and Technology (NIST) et depuis 2001, les évaluations[7] portent sur la reconnaissance de la parole spontanée conversationnelle téléphonique. De facto, l'intérêt porté à cette problématique a donné lieu à des corpora de transcriptions de conversations téléphoniques (Switchboard Credit Card), Verbmobil[8]) assurant ainsi l'expansion des outils de TAL pour ces textes.

Dans le cadre de cette thèse, nous proposons une approche d'EI pour des transcriptions manuelles de la parole spontannée conversationnelle. Nous la validons en implémentant trois étapes des quatre proposées, c'est à dire la segmentation des conversations, leur étiquetage sémantique et l'apprentissage des patrons d'extraction pour extraire les faits individuels[9]. La dernière étape correspond à la résolution des coréférences et n'est pas développée dans le cadre de cette thèse dans la mesure où elle représente une problématique à part entière commune à différentes applications du TAL telles que le résumé automatique.

Les trois premières étapes ont été validées sur des transcriptions manuelles de conversations téléphoniques dans le domaine de la recherche et sauvetage maritime. Ces textes ont

[5] Ce corpus est composé de textes littéraires, d'articles de journaux ainsi que de transcriptions de réunions ou conversations. Ce corpus est distribué par le Linguistic Data Consortium (LDC), `http://www.ldc.upenn.edu/`

[6] `http://www.ncbi.nlm.nih.gov/`

[7] `http://www.nist.gov/speech/publications/index.htm`

[8] Evaluations and Language ressources Dsitribution Agency (ELDA), `http://www.elda.org/`

[9] Un fait individuel ("individual fact")[Grishman, 1998] est une information pertinente extraite à partir du texte et obtenue sans recours à un mécanisme d'inférence (résolution de coréférences ou inférence à partir des connaissances du monde ou du domaine).

été fournis par le Centre de Recherche de la Défense Valcartier (CRDV) dans le cadre du projet SARPlan[10] qui, à l'origine, a été proposé pour la conception d'un outil d'aide à la décision.

## 1.2 Problématique

Le but de l'extraction d'information (EI) est de structurer des informations pertinentes à un domaine particulier. Il s'agit de repérer des instances d'une classe d'événements ou de relations et d'en extraire ses attributs. Les informations factuelles sont d'abord extraites, pour ensuite servir à inférer des faits plus complexes par le biais d'inférences ou par la résolution des coréférences.

Au début des années 90, les MUC ont joué un rôle important dans la formalisation du processus d'EI dans les textes écrits non spécialisés tels que les dépêches journalistiques. L'approche standard développée pour ces types de textes repose sur deux principes :

1. L'information pertinente est véhiculée par des groupes nominaux qui se rapportent essentiellement aux noms de personnes, d'organisations et de lieux.
2. La phrase est une unité linguistique syntaxico-sémantique maximale contenant une relation **sujet-verbe-objet**[11] qui garantit la localité des constituants syntaxiques véhiculant l'information pertinente.

Dans un référentiel qui classe les textes selon leur degré de spécialisation et leur rhétorique, les textes étudiés lors des MUC se retrouvent parmi les moins spécialisés et les plus structurés. Il s'ensuit que l'extension de l'approche d'EI à types de textes plus spécialisés et moins structurés tels que les textes conversationnels spécialisés nécessite l'étude des particularités de ces textes afin de vérifier que les deux principes sur lesquels repose l'approche sont applicables. Ainsi, nous abordons cette problématique en répondant aux questions suivantes :

1. Quelles sont les caractéristiques linguistiques des textes conversationnels spécialisés qui violent les principes de base sur lesquels repose l'approche standard ?

[10] `http://www.drev.dnd.ca/f/actualitesdisplay_f.asp?lang=f&page=33&news=61`
[11] La majorité des travaux en extraction d'information s'intéresse aux contructions **sujet-verbe-objet**.

2. Quelles sont les étapes du processus d'EI à reconsidérer ou à ajouter ?

Nous proposons une approche d'EI adaptée aux textes conversationnels spécialisés et nous validons cette approche par l'implémentation d'une méthode d'apprentissage de patrons d'extraction, étape centrale dans le développement d'un système d'EI.

### 1.2.1 Exemple de texte conversationnel spécialisé

Les textes considérés dans notre étude sont des transcriptions de conversations téléphoniques du domaine de la recherche et sauvetage (*Search And Rescue* ou SAR). Ces conversations ont été transcrites manuellement et fournies par le CRDV. Elles peuvent être : (1) des rapports d'incidents survenus en mer tels qu'un bateau porté disparu ou perdu en mer, (2) des discussions pour planifier des missions de recherche ou de sauvetage, comme l'allocation d'avions et de bateaux pour la recherche, ou (3) des comptes-rendus d'une mission de sauvetage et les résultats de cette mission. Il peut aussi s'agir d'une combinaison de ces trois cas.

Le tableau 1.1 est un exemple de texte conversationnel. Nous l'appellerons Overdue boat car nous nous y réfererons tout au long de cet ouvrage. Pour faciliter les références à ses énoncés, nous avons ajouté un numéro pour chaque tour de parole et un identificateur du locuteur.

Les expressions soulignées dans la conversation sont les informations pertinentes au domaine de la recherche et sauvetage. Parmi celles-ci, nous distinguons des noms de personnes (`Mr. Wellington`), d'organisations (`Maritime operation centre`) et de lieux (`the South Coast of Newfoundland`), ainsi que le type d'incident (`overdue boat`), le nom d'avions et bateaux alloués pour la recherche (`DFO`[12] `King Air, Challenger, Hurk`) et le bateau objet de l'incident (`Doray`).

[12] Ce terme est un acronyme de `Department of Fisheries and Oceans.`

| No | Loc. | Énoncé |
|---|---|---|
| 1 | a : | Maritime operation centre, (INAUDIBLE) hello. |
| 2 | b : | hi, Mr. Wellington, it's captain Mr. VanHorn |
| 3 | a : | yes. |
| 4 | b : | ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the South Coast of Newfoundland, just in the area quite between Fortune Bay and Trepassey. |
| 5 | b : | it's on the south east coast of Newfoundland. |
| 6 | b : | this is been going on for, for 24 hours that the case has, or almost anyway, and we had an DFO King Air up flying this morning |
| 7 | b : | they did a radar search for us in that area. |
| 8 | a : | yes. |
| 9 | b : | and their search turned up nothing. |
| 10 | a : | yeah. |
| 14 | b : | so I'm wondering about the possibility of attempting it with a different platform perhaps someone with even other sensors other than the radar and, in fact, someone with a, with a radar that'll be a little more sensitive. |
| 15 | b : | before I used the Challenger, I 'll use a Hurk. |
| 20 | a : | ...do you want this thing fired up now or you wanna wait till the Big Boys come in to work tomorrow morning ? |
| 21 | b : | well, I would like it if possible, I'd like them to, to be airborne at first light. |
| 22 | a : | ok. |
| 23 | a : | ok then ( INAUDIBLE ). |
| 28 | a : | i think that's about all the information I need to know so, it's a... |
| 29 | b : | it's... |
| 30 | a : | an overdue vessel ? |
| 31 | b : | overdue 20-feet open boat, a Doray, with a 10-horse power upboard, one person on board. |
| 32 | a : | ok. |
| 33 | a : | so, it's up to the South East Coast ? |
| 34 | b : | right. |
| 35 | b : | and it started 18 Zulu on 8, so like 24 hours ago. |
| 36 | a : | ok. |
| 37 | b : | thanks. |

TAB. 1.1 – Exemple d'une conversation transcrite manuellement. Le thème général de cette conversation est la signalisation d'un bateau porté disparu mer (Overdue boat) et le déclenchement d'une mission de recherche. Les mots soulignés sont les informations pertinentes à ce domaine que nous voulons identifier automatiquement.

| No | Loc | Énoncé |
|---|---|---|
| 1 | a : | Maritime operation centre, (INAUDIBLE) hello. |
| | | ORGANISATION |
| 2 | b : | hi, Mr. Wellington, it's captain Mr. VanHorn |
| | | PERSON PERSON |
| 3 | a : | yes. |
| 4 | b : | ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the South Coast of Newfoundland, LOCATION just in the area quite between Fortune Bay and Trepassey. LOCATION |
| 5 | b : | it's on the south east coast of Newfoundland. LOCATION |

TAB. 1.2 – Entités nommées dans un extrait de la conversation Overdue boat.

### 1.2.2 Extraction d'information à partir de textes conversationnels spécialisés

L'approche standard d'EI développée pour les textes journalistiques est un processus en cascade qui peut être divisé en trois étapes :

**L'extraction des entités nommées** C'est une étape d'étiquetage sémantique simplifiée pour trouver les noms d'organisations (ORGANISATION), de lieux (LOCATION), de personnes (PERSON) ainsi que certaines expressions numériques telles que les dates (DATE). Les expressions soulignées dans l'extrait de conversation du tableau 1.2 en sont des exemples. La classe sémantique de l'entité nommée est indiquée en dessous de chacune.

Dans les applications d'EI à partir de textes non spécialisés, les entités nommées représentent l'essentiel des entités à annoter sémantiquement.

**Extraction des faits individuels** Cela correspond à une étape de filtrage utilisant des patrons d'extraction générés manuellement ou automatiquement (section 1.2.5) qui sélectionnent les expressions pertinentes au domaine de l'application. Ces patrons permettent de déterminer les événements (le *quoi*) pertinents à un domaine particulier

ainsi que la relation entre ces entités ou leur rôle (*qui fait quoi, où, quand*) relativement à ces événements. Un exemple de patron d'extraction est présenté à l'exemple (1).

(1) `VESSEL is on LOCATION`

Le patron (1) exprime la relation entre le concept VESSEL qui est une entité pertinente au domaine et le concept LOCATION qui est le lieu où le bateau a été observé pour la dernière fois.

**La résolution des coréférences et l'inférence de faits complexes** La première met en relation les anaphores et leurs antécédents (entités ou événements). Dans l'exemple suivant, le pronom `it` fait référence à l'expression `overdue boat` :

`it`'s on the south east coast of Newfoundland
LOCATION

⇓

`overdue boat` is on the south east coast of Newfoundland
VESSEL LOCATION

La seconde infère des réponses non explicitées dans le texte à partir des faits individuels.

Les informations obtenues à partir des deux dernières étapes sont utilisées pour générer des formulaires qui sont des représentations structurées des événements les plus pertinents à un domaine particulier. Des exemples de formulaires[13] remplis à partir de la conversation Overdue boat sont donnés au tableau 1.3. Les informations contenues dans les formulaires sont extraites manuellement et sont sélectionnées parmi les groupes de mots soulignés dans la conversation Overdue boat.

### 1.2.3 Analyse des réponses de champs de formulaires

Le tableau 1.3 illustre un ensemble de formulaires fournis par le CRDV et remplis à partir de la conversation Overdue boat. Les formulaires :

[13]Un formulaire est une représentation tabulaire de l'information.

| *Missing-Object* | |
|---|---|
| 1 ID | *OBJECT1* |
| 2 Category | boat |
| 3 Type | 20-feet Doray |
| 4 Engine-type | 10 horse-power |
| | upboard |
| 5 Persons-on-board | 1 |

| *Search-Unit1* | |
|---|---|
| 1 Id | *UNIT2* |
| 2 Brand | Hercules |
| 3 Organization | DFO |
| 4 Detection-equipment | radar |
| 5 Status | *CONSIDERED* |
| 6 Result | |
| 7 Date | tomorrow morning, first light |

| *Incident* | |
|---|---|
| 1 Controller | Captain Mr. Van Horn |
| 2 Type | overdue |
| 3 Location | South East Coast of Newfoundland, between Fortune Bay and Trepassey |
| 4 Date | 18 Zulu n 8, 24 hours ago |
| 5 Missing-object | *OBJECT1* |
| 6 Search-missions | *MISSION2* |

| *Search-Mission* | |
|---|---|
| 1 Id | *MISSION2* |
| 2 Search-unit | *UNIT1, UNIT2* |
| 3 Region | South East Coast of NewFoundland |
| 4 Type | radar search |
| 5 Status | being planned |

| *Search-Unit2* | |
|---|---|
| 1 Id | *UNIT1* |
| 2 Brand | King Air |
| 3 Organization | DFO |
| 4 Detection-equipment | radar |
| 5 Status | |
| 6 Result | nothing |
| 7 Date | this morning |

Tab. 1.3 – Formulaires remplis manuellement à partir de la conversation Overdue boat.

***Missing-object*** regroupe la catégorie de l'objet recherché (CATEGORY), le type de l'objet (TYPE), ses caractéristiques techniques (ENGINE-TYPE) et le nombre de personnes à bord (PERSONS-ON-BOARD).

***Search-Unit1*** **et** ***Search-Unit2*** rassemblant les informations sur le véhicule utilisé pour la recherche (BRAND), l'équipement utilisé pour la recherche (DETECTION-EQUIPEMENT), la disponibilité du véhicule (STATUS), le résultat de la recherche (RESULT) et la date de début de la recherche (DATE).

***Incident*** est un formulaire qui contient les informations sur la personne en charge de la mission (CONTROLLER), le type de l'incident (TYPE), la date à laquelle a été rapporté l'incident (DATE), l'endroit présumé de l'incident (LOCATION), un pointeur sur l'objet recherché (MISSING-OBJECT) et un pointeur sur les missions allouées pour la recherche du bateau (SEARCH-MISSIONS).

***Search-Mission*** réunit les informations relatives à la mission de recherche, c'est-à-dire la date du début de la mission (DATE), le lieu de la mission (REGION), le type de recherche (DETECTION), l'état d'avancement de la mission (STATUS) et un pointeur sur les ressources allouées pour cette mission (SEARCH-UNIT) et le formulaire *Search-Unit* rassemble les informations sur la ressource allouée.

Les réponses des champs des formulaires du tableau 1.3 sont de trois types. Premièrement, les réponses en majuscules italiques suivies d'un chiffre sont des pointeurs vers d'autres formulaires. Ainsi, la réponse *OBJECT1* du champ *Incident*:MISSING-OBJECT[14] renvoie au formulaire *Missing-object*. Les réponses en majuscules qui ne sont pas suivies d'un chiffre sont obtenues par inférence. Par exemple, la réponse *CONSIDERED* du champ *Search-Unit1*:STATUS ne figure pas explicitement dans la conversation Overdue boat, mais elle est obtenue par inférence à partir de l'énoncé `15 I'll use a Hurk`. Troisièmement, les réponses en `typewriter face` sont des faits individuels explicitement présents dans la conversation. Notre analyse des failles des approches standard porte sur ce type de réponses.

[14] Cette notation *Formulaire*:CHAMP renvoie au champ CHAMP du formulaire *Formulaire*.

Nous remarquons que certaines réponses des champs de formulaires sont obtenues en combinant les mots soulignés de différents énoncés de la conversation, comme c'est le cas pour les champs *Incident*:LOCATION et *Search-Unit1*:DATE.

D'autres réponses sont des mots ou des groupes de mots extraits d'énoncés qui ne présentent pas une structure syntaxique grammaticale : c'est le cas des champs *Incident*:DATE, *Missing-object*:TYPE et *Missing-object*:ENGINE-TYPE.

Enfin, nous remarquons que certaines réponses, par exemple celle du champ *Search-Unit1*:DATE, sont problématiques car elles ne correspondent pas à un terme du domaine. Ce sont des expressions sémantiquement similaires à des termes spécialisés répertoriés dans un lexique du domaine. Dans notre exemple, l'expression `first light` peut être considérée comme une expression sémantiquement similaire à l'expression `morning` habituellement utilisée dans ces textes. Cet exemple soulève un problème connu en extraction de termes qui se rapporte aux variantes terminologiques. Il résulte de l'utilisation de mots qui n'apparaissent pas dans le vocabulaire du domaine mais qui sont sémantiquement similaires à un terme de ce vocabulaire.

Ainsi, pour identifier les classes d'événements et relations pertinentes à un domaine, il est nécessaire de tenir compte des particularités suivantes des textes conversationnels :

**La représentation fragmentaire de l'information** L'information peut s'étaler sur plus d'un tour de parole moyennant des références pronominales. Le tableau 1.4 montre des exemples d'information fragmentaire. Dans le premier exemple (énoncés `4` à `6`), le lieu de l'incident est indiqué dans deux tours de parole différents, tandis que dans le deuxième exemple (énoncés `7` à `10`), le premier tour de parole indique qu'une recherche par radar a été effectuée et le résultat de cette recherche est introduit par une référence anaphorique dans un second tour de parole. Enfin, dans le troisième exemple (énoncés `20` à `22`), la fragmentation résulte de l'utilisation du style interrogatif.

Ainsi, la réponse `south east coast of Newfoundland, between Fortune Bay and Trepassey` contenue dans le champ *Incident*:LOCATION est obtenue en combinant le contenu des énoncés `4` et `5` car l'énoncé `5` est une reprise partielle avec correction de

| |
|---|
| 4 b : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the South Coast of Newfoundland, just in the area quite between Fortune Bay and Trepassey.<br>5 b : it's on the south east coast of Newfoundland.<br>6 b : this is been going on for, for 24 hours that the case has, or almost anyway, and we had an DFO King Air up flying this morning |
| 7 b : they did a radar search for us in that area.<br>8 a : yes.<br>9 b : and their search turned up nothing.<br>10 a : yeah. |
| 20 a : ...do you want this thing fired up now or you wanna wait till the Big Boys come in to work tomorrow morning ?<br>21 b : well, I would like it if possible, I'd like them to, to be airborne at first light.<br>22 a : ok. |

TAB. 1.4 – Illustration de la représentation fragmentaire de l'information. La présence d'anaphores pronominales en position de sujet et les paires de question-réponse sont des cas typiques de fragmentation de l'information.

l'énoncé `4`, tandis que pour le champ *Search-Unit2*:DATE, la réponse est construite de manière itérative lors des échanges `20` et `21`. L'énoncé `20 a` spécifie une date (`tomorrow morning`), tandis que l'énoncé `20 b` précise la période de la journée pour débuter la mission de recherche (`first light`).

Enfin, certaines réponses sont obtenues après résolution de la coréférence pronominale en position de sujet, comme c'est le cas pour le champ *Incident*:LOCATION où l'expression `south east coast of Newfoundland` est rattachée au lieu de l'incident grâce à la résolution du pronom `it` dans l'énoncé `5`. Nous montrons que la pronominalisation du thème[15] est un phénomène très présent lors du développement d'une unité théma-

[15]La notion de thème dans le cadre de cette thèse se rapporte au sujet d'une unité thématique. La section 4.3.1 définit clairement cette notion.

tique[16]. Aussi, nous montrons que la résolution de l'anaphore pronominale en position de sujet fait émerger des relations pertinentes masquées par la pronominalisation du thème. Par exemple, la relation VESSEL `is on` LOCATION décrite à l'étape d'extraction des faits individuels (section 1.2.2) n'apparaît qu'après la résolution de l'anaphore pronominale `it`. Ces exemples mettent en évidence la nécessité de définir une *unité d'extraction* tenant compte de l'aspect fragmentaire de l'information.

**Les irrégularités langagières de l'oral** Ce sont les répétitions, omissions et reprises de mots ou de partie d'un groupe de mots. Elles altèrent la structure syntaxique des énoncés et rendent difficile, sinon impossible, la génération d'une analyse syntaxique complète d'un énoncé. La présence de répétitions de mots ou de syntagmes ou les omissions invalident la présemption d'une position définiedes syntagmes dans la relation **sujet-verbe-objet** et rend difficile la déduction d'une relation **prédicat-argument** à partir de la strcuture syntaxique.

```
28 a : i think that's about all the information I need to know so, it's
       a...
29 b : it's...
30 a : an overdue vessel ?
31 b : overdue 20-feet open boat, a Doray, with a 10-horse power
       upboard, one person on board.
32 a : ok.
```

TAB. 1.5 – Illustration d'une interruption et d'omissions de composantes syntaxiques qui engendrent des relations syntaxiques incomplètes.

Ainsi, dans l'énoncé `31` du tableau 1.5, il n' y a pas de sujet, ni de verbe liant le bateau recherché et sa description. La relation **sujet-verbe-objet** cruciale pour identifier les rôles thématiques des objets est incomplète.

[16]Une unité thématique est un segment d'une conversation qui porte sur un même sujet qui se distingue de celui du segment précédent et de celui du suivant.

### 1.2.4 Failles des approches standard d'EI

Nous distinguons deux problèmes liés aux deux premières étapes de l'approche standard d'extraction appliquées aux textes conversationnels spécialisés :

1. Les problèmes rencontrés à l'étape d'extraction des entités nommées touchent la couverture des entités pertinentes au domaine. Les entités nommées sont essentiellement des groupes nominaux pertinents pour le domaine mais indépendants des événements recherchés[17]. Ce sont souvent des noms propres qui sont identifiés grâce à une liste ou un lexique. Elles correspondent à trois classes sémantiques qui renvoient à des noms de personnes (entité PERSONNE), d'organisations (entité ORGANISATION) et de lieux (entité LIEU). Dans les textes spécialisés, telle que la conversation Overdue boat, d'autres types d'expressions présentent les mêmes caractéristiques que les entités nommées. Par exemple, les noms d'avions ou les types d'équipement de détection sont des expressions pertinentes pour le domaine et indépendantes d'un événement en particulier. Ils représentent des termes spécialisés du domaine de la recherche et sauvetage. Il est clair que l'étape d'extraction des entités nommées est une tâche particulière aux textes de type dépêches journalistiques. Le passage à des textes spécialisés nécessite un autre type d'étiquetage sémantique. L'extraction des entités nommées devient alors davantage un processus d'extraction des termes du domaine.

2. Le deuxième problème concerne l'étape d'extraction des faits individuels. Le succès de cette étape repose sur deux éléments : la détection des entités pertinentes du domaine (étape d'extraction des entités nommées pour l'approche standard) et la couverture des relations pertinentes (étape de conception des patrons d'extraction). Nous avons déjà discuté du premier élément au point précédent. La couverture des relations pertinentes à partir de textes conversationnels spécialisés est une problématique à part entière car elle dépend de la spécialisation des textes et de leur rhétorique.

[17] Ces entités sont les principaux attributs des classes d'événements, mais à l'étape d'extraction des entités nommées aucune relation ne les lie aux événements. Ce sont les étapes d'extraction des faits individuels et de résolution des coréférences qui établissent les liens entre entités nommées et événements.

### 1.2.5 Apprentissage de patrons d'extraction à partir de textes conversationnels

Un patron d'extraction est une structure linguistique qui introduit des contraintes syntaxiques (position des composantes **sujet**, **verbe**, **objet** dans la relation **sujet-verbe-objet**) et sémantiques (classes de mots) permettant le filtrage d'un sous-ensemble d'énoncés qui contiennent les faits individuels pertinents au domaine d'application.

L'analyse des réponses des formulaires (section 1.2.3) montre que la conception de cette structure dépend de deux éléments :

**Le domaine d'application** Il définit le degré de spécialisation des textes. Plus un texte est spécialisé, plus il se caractérise par un vocabulaire contenant davantage de termes qui appartiennent à différentes classes sémantiques du domaine. Ces classes sont utilisées dans la description des patrons d'extraction. En biomédecine, par exemple, ce sont les noms de gènes et de protéines, tandis que dans les dépêches journalistiques ce sont les noms de personnes, d'organisations et de lieux. Le patron VESSEL `is on` LOCATION (section 1.2.2) est un exemple qui illustre l'utilisation des classes sémantiques dans la description de patrons d'extraction (classes VESSEL et LOCATION).

**La rhétorique** Elle a un impact sur les contraintes syntaxiques. La phrase, unité linguistique utilisée dans l'approche standard d'EI, est bien délimitée et est conforme à des règles de grammaire connues garantissant la présence de la relation **sujet-verbe-objet** et la position des constituants syntaxiques.

À l'opposé, les conversations sont des transcriptions de l'oral spontané qui ne présentent pas toujours une relation **sujet-verbe-objet**. Ce sont plutôt des tours de parole qui peuvent être entrecoupés et où l'on retrouve différentes irrégularités langagières de l'oral telles que les répétitions et les omissions de mots. En particulier, le bruit[18] introduit par ces irrégularités interfère dans la structure linguistique des patrons d'extraction et modifie la relation syntaxique entre ses éléments, soit :

[18] Ce sont les mots superflus ou au contraire l'absence de mots introduits par les irrégularités de l'oral.

- En insérant des éléments superflus (cas des répétitions et reprises), comme dans l'exemple suivant :

  (2) `sujet(them) répétition(to), verbe(to be airborne) objet(at first light)`

  Dans cet exemple, les faits individuels à extraire sont `them to be aiborne at first light`, mais la particule `to` interfère dans la disposition des constituants de la relation **sujet-verbe-objet**.

- En supprimant un ou plusieurs éléments (par exemple un groupe nominal `gn` ou un verbe `v`) de la relation syntaxique (cas des omissions), comme dans l'exemple (3) où les éléments entre crochets sont absents de l'énoncé et de la relation **sujet-verbe-objet** qui aurait permis de rattacher `24 hours ago` au champ *Incident*:DATE (tableau 1.3) :

  (3) `it started 18 Zulu on 8, so [sujet(it) verbe(started)] objet(like 24 hours ago)`

D'autre part, le cas du champ *Search-Unit1*:DATE, qui est une combinaison des informations présentes dans la question et sa réponse (section 2.3.2), montre l'importance de la prise en compte du contexte, c'est-à-dire la question, pour comprendre et situer la réponse et permettre la conception des patrons d'extraction. À ce niveau, nous retenons deux structures importantes :

- Les paires d'adjacence distinguant les tours de parole dépendant l'un de l'autre pour une représentation cohérente de l'information, comme dans le cas des paires question-réponse où l'information exprimée dans la deuxième partie complète celle exprimée dans la première partie.

- Les unités thématiques qui sont importantes pour traiter la pronominalisation du thème, phénomène omniprésent dans les conversations et réduisant la visibilité des relations pertinentes.

## 1.3 Proposition de thèse

Le but de ce travail est de proposer une approche d'EI pour les textes conversationnels spécialisés et de la valider en implémentant les trois premières étapes de notre approche, c'est-à dire la segmentation linguistique et thématique, l'étiquetage sémantique et l'apprentissage de patrons d'extraction (Figure 1.1). Cette étude s'articule autour de quatre thèmes :

1. **Analyse linguistique des textes conversationnels spécialisés** afin de déterminer les caractéristiques structurelles et linguistiques déterminantes pour la tâche d'extraction. Le but de cette étude est la spécification d'une unité linguistique adéquate pour la tâche d'extraction. Elle dresse également un inventaire des propriétés syntaxiques et sémantiques de l'information recherchée pour ce type de textes. Le résultat de cette étude est un ensemble de lignes directrices qui conditionnent notre approche d'EI pour ces textes.

2. **Segmentation des textes conversationnels en unités linguistiques et thématiques**. Nous proposons une méthodologie pour la détection des frontières d'unités linguistiques. La segmentation proposée est motivée par la recherche d'une relation **prédicat-arguments** pour répondre à des questions telles que "qui fait quoi ?". La segmentation en unités thématiques facilite la résolution des anaphores pronominales engendrée par la pronominalisation du thème.

3. **Étiquetage sémantique des textes conversationnels**. L'analyse des caractéristiques syntaxiques et sémantiques de l'information ciblée permet de définir les connaissances *a priori* nécessaires à la tâche d'extraction. Dans notre cas, nous modélisons ces connaissances dans une ontologie du domaine. Nous proposons une approche d'étiquetage sémantique robuste qui détecte les termes du domaine et leurs variantes.

4. **L'apprentissage de patrons**, à partir de séquences de mots annotés sémantiquement avec les concepts de l'ontologie. Nous proposons une approche d'apprentissage des patrons d'extraction basée sur les modèles de Markov [Rabiner, 1989; MacDonald et Zucchini, 1997] pour gérer le bruit introduit par les irrégularités langagières de l'oral.

La figure 1.1 décrit notre approche d'EI ainsi que les étapes que nous proposons de valider dans le cadre de cette thèse. Les rectangles en ligne continue sont les modules que nous développons, tandis que le rectangle en ligne pointillée représente l'étape de résolution des coréférences que nous ne développerons pas étant donné l'étendue de cette problématique.

## 1.4 Contributions de la thèse

Nous proposons une approche d'EI adaptée aux textes conversationnels spécialisés se basant sur trois notions clé :

**L'ontologie du domaine** qui organise les termes du domaine en classes sémantiques. L'utilisation d'une ontologie permet une meilleure généralisation des patrons car l'apprentissage se fait sur des classes sémantiques, donc des classes de mots, au lieu des termes.

**La segmentation des conversations** en unités linguistiques et thématiques. La première segmentation définit une unité linguistique pour l'apprentissage des patrons. Cette unité est associée aux paires d'adjacence qui respectent la dépendance des tours de parole et est donc un moyen de définir une unité sémantique maximale. Cette situation s'applique surtout aux paires question-réponse. La segmentation en unités thématiques facilite la résolution des anaphores pronominales en position de sujet et contribue à faire émerger les relations pertinentes dissimulées par la pronominalisation du thème. Le but de cette étape est de permettre une meilleure couverture des relations pertinentes au domaine.

**Les relations prédicat-arguments** qui définissent des contraintes telles que celles imposées par la relation **sujet-verbe-objet** pour l'identification des rôles thématiques. Toutefois, ces contraintes s'opèrent au niveau sémantique en déterminant les associations entre les classes sémantiques. Cela lève la contrainte de contiguïté des constituants syntaxiques imposée par la relation **sujet-verbe-objet**.

Ces trois notions motivent les principales étapes de notre approche présentée à la figure 1.1. Précisément, notre contribution porte sur trois volets :

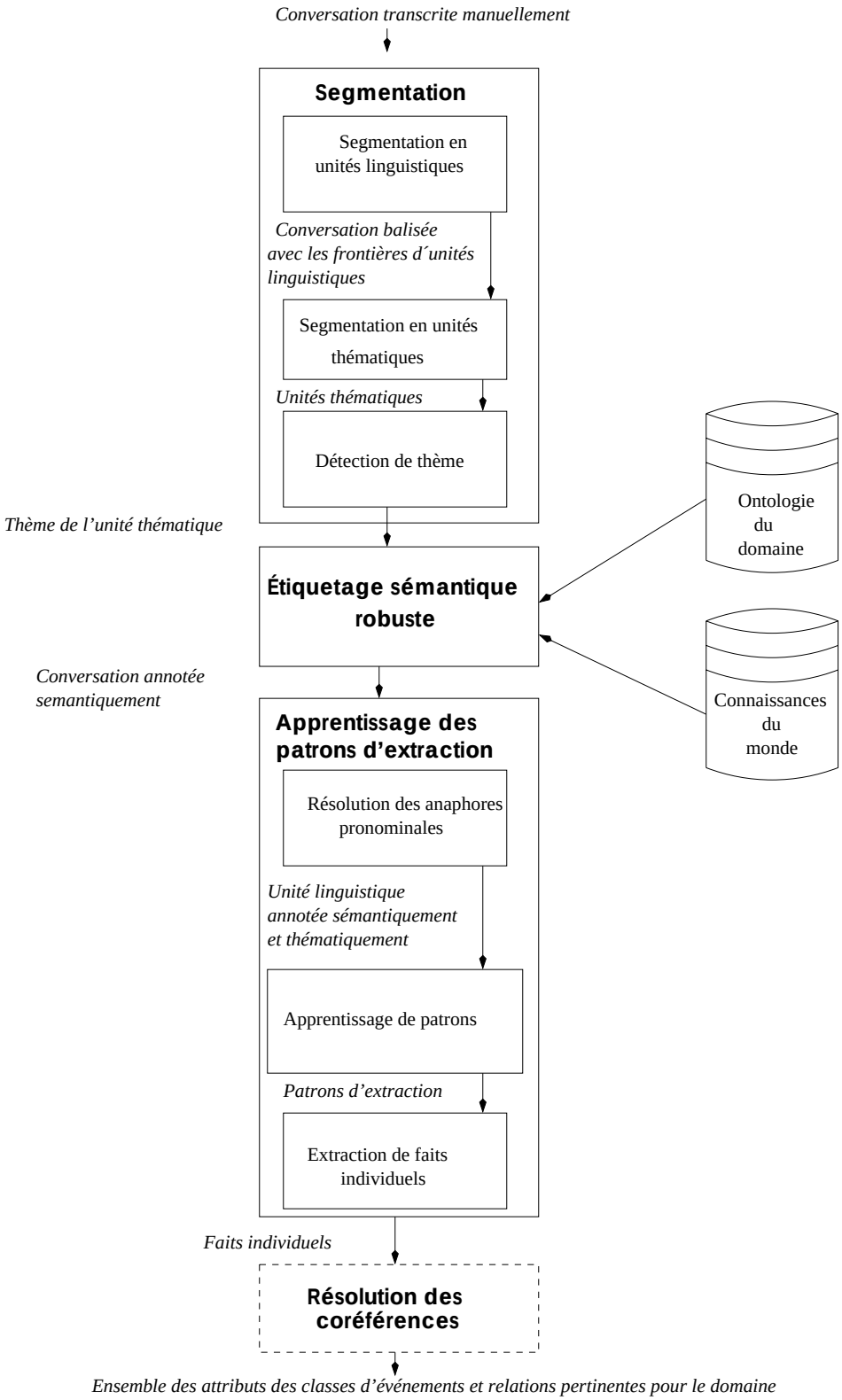


FIG. 1.1 – Étapes de notre approche d'EI à partir de textes conversationnels spécialisés.Les rectangles en ligne continue sont les modules développés dans le cadre de cette thèse. Le cadre en ligne pointillée montre l'étape considérée dans notre approche mais qui n'a pas été développée.

1. L'EI à partir des textes spécialisés :
    - L'extension de l'extraction des entités nommées vers l'étiquetage sémantique de termes du domaine de l'application.
    - La conception et l'implémentation d'un module d'étiquetage sémantique robuste pour l'extraction des termes du domaine et de leurs variantes terminologiques.
2. L'EI à partir des textes conversationnels :
    - L'analyse linguistique des textes conversationnels et la mise en évidence des caractéristiques syntaxiques des informations qui véhiculent l'information.
    - Le choix d'une unité linguistique adéquate pour l'extraction tennant compte de l'aspect interactionnel des conversations et la conception et l'implémentation d'un module de segmentation en unités linguistiques.
    - La définition de la notion d'unité thématique pour faciliter la résolution des anaphores pronominales en position de sujet et l'implémentation du segmenteur.
3. L'apprentissage automatique de patrons d'extraction à partir de textes conversationnels spécialisés :
    - L'implémentation d'une étape de résolution des anaphores pronominales pour gérer la pronominalisation du thème.
    - L'utilisation de modèles de Markov pour gérer la présence de bruit qui modifie la structure syntaxique des énoncés. Nous avons entraîné ces modèles pour générer des relations **prédicat-arguments** qui définissent les relations pertinentes du domaine.

## 1.5 Description des chapitres de la thèse

Le chapitre 2 décrit les principales caractéristiques de notre corpus qui rendent les approches standard d'EI inappropriées. C'est une analyse linguistique des textes conversation-

nels qui aboutit à un ensemble d'hypothèses conditionnant notre approche d'EI.

Le chapitre 3 décrit les étapes du processus standard d'EI. Nous discutons des techniques utilisées par les systèmes les plus performants et nous mettons l'emphase sur leur pertinence pour les textes conversationnels spécialisés. Nous reprenons les hypothèses formulées dans le chapitre 2 pour proposer notre approche d'EI à partir des textes conversationnels spécialisés.

Le chapitre 4 présente la segmentation en unités linguistiques, en unités thématiques et la détection de thème facilitant la tâche d'extraction. Ce chapitre reprend en partie nos travaux [Boufaden *et al.*, 2001; Boufaden *et al.*, 2002] présentés dans le cadre des conférences TALN (Traitement Automatique du Langage Naturel) et NLPRS (Natural Language Processing Pacific Rim Symposium). La partie théorique contient les fondements linguistiques et mathématiques qui ont permis l'élaboration des modules de segmentation. Nous présentons les résultats des expériences effectuées sur notre corpus pour chacun de ces modules.

Le chapitre 5 décrit notre approche pour la construction de l'ontologie du domaine. Nous définissons les classes sémantiques utilisées pour l'annotation.

Le chapitre 6 présente le module d'étiquetage sémantique robuste développé. La partie théorique explicite les fondements mathématiques du module d'étiquetage sémantique robuste développé pour l'annotation des textes conversationnels. Cette partie du chapitre reprend les résultats de nos travaux [Boufaden, 2003] présentés à la conférence ACL (Association for Computational Linguistics). Également, nous présentons le procédé utilisé pour l'étiquetage sémantique des mots sémantiquement similaires aux termes du domaine, c'est-à-dire les variantes terminologiques. Dans cette partie, nous reprenons l'approche et les résultats décrits dans les travaux [Boufaden *et al.*, 2004a; Boufaden *et al.*, 2004b] que nous avons présenté à LREC (Language Ressources and Evaluation Conference) et TALN.

Le chapitre 7 met en relation les différents systèmes présentés dans les chapitres 4 et 5 et 6. Nous développons un module de résolution des anaphores pronominales pour faire émerger plus de relations **prédicat-arguments** et augmenter la couverture des événements pertinents. Nous proposons d'entraîner des modèles de Markov pour l'apprentissage de ces relations. Une partie théorique décrit les fondements mathématiques de chacun des modules.

Nous terminons le chapitre par l'évaluation des patrons appris en comparant les faits individuels extraits avec ceux fournis par le CRDV.

Enfin, dans le chapitre 8, nous concluons cette thèse par une discussion sur nos résultats et contributions et nous présentons nos axes futurs de recherche.

# Chapitre 2

# Caractéristiques des textes conversationnels

## 2.1 Introduction

Dans ce chapitre, nous proposons l'étude des caractéristiques des textes conversationnels dans une perspective d'EI. Une analyse linguistique comparative entre les textes structurés non spécialisés tels que les dépêches journalistiques et notre corpus de textes conversationnels spécialisés met en évidence les caractéristiques qui violent les principes sur lesquels repose l'approche standard d'EI (section 1.2).

Nous commençons par une description de notre corpus. Ensuite, nous nous intéressons aux constituants syntaxiques qui véhiculent l'information pertinente, c'est-à-dire les faits individuels. Nous étudions le choix d'une unité linguistique pour l'EI à partir de textes conversationnels et nous soulignons l'importance de la composante thématique pour la résolution des anaphores, étape nécessaire pour faire émerger les relations pertinentes masquées par la pronominalisation du thème.

| No | Loc | Énoncé |
|---|---|---|
| 1 | a : | Maritime operation centre, (INAUDIBLE) hello. |
| 2 | b : | hi, Mr. Wellington, it's captain Mr. VanHorn |
| 3 | a : | yes. |
| | | ..... |
| 4 | b : | ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the South Coast of Newfoundland, just in the area quite between Fortune Bay and Trepassey. |
| 5 | b : | it's on the south east coast of Newfoundland. |
| | | ..... Incident ..... |
| 6 | b : | this is been going on for, for 24 hours that the case has, or almost anyway, and we had an DFO King Air up flying this morning |
| 7 | b : | they did a radar search for us in that area. |
| 8 | a : | yes. |
| 9 | b : | and their search turned up nothing. |
| 10 | a : | yeah. |
| | | ..... |
| 14 | b : | so I'm wondering about the possibility of attempting it with a different platform perhaps someone with even other sensors other than the radar and, in fact, someone with a, with a radar that'll be a little more sensitive. |
| 15 | b : | before I used the Challenger, I 'll use a Hurk. |
| 20 | a : | ...do you want this thing fired up now or you wanna wait till the Big Boys come in to work tomorrow morning? |
| 21 | b : | well, I would like it if possible, I'd like them to, to be airborne at first light. |
| 22 | a : | ok. |
| 23 | a : | ok then ( INAUDIBLE ). |
| 28 | a : | I think that's about all the information I need to know so, it's a... |
| | | ..... |
| 29 | b : | it's... |
| 30 | a : | an overdue vessel? |
| 31 | b : | overdue 20-feet open boat, a Doray, with a 10-horse power upboard, one person on board. |
| 32 | a : | ok. |
| 33 | a : | so, it's up to the South East Coast? |
| 34 | b : | right. |
| 35 | b : | and it started 18 Zulu on 8, so like 24 hours ago. |
| 36 | a : | ok. |
| 37 | b : | thanks. |

TAB. 2.1 – Conversation Overdue boat du chapitre 1 annotée avec les frontières de segments thématiques déterminés manuellement.

## 2.2 Description du corpus

Notre corpus est une collection de 95 conversations téléphoniques[1] du domaine de la recherche et sauvetage totalisant le modeste nombre de 30 000 mots. Ces conversations on été transcrites manuellement par des personnes qui ne sont pas familières au domaine de la recherche et sauvetage. Nous avons relevé 2,3 % de taux d'erreurs de transcriptions, certaines sont des erreurs d'orthographe comme dans `we have an south east` *`flowing`* à la place de `we have an south east` *`blowing`* tandis que d'autres sont dues au manque de connaissances du domaine comme dans `it's an open` *`Doray`* à la place de `it's an open` *`Dory`*.

Les noms de personnes ont été modifiés pour respecter la confidentialité des conversations ce qui a causé des erreurs telles que des constructions non standards comme `captain Mr.` ou certaines inconsistences dans les noms de personnes utilisés.

Le tableau 2.1 reprend l'exemple de conversation que nous avons appelée Overdue boat au chapitre 1. Les conversations ont été segmentées par des tierces personnes. Les énoncés correspondent à des unités prosodiques identifiées par la chute du ton du locuteur signalant la fin d'une production. Elles coïncident souvent avec la fin d'une contribution ou avec le changement de locuteur. Les virgules correspondent à des temps de pause brefs, les points de suspension à une interruption et le mot `INAUDIBLE` indique un segment de l'énoncé qui n'est pas transcrit. Les locuteurs sont souvent des personnes qui se connaissent et de fait cela introduit un style informel où les locuteurs n'observent pas toujours les règles sociales de conversation [Grice, 1975] régissant les échanges. En particulier, un locuteur peut s'octroyer plusieurs tours de parole sans attendre ou interrompre son interlocuteur. Par conséquent, certains segments sont dans un style narratif (énoncés 4-7, tableau 2.1) correspondant à une étape où le locuteur communique l'ensemble des informations qu'il détient.

Certaines conversations parlent d'un même incident mais sont complémentaires. Les locuteurs parlent d'un incident et s'appellent par téléphone au fur et à mesure que de nou-

[1] 2 % des conversations sont des appels téléphoniques impliquant deux locuteurs et une tierce personne. Ces conversations ont été retirées du corpus car elles sont problématiques pour l'étape de segmentation.

velles informations apparaissent. Les différents appels forment différentes conversations qui tournent autour du même événement. Cela signifie qu'un objet ou événement peut être référencé pendant toute une conversation sans qu'il ne soit explicitement nommé.

### 2.2.1 Irrégularités langagières de l'oral

Le taux d'erreur des transcriptions (mauvaise reconnaissance des mots, fautes d'orthographe, etc.) est de 3 %. Le tableau 2.2 présente certaines statistiques sur le nombre d'irrégularités langagières de l'oral calculées sur 9 conversations tirées de notre corpus. Trois types d'irrégularités sont répertoriés sur les énoncés de plus de deux mots (369 énoncés) :

**Les répétitions** qui sont des suites de mots identiques et de même catégorie syntaxique. Une répétition peut être partielle (partie d'un syntagme) ou totale (syntagme complet). Les mots soulignés dans l'exemple (1) illustrent une répétition partielle :

(1) `well, I would like it if possible, I'd like them to, to be airborne at first light.`

**Les omissions** ou l'absence d'un constituant syntaxique tel que le sujet. Dans l'exemple (2), les constituants syntaxiques omis sont le groupe nominal qui représente le sujet et le verbe. Les constituants et particules omis sont entre crochets. Nous avons reconstitué ces éléments pour mettre en évidence l'effet des omissions sur la structure syntaxique de l'énoncé :

(2) $[it]_{gn}$ $[is]_{v}$ $[an]_{det}$ `overdue 20-feet open boat,` $[it]_{gn}$ $[is]_{v}$ `a Doray, with a 10-horse power upboard,` $[and]_{conj}$ `one person on board.`

Les indices *gn*, *v*, *det* et *conj* repésentent respectivement un groupe nominal, un verbe, un déterminant et une conjonction.

**Les reprises** qui sont des interruptions suivies d'un nouveau syntagme. Il y a alors abandon d'un constituant syntaxique et début d'un nouveau constituant. L'énoncé de l'exemple (3) est une reprise qui commence à la position indiquée par le symbole ↑.

| Irrégularité | Ensemble des énoncés (369) | Énoncés pertinents (80) |
|---|---|---|
| Répétitions | 15,0 % | 10,0 % |
| Omissions de mots | 9,2 % | 12,5 % |
| Reprises | 9,7 % | 12,5 % |

TAB. 2.2 – Taux de différentes irrégularités langagières de l'oral dans l'échantillon de 9 conversations.

(3) `this is been going on for, for 24 hours that the case has ↑, or almost anyway, and we had an DFO King Air up flying this morning.`

Le bruit introduit par les irrégularités langagières de l'oral représente un obstacle pour l'analyse syntaxique complète d'un énoncé et par conséquent pour la construction de la relation **sujet-verbe-objet**.

Pour remédier à ce problème, plusieurs approches d'analyse syntaxique partielle ont été développées [Lavie *et al.*, 1997; Ballim et Russell, 1994]. Certaines utilisent un traitement en aval de l'analyse syntaxique pour le recouvrement de l'analyse complète, tandis que d'autres intègrent des heuristiques pour le traitement des irrégularités durant l'analyse syntaxique [Boufaden *et al.*, 1998; Langer, 1990] afin de générer directement une analyse complète.

Pour nos textes, nous avons retenu l'option d'une analyse syntaxique partielle. Aucun traitement n'est effectué pour obtenir une analyse complète. L'avantage de ce choix est d'éviter de supprimer certains constituants qui contiennent de l'information pertinente. Toutefois, cela engendre une surgénération de constituants syntaxiques qui peut conduire à l'extraction de plusieurs réponses pour un même champ.

Plusieurs irrégularités sont liées, notamment les répétitions et omissions de mots. De ce fait, les taux que nous obtenons ne peuvent être cumulés pour obtenir le taux global d'irrégularités dans l'échantillon de corpus. Sur les 369 énoncés de plus de 2 mots, 142 contiennent au moins une irrégularité, soit 38,5 %. Enfin, les interruptions causées par l'intervention d'un autre locuteur représentent plus de 7,8 % des énoncés.

### 2.2.2 Vocabulaire du domaine

Les textes spécialisés sont des sous-langages caractérisés par un vocabulaire spécialisé et parfois une syntaxe particulière. Les rapports météorologiques et les articles scientifiques du domaine de la pharmacologie sont des exemples de sous-langages.

Notre corpus est composé de textes spécialisés dans le domaine de la recherche et sauvetage maritime. Ces textes se caractérisent par :

- Un vocabulaire du domaine composé de mots et expressions qui sont :
  - Le type d'incident, tels que `missing` ou `overdue`.
  - La cause d'un incident, tels que `dead battery`, `fire` ou `engine failure`.
  - Le moyen utilisé pour signaler un incident, tels que `witness report`, `flares` et `ELT` (Emergency Locator Transmitter).
  - Le moyen utilisé pour la recherche, tels que `radar`, `SARSAT` (Search And Rescue Satellite-Aided Tracking), `goggles` et `divers`.
  - Les noms d'avions et de bateaux alloués pour la recherche, tels que `Aurora`, `King Air` et `Challenger`.
  - La disponibilité d'une ressource, tels que `ready` ou `available`.
  - Le type d'un bateau, tels que `Zodiacs`, `20-footer` ou `yacht`.
  - La description d'un bateau, tels que `red` ou `trims`.
  - L'état d'un bateau, tels que `sinking` ou `drifting`.
  - Les conditions météorologiques, tels que `foggy`, `fog`, `visibility`, `sea` et `haze`.
  - L'état d'avancement et le résultat d'une mission de recherche, tels que `completed` ou `nothing`.

  Certaines de ces expressions sont des termes (`ELT`, `SARSAT`), tandis que d'autres sont des mots communs qui ont une sémantique particulière dans le contexte du domaine de la recherche et sauvetage (`flares` ou `gun` qui tous deux sont des indicateurs d'appels de détresse).
- Des constructions syntaxiques particulières pour représenter les positions en termes de latitude et longitude, tels que `47 degrees 23 minutes North` ou `4640 0 North`,

5409 4 Wetecker.

La plupart des termes et expressions sont des groupes nominaux, notamment des noms de bateaux, avions, organisations et lieux, bien que ce ne soit pas toujours le cas, en particulier pour les conditions météorologiques comme en témoignent les extraits du tableau 2.3.

| No | Loc | Énoncé |
|---|---|---|
| 2 | a : | Ha, we got a fairly ( INAUDIBLE ) for tomorrow. |
| | b : | Oh yeah. |
| 6 | b : | Ha, the weather is still shitty. |
| | a : | Yeah. |
| 29 | a : | I know that the, the weather here is foggy but I just talked to somebody in the community ( INAUDIBLE ) and they got at least 10 miles of visibility there. |
| 53 | b : | Weather on scene ? |
| | a | It was fairly good, it was, the visibility was ( INAUDIBLE ) 1 mile. |
| | b | OK. |

TAB. 2.3 – Extraits de quatre conversations où les conditions météorologiques sont exprimées par des adjectifs.

Enfin, certains concepts du domaine tels que les concepts INCIDENT et WEATHER-CONDITIONS sont parfois exprimés avec des variantes des termes du domaine. Cette caractéristique est une conséquence du caractère spontané des conversations. Ainsi, un même type d'incident, par exemple une panne de moteur, peut être exprimé par les expressions they're having a tough time starting their engine ou it's for an engine failure et le beau temps peut être exprimé par des expressions aussi variées que the weather is clear ou the weather is pretty calm.

L'identification des expressions sémantiquement similaires au vocabulaire du domaine est une problématique similaire à celle de la détection des variantes terminologiques. Nous revenons sur ce problème dans le chapitre 6 et proposons une approche pour la détection des variantes du vocabulaire.

## 2.3 Catégorie syntaxique des faits individuels

Dans cette section, nous nous intéressons aux variations lexicales entre les textes écrits et conversationnels afin de déterminer quels constituants syntaxiques (groupes nominaux, verbes et adjectifs) véhiculent l'information pertinente et constituent des faits individuels.

Plusieurs travaux en linguistique [Biber, 1988; Halliday, 1989] ont étudié les variations linguistiques entre les textes écrits et parlés. Bien que certaines divergences existent entre linguistes, il existe un consensus sur certaines caractéristiques lexicales qui distinguent ces deux types de textes. Nous retenons quatre caractéristiques importantes pour la tâche d'EI :

1. Le taux de noms est corrélé avec la densité d'information d'un texte. Les groupes nominaux sont les constituants de base qui véhiculent l'information [Biber, 1988; Halliday, 1989].

2. Le taux d'occurrence du verbe `be` comme verbe principal et le taux d'adjectifs[2] indiquant le style d'expression (concis/élaboré) utilisé dans le texte. D'une part, les constructions de type **gn-verbeBE-adjectif** sont des constructions simples, caractéristiques des textes oraux et utilisées pour exprimer de l'information. D'autre part, les adjectifs attributifs[3] sont généralement utilisés pour modifier le nom [Biber, 1988].

3. Le taux de nominalisation qui distingue les constructions verbales caractéristiques des textes oraux (*being treated*) de celles basées sur les noms caractéristiques des textes écrits (*having treatment*).

4. Le ratio type/occurrence[4] est caractéristique de la variété du vocabulaire et indique l'importance des variations langagières. D'un point de vue de l'EI, il est plus difficile de repérer les événements pertinents lorsque ceux-ci sont exprimés dans diverses formes lexicales.

L'exemple de la figure 2.1 illustre la différence de style entre les textes écrits et les textes

[2]En particulier, les adjectifs prédicatifs, qui apparaîssent après les verbes tels que BE, SEEM.
[3]Un adjectif attributif précède toujours un nom et non après les verbes tels que BE, SEEM.
[4]C'est le nombre d'unités lexicales (chaînes de caractères entre deux espaces distinctes) sur le nombre total de toutes les chaînes de caractères.

**Phrase dans un texte écrit :**
*The use of this method of control* unquestionably leads to safer and faster train running in *the most adverse weather conditions*.

**Variante de cette phrase en langage parlé :**
*If this method of control* is used trains will unquestionably (be able to) run more safely and faster (even) when *the weather conditions are most adverse*.

FIG. 2.1 – Phrase extraite d'un texte écrit et sa variante exprimée en langage oral [Halliday, 1989, p. 79] pour illustrer les variations linguistiques entre ces deux modes de communication.

oraux. En particulier, nous remarquons le remplacement des groupes nominaux par des constructions **gn-verbeBE-adjectifPrédicatif** :

- ***the use of this method of control*** est remplacé par `If the method of control is used`.
- ***the most adverse weather conditions*** est remplacé par `weather conditions are the most adverse`.

### 2.3.1 Variations lexicales du corpus

Dans le tableau 2.4, nous présentons les scores des variations lexicales que nous avons retenues pour cette étude. Chaque score correspond à un pourcentage calculé pour trois types de textes : notre corpus (conversations téléphoniques spécialisées), des dépêches journalistiques et des conversations téléphoniques non spécialisées. Les scores rapportés pour les deux derniers types de textes ont été tirés de l'ouvrage de Biber [Biber, 1988] et ont été calculés pour 44 dépêches journalistiques extraites du corpus LOB[5] et 27 conversations téléphoniques extraites du corpus London-Lund[6] pour mieux situer notre corpus.

[5] `http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html`
[6] `http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTML`

| | Corpus | Conv. téléphoniques | Dép. journalistiques |
|---|---|---|---|
| Noms | 17,5 % | 13,5 % | 22,0% |
| Be verbe principal | 7,6 % | 4,3 % | 2,1 % |
| Adjectifs prédicatifs | 1,3 % | 0,6 % | 0,3 % |
| Nominalisation | 0,8 % | 0,7 % | 1,9 % |
| Ratio type/occurence | 5,8 | 4,6 | 5,5 |
| Pronom It | 7,3 % | 2,2 % | 0,6 % |

TAB. 2.4 – Variations linguistiques entre notre corpus, des conversations téléphoniques non spécialisées et des dépêches journalistiques. Les scores pour ces deux derniers types de textes sont donnés par Biber [1988].

### 2.3.2 Analyse des scores

Notre analyse de corpus tient compte des questions suivantes :

1. Quelle est la nature syntaxique des expressions qui constituent les faits individuels à extraire ?
2. Comment assurer la bonne couverture des événements et relations pertinents compte tenu de la taille modeste du corpus et la variété du vocabulaire ? Faut-il choisir une approche lexicale dans la définition des patrons ou, plutôt, favoriser une approche basée sur les classes de mots ?

Nous tirons trois constatations :

1. Le taux de noms dans notre corpus se situe entre les taux de noms observés pour les dépêches journalistiques et celui des conversations téléphoniques non spécialisées. Une façon d'expliquer ce score est, d'une part, que notre corpus est à but informatif et par conséquent présente une grande densité d'information qui est en partie véhiculée par les noms. D'autre part, le score élevé du verbe `be` et des adjectifs prédicatifs **gn-verbeBE-adjectifPrédicatif** montre qu'une importante partie de l'information est exprimée à travers les adjectifs prédicatifs qui modifient les noms. Les exemples (4), (5) et (6) montrent quelques constructions de type **gn-verbeBE-adjectifPrédicatif** tirées de notre corpus.

(4) `They can't do a visual search, their radar is degraded, but I've asked them for a radar search anyway.`

(5) `So, if you think the Aurora is suitable, lets go for it.`

(6) `I know that the, the weather here is foggy but I just talked to somebody in the community ( INAUDIBLE ) and they got at least 10 miles of visibility there.`

2. Le taux de nominalisation similaire à celui observé pour les conversations d'ordre général indique que les énoncés sont moins élaborés que les phrases des textes écrits [Biber, 1988]. À partir des observations (1) et (2), nous tirons un première constatation : outre les noms, les verbes et les adjectifs sont importants pour la tâche d'extraction car ce sont des constituants syntaxiques qui véhiculent de l'information pertinente.
3. La diversité du vocabulaire est aussi importante dans notre corpus que dans les dépêches journalistiques. Cela implique que les relations et événements pertinents peuvent être exprimés dans des termes variés difficiles à couvrir de manière exhaustive. Ce problème connu en EI à partir des textes structurés est une des principales limites de l'approche standard d'EI.

## 2.4 Une unité syntaxico-sémantique maximale pour les textes conversationnels ?

La structure des patrons d'extraction repose sur la relation **sujet-verbe-objet** qui établit les rôles thématiques des constituants syntaxiques d'une phrase. Le choix d'une unité pour l'EI doit être motivé par l'existence de cette relation. Pour les textes écrits, l'unité utilisée en EI est la phrase, laquelle est une unité syntaxico-sémantique maximale, délimitée par une ponctuation explicite et contenant une relation syntaxique **sujet-verbe-objet**.

Contrairement au texte écrit, la conversation est une forme d'expression interactive et spontanée. Les échanges entre locuteurs s'effectuent selon des règles sociales connues [Sacks

*et al.*, 1974]. Un locuteur interagit en s'assurant que ses propos restent cohérents avec ceux des autres participants. L'interprétation d'un énoncé est donc fortement influencée par le contexte d'énonciation.

Ainsi, il est difficile d'établir un parallèle entre la phrase en tant qu'unité syntaxico-sémantique maximale utilisée pour les textes écrits et l'énoncé qui en soi ne bénéficie pas d'un consensus [Traum et Heeman, 1997]. Le choix d'une unité de traitement pour la tâche d'EI à partir des textes conversationnels est compliqué par deux facteurs :

1. L'absence de consensus sur la définition de ce que représente un énoncé.
2. L'aspect interactionnel des textes conversationnels qui soulève deux questions importantes relativement à la tache d'extraction. Quelle est l'unité linguistique qui garantit des propriétés syntaxico-sémantiques maximales ? Faut-il se limiter à l'énoncé ou plutôt considérer les paires d'adjacence ?

### 2.4.1 Absence de consensus sur l'énoncé

Plusieurs linguistes et psycholinguistes se sont penchés sur la définition d'un énoncé. Certains [Sacks *et al.*, 1974; Ford et Thompson, 1991; Fries, 1952; Nakajima et Allen, 1993] associent l'énoncé au tour de parole qui est une suite ininterrompue de mots produits par un seul locuteur. Ce découpage est relativement facile, mais présente l'inconvénient de contenir parfois plus d'une relation **sujet-verbe-objet** ou au contraire, une relation incomplète (cas d'une interruption par l'interlocuteur).

D'autres [Halliday et Hassan, 1976; Gross *et al.*, 1993; Takagi et Itahashi, 1996] le définissent comme une unité prosodique qui est un segment délimité par des informations telles que l'intonation. Ce découpage est motivé par l'idée que les changements de ton coïncident avec la fin d'une contribution [Heeman, 1999]. Un inconvénient de ces découpages est que, dans les situations de chevauchement de la parole, il est difficile de déterminer l'endroit de la frontière de l'unité prosodique.

Enfin, certains [Stolcke et Shriberg, 1996; Meeter et Iyer, 1996; Levelt, 1989] définissent l'énoncé comme une unité linguistique ayant des propriétés syntaxiques et/ou sémantiques.

Ce découpage est approprié pour les applications de TAL car il garantit la présence de la relation **sujet-verbe-objet**. Toutefois, ce découpage est difficile à réaliser puisque l'identification de la frontière de l'unité linguistique repose sur la complétude syntaxico-sémantique de l'énoncé. Rappelons que les irrégularités langagières de l'oral constituent un obstacle à l'analyse syntaxique complète d'un énoncé.

Un moyen de pallier ce problème est d'utiliser des paires d'adjacence qui sont deux tours[7] de parole contigus, produits par des locuteurs différents. La production de la première partie de la paire exerce une contrainte sur la seconde et justifie sa production. Levinson [1983] définit une paire d'adjacence comme suit :

> *An adjacency pair is a unit of conversation that contains an exchange of one turn each by two speakers. The turns are functionally related to each other in such a fashion that the first turn requires a certain type or range of types of second turn.[Levinson, 1983, p. 303-304]*

Nous proposons d'utiliser la paire d'adjacence, comme unité linguistique[8] car, d'une part, elle présente des propriétés sémantiques intéréssantes. D'autres part, elle est délimitée par des marques lexicales et prosodiques qui facilitent sa détection automatique. Par ailleurs c'est l'unité de base utilisée pour l'analyse des conversations [Levelt, 1989] et pour leur traitement automatique[Stolcke et Shriberg, 1996].

Des exemples de paires d'adjacence sont les question-réponse (énoncés `20-21` et `30-31`, tableau 2.1), ou assertion-acquiescement (énoncés `7-8` et `9-10`).

### 2.4.2 Analyse du discours : rôle de la structure thématique

L'importance de la dimension discursive pour l'analyse des conversations a été soulignée dans plusieurs ouvrages [Levelt, 1989; Sacks *et al.*, 1974; Maynard et Zimmerman, 1984; Grosz et Sidner, 1986]. En particulier, Grosz et al. [Grosz et Sidner, 1986] parlent de structure

[7]

[8] Dans la section **??** nous montrons que l'unité linguistique varie selon le style du discours. Dans le contexte particulier de nos conversations, certaines parties de conversations ont un style conversationnel, tandis que d'autres ont un style narratif. Pour tenir compte de ces spécificités, nous avons modifié la définition de l'unité linguistique pour qu'elle soit une paire d'adjacence ou un tour de parole selon le style du discours.

thématique hiérarchique d'une conversation. Une conversation est organisée en thèmes qui correspondent à différents buts informatifs. Chaque thème est une séquence d'énoncés qui s'articulent autour d'un même sujet ou *focus*. Par exemple, la conversation du tableau 2.1 est segmentée en quatre unités thématiques. Les deuxième, troisième et quatrième segments portent respectivement sur l'incident, la mission de recherche et l'allocation d'une unité de recherche.

En EI à partir de textes structurés, l'analyse du discours est prise en compte à l'étape de résolution des coréférences après l'extraction des faits individuels. Le but de cette étape est de rattacher les informations qui sont des descriptions partielles d'événements ou des relations qui n'ont pu être extraites à l'étape précédente, c'est-à-dire l'extraction des faits individuels à cause des coréférences. Les coréférences sont davantage utilisées dans les textes conversationnels que dans les textes structurés, comme le montre par exemple la différence des taux de pronoms `It` (tableau 2.4). L'augmentation des coréférences est due notamment à la pronominalisation du thème dans les énoncés qui composent une unité thématique.

W. Levelt souligne que durant le développement d'une unité thématique, souvent l'objet principal du thème développé est référencé par une anaphore pronominale en position de sujet :

> *When the speaker's purpose is to expand the addressee's knowledge about something, the message will highlight this topic concept, to distinguish it from the comment that is made about it [...] in a syntactic prominent position. [...] the topic is encoded as grammatical subject.* [Levelt, 1989, p. 260-265]

La pronominalisation du thème est un moyen pour le locuteur d'indiquer que des propos sont thématiquement cohérents.

D'un point de vue de l'EI, certaines relations pertinentes sont masquées par la pronominalisation d'un de leurs arguments. Conséquemment, ces derniers sont inaccessibles à l'étape d'extraction des faits individuels.

Pour l'étape d'apprentissage des patrons, cela implique que ces relations ne seront pas prises en compte dans le processus de généralisation des patrons d'extraction. Ainsi, la pro-

nominalisation a un impact sur la complétude de la couverture des schémas d'extraction et sur la performance de l'extraction des faits individuels.

Le tableau 2.5 est un exemple de ces cas, où la relation VESSEL `is` on LOCATION est masquée par le pronom `it`.

| No Loc Énoncé |
|---|
| 4 b : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat (VESSEL) on the South Coast of Newfoundland (LOCATION), just in the area quite between Fortune Bay and Trepassey (LOCATION). |
| 5 b : it (VESSEL) 's on the south east coast of Newfoundland (LOCATION). |

TAB. 2.5 – Exemple d'unité thématique dont l'objet principal est le bateau en retard. Cette unité est extraite de la conversation Overdue boat.

## 2.5 Synthèse

L'analyse linguistique de notre corpus met en évidence cinq problématiques reliées à l'EI à partir des textes conversationnels spécialisés dont il faut tenir compte dans la conception de notre approche d'EI :

1. Le vocabulaire du domaine est composé de mots et termes spécialisés définissant les concepts pertinents qui sont différents des entités nommées. La couverture de ce vocabulaire nécessite une étape d'étiquetage sémantique qui va au-delà de l'extraction des entités nommées tels que définies dans les MUC.
2. Certains concepts sont formulés en utilisant des expressions sémantiquement similaires au vocabulaire du domaine. L'étiquetage sémantique de ces variantes permet une meilleure couverture des entités pertinentes.
3. Les textes conversationnels s'organisent autour de tours de parole. Pour tenir compte de l'aspect interactif de ces textes, il est nécessaire de définir une unité linguistique qui tient compte de certaines dépendances telles que celles présentes dans les paires

question-réponse. La paire d'adjacence est une unité linguistique qui tient compte de ces particularités.

4. L'inférence d'une relation **prédicat-argument** à partir de la relation synatxique **sujet-verbe-objet** est compliquée par la présence des irrégularités langagières de l'oral qui sont un obstacle à la génération d'une analyse syntaxique complète. Par conséquent, les rôles thématiques ne sont pas accessibles à partir de la structure synatxique. Nous proposons d'accéder aux rôles thématiques en apprenant les relations **prédicat-argument** à partir d'unités linguistiques annotées sémantiquement.
5. La pronominalisation du thème est un phénomène très présent dans les textes conversationnels. Elle a pour effet de masquer certaines relations pertinentes, ce qui diminue le nombre de relations disponibles pour l'apprentissage des patrons. Cela a des conséquences négatives sur le processus de généralisation à l'étape d'apprentissage et sur la qualité de la couverture des relations pertinentes générées automatiquement.

## 2.6 Conclusion

Dans ce chapitre, nous avons analysé l'effet des caractéristiques linguistiques des textes conversationnels sur le processus d'EI et plus particulièrement sur l'étape d'extraction des entités pertinentes au domaine et sur l'étape d'apprentissage des patrons d'extraction. Le but de cette analyse est de mettre en évidence les particularités de ces textes pour en tenir compte dans la description de notre approche d'EI.

Dans ce qui suit, le chapitre 3 décrit les principales techniques utilisées pour l'extraction des entités nommées et l'apprentissage des patrons d'extraction. Nous choisissons les techniques qui répondent le mieux aux problématiques décrites dans la section 2.5 pour définir notre propre approche d'EI. Enfin, nous décrivons l'architecture de notre module d'apprentissage de patrons d'extraction.

# Chapitre 3

# Extraction d'information à partir de textes conversationnels

## 3.1 Introduction

Le but de l'EI est de générer une représentation tabulaire des informations pertinentes sélectionnées dans un texte. C'est aussi un moyen de distiller de grandes quantités d'information pour fournir un inventaire des faits saillants et de leurs relations pour un domaine en particulier.

Dans ce chapitre, nous présentons une étude sur la portabilité de la technologie de l'EI vers des textes conversationnels spécialisés. Nous mesurons l'effet des caractéristiques des textes conversationnels spécialisés sur une approche standard d'EI. Nous discutons des techniques les plus utilisées pour ces étapes d'un point de vue de leur adéquation pour le traitement de textes de l'oral. Enfin, nous concluons ce chapitre par la proposition d'une approche d'EI pour les textes conversationnels spécialisés.

## 3.2 Historique des systèmes d'EI

Les premiers systèmes d'EI ont été développés au début des années 80 dans le but d'analyser des documents narratifs dans des domaines spécialisés tels que la chimie [Reeker *et al.*, 1983] et la médecine [Sager *et al.*, 1987]. La tâche d'extraction consistait à extraire des faits individuels pour peupler une base de données. Toutefois, la majeure partie des travaux en EI a été réalisée dans le cadre des campagnes d'évaluation Message Understanding Conference [Hirschman, 1991] où la tâche était d'analyser des textes non spécialisés tels que les dépêches journalistiques. Ces campagnes ont abouti à l'adoption d'une approche standard d'EI largement inspirée du système FASTUS développé par Hobbs *et al.* [1997].

Dans ce qui suit, nous présentons un bref historique de ces campagnes, les tâches qui ont été définies dans leur cadre et les métriques d'évaluation avec les meilleurs scores obtenus lors des différentes campagnes.

### 3.2.1 Historique des MUC

Les deux premières conférences MUCKI[1] été organisées et financées par la Defense Advanced Research Projects Agency (DARPA) et le Naval Ocean Systems Center (NOSC) dans le but d'extraire des informations à partir de messages narratifs provenant de la marine militaire. MUCKI s'est déroulée en 1987 et était plutôt une conférence exploratoire. Ce n'est qu'à partir de la deuxième conférence, en 1989, que des normes ont été définies pour l'évaluation de cette tâche. En 1991, une nouvelle série de conférences plus ambitieuses que les précédentes a commencé dans le cadre du programme TIPSTER aussi financé par la DARPA et qui incluait deux autres conférences : TREC (Text Retrieval Conference) consacrée à l'évaluation des systèmes de recherche d'information ainsi que SUMMAC (SUMMArisation Conference) consacrée aux systèmes de résumé automatique. La différence entre ces deux séries se situe au niveau de la complexité des textes analysés, la dimension du corpus, la nature et la difficulté de la tâche d'extraction ainsi que la façon avec laquelle les réponses sont évaluées. Ainsi, les deux premières conférences de la série MUC, MUC-3 en 1991 et MUC-

[1] Les premières conférences MUC portaient le nom de MUCK, où le K fait référence à *knowledge*.

4 en 1992, ont porté sur des dépêches journalistiques à propos d'attentats terroristes. À ces conférences, la tâche d'extraction consistait à remplir un seul formulaire qui englobait toutes les informations pertinentes. Lors de MUC-5, en 1993, les textes analysés étaient des dépêches journalistiques contenant des annonces de partenariat ainsi que des annonces de produits micro-électroniques, et ce dans deux langues : l'anglais et le japonais. Ce n'est que lors de MUC-6, en 1995, que la tâche d'extraction a été divisée en trois sous-tâches incluant deux nouvelles tâches : l'extraction des entités nommées et la résolution des coréférences. Les textes analysés étaient des dépêches extraites du Wall Street Journal rapportant les co-entreprises et les événements de succession de postes de responsabilité. Enfin, la dernière conférence MUC-7, en 1998, a porté sur des comptes-rendus commerciaux et sur des lancements de fusées. Cette conférence est la dernière de la série et elle dresse un bilan des points forts et des limites de cette technologie.

### 3.2.2 Description des tâches

Les premiers MUC (MUC-3, MUC-4, MUC-5) se sont limités à évaluer la tâche de remplissage d'un formulaire (*scenario template*), mais une telle évaluation s'est avérée trop globale pour permettre une analyse détaillée. Ce n'est que lors de MUC-6 que la tâche d'extraction a été divisée en trois sous-tâches :

1. Extraction des entités nommées. Permet d'identifier les principaux actants dans le domaine de l'application.
2. Extraction des faits individuels pour remplir les formulaires *Element*. Ces derniers sont des formulaires rattachés aux actants tels que les personnes et organisations. Le tableau 3.1 présente un exemple de ces formulaires. L'approche d'EI que nous proposons permet de remplir ce type de formulaire.
3. Résolution des coréférences pour remplir les formulaires *Scenarios* qui décrivent les événements et les relations entre les entités d'un document. Un exemple de formulaire *Scenario* relatif à un événement de succession est présenté au tableau 3.2. Les informations qui remplissent les champs sont obtenues par la combinaison des faits individuels

| *Organisation* | |
|---|---|
| 1 ORG-NAME | Nom de l'organisation. |
| 2 ORG-ALIAS | Alias de cette organisation. |
| 3 ORG-DESCRIPTOR | Groupe nominal qui décrit l'organisation ou lui fait référence. |
| 4 ORG-TYPE | Organisation gouvernementale ou corporative, par exemple. |
| 5 ORG-LOCAL | Endroit où est localisée l'organisation. |
| 6 ORG-COUNTRY | Pays ou région où se situe l'organisation. |

TAB. 3.1 – Exemple de formulaire *Element* tel que défini lors de MUC-6.

| *Succession* | |
|---|---|
| 1 SUCCESSION-ORG | Pointeur sur l'organisation jouant un rôle dans l'événement de succession. |
| 2 POST | Titre de l'ancien/nouveau poste occupé. |
| 3 IN-AND-OUT | Pointeur sur le formulaire qui contient les informations relatives à la personne qui quitte/accepte le poste. |
| 4 VACANCY-REASON | La raison pour laquelle le poste est vacant ou deviendra vacant. |

TAB. 3.2 – Exemple de formulaire *Scenario* pour un événement de succession tel que défini lors de MUC-6.

extraits à l'étape précédente. La combinaison inclut la résolution des coréférences et l'inférence de faits en exploitant des connaissances du monde.

### 3.2.3 Évaluation des systèmes d'extraction d'information

Les métriques utilisées pour l'évaluation des systèmes d'extraction d'information ont été inspirées de celles utilisées en recherche d'information. Hobbs [2002] les présente d'une façon très éloquente :

> *Recall is a measure of completeness, precision of correctness. When you promise to tell the whole truth, you are promising 100% recall. When you promise to tell nothing but the truth, you are promising 100% precision.*

De manière plus formelle, ces métriques sont :

- Le rappel, $R$, représentant le nombre de réponses correctes générées par le système divisé par le nombre de réponses fournies par les évaluateurs.
- La précision, $P$, représentant le nombre de réponses correctes générées par le système divisé par le nombre de réponses générées par le système.
- Le F-score qui combine ces deux métriques :

$$F = \frac{(\beta^2 + 1.0)PR}{\beta^2 P + R}$$

où $\beta$ est l'importance relative associée à la précision par rapport au rappel.

Le tableau 3.3 présente les meilleurs scores enregistrés pour les tâches énumérées dans la section 3.2.2 lors des différents MUC.

L'analyse des résultats obtenus pour la tâches d'extraction des événements (formulaire *Scenario*) montre trois raisons expliquant le faible F-score pour cette tâche :

1. Les informations distantes. Cela se produit lorsqu'il y a une apposition qui est un complément d'information sur un nom, comme dans `Mr. James, 57 years old, is stepping down as chief executive officer`. Dans ce cas d'apposition, le segment de phrase `57 years old` s'interpose dans la relation syntaxique **sujet-verbe-objet** et empêche l'application du patron PERSON `is stepping down as` POSITION.
2. La difficulté à couvrir l'ensemble des classes d'événements et relations pertinentes à un domaine. D'une part, les variations langagières font qu'un événement peut être exprimé sous plusieurs formes lexicales. D'autre part, les exemples de ces variations langagières n'apparaissent pas suffisamment dans les textes : il y a un problème de sous-représentation des données.
3. Le report de l'étape de résolution des coréférences, en particulier celles impliquant les entités nommées, après l'étape d'extraction des faits individuels. Les coréférences masquent certaines relations pertinentes, réduisant ainsi la couverture de ces constructions (section 2.3.2). Par conséquent, certaines informations relatives aux entités nommées ne sont pas détectées à l'étape d'extraction des faits individuels.

| Tâche | Extraction des entités nommées | Résolution des coréférences | Formulaire *Element* | Formulaire *Scenario* |
|---|---|---|---|---|
| MUC-3 | | | | 69 % |
| MUC-4 | | | | 56 % |
| MUC-5 | | | | 53 % |
| MUC-6 | 97 % | 71 % | 80 % | 57 % |
| MUC-7 | 94 % | 62 % | 87 % | 51 % |
| Humain | 98 % | | | 97 % |

TAB. 3.3 – Meilleurs scores obtenus pour les différentes tâches avec $F(\beta = 0,5)$. Les scores des tâches d'extraction des entités nommées et de résolution des coréférences apparaissent uniquement pour les deux dernières MUC puisque ces tâches ont été introduites lors de MUC-6.

### 3.2.4 Enjeux et défis de l'EI

Les travaux en EI se sont divisés en deux axes principaux selon le but recherché :

**L'amélioration du niveau de performance des systèmes** Le projet ACE (Automatic Content Extraction)[2] est un exemple des efforts développés pour l'étude de l'intégration des technologies d'EI au sein d'autres applications du TAL tels que les moteurs de recherche d'information ou les systèmes de génération de résumés.

**La conception de systèmes portables** La portabilité soulève plusieurs problèmes car le changement de domaine nécessite généralement la construction d'un nouveau lexique et la conception de nouveaux patrons d'extraction, deux étapes qui forment le goulot d'étranglement du processus d'EI.

Le fait d'effectuer l'EI à partir de textes spécialisés tels que des articles en biomédecine ou des rapports météorologiques a un effet direct sur les classes sémantiques de l'information recherchée et par conséquent sur les procédés utilisés. Par exemple, depuis le début des années 2000, des travaux en EI[3] dans le domaine de la biomédecine [Hobbs, 2002; Grishman *et al.*, 2002] ou sur le "réchauffement de la planète" [Aitken, 2002] ne parlent pas d'extraction d'entités nommées, mais plutôt d'extraction

[2] `http://www.itl.nist.gov/iad/894.01/tests/ace/`

[3] Ces travaux ont été présentés entre autres lors du *workshop* ACL sur le traitement des textes en biomédecine. `http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm`

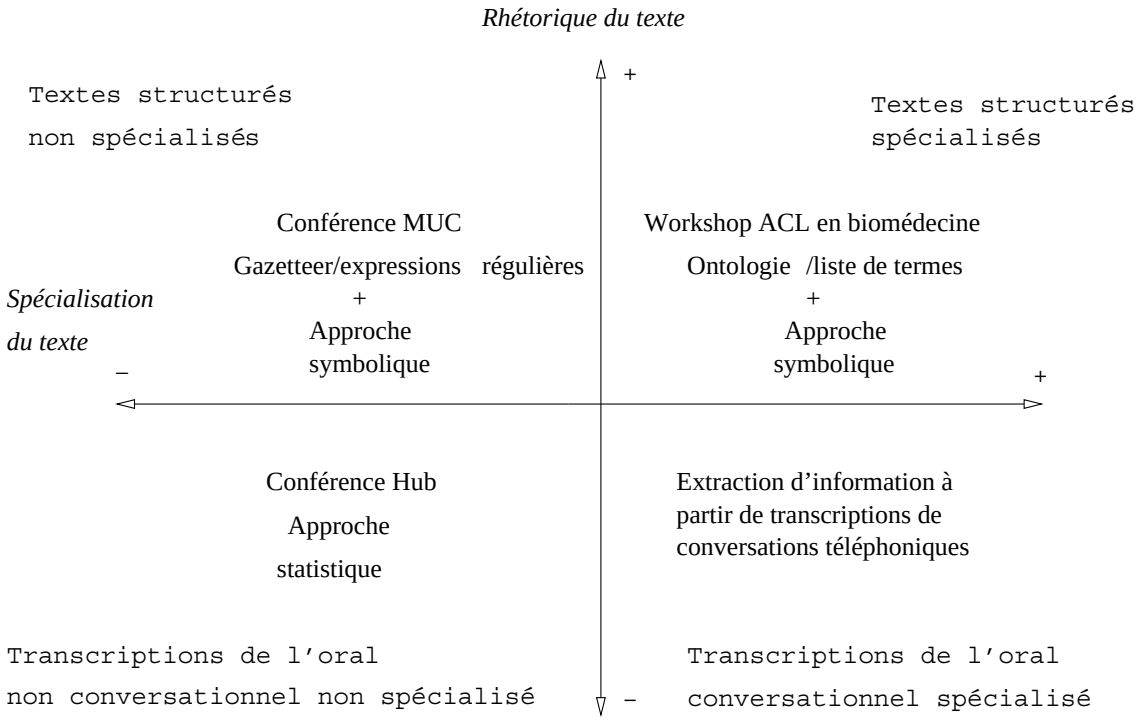


FIG. 3.1 – Types de textes étudiés en extraction d'information

de termes.

D'autre part, la rhétorique des textes analysés, qui va des textes écrits bien structurés tels que les dépêches journalistiques jusqu'aux transcriptions de l'oral peu structurées a une conséquence sur les techniques d'extraction des entités nommées et des faits individuels. Par exemple, lors des campagnes d'évaluations HUB-4 (1998) et HUB-5[4] (1999) centrées sur l'extraction d'entités nommées à partir de transcriptions de monologues de type bulletins de journal télévisé, les systèmes les plus performants ont utilisé des modèles de Markov pour gérer les irrégularités langagières [Bikel *et al.*, 1997; Kubula *et al.*, 1998].

La figure 3.1 fait la synthèse des textes étudiés en EI, des approches les plus utilisées ainsi que les conférences et *workshops* qui y ont été consacrés.

[4] `http://www.nist.gov/speech/tests/ie-er/er_99/er_99.htm`

## 3.3 Approche classique d'EI

Les dépêches journalistiques ont largement été étudiées durant les MUC. L'intérêt pour ces textes réside dans la disponibilité de corpora tels que le British National Corpus (BNC), mais aussi dans leurs caractéristiques linguistiques. D'une part, la majeure partie de l'information pertinente est véhiculée par des groupes nominaux facilement identifiables : les entités nommées. D'autre part, la structure des phrases suit des règles de grammaire définies qui garantissent la présence d'une relation **sujet-verbe-objet**. Cette dernière propriété assure la localité de l'information, ce qui facilite sa détection.

La majorité des systèmes d'EI procèdent selon trois étapes en cascade [Appelt *et al.*, 1995; Hobbs *et al.*, 1997] :

1. Extraction des entités nommées qui sont des groupes nominaux facilement identifiables et indépendants des événements recherchés, tels que les noms de lieux, d'organismes et de personnes.
2. Extraction des faits individuels en comparant la structure argumentale de la relation **sujet-verbe-objet** des phrases avec celle des patrons d'extraction. Une phrase est compatible avec un patron quand les verbes respectifs font partie d'une même classe de verbes prédéfinie et quand les arguments sont de même nature sémantique, par exemple deux entités nommées de même type.
3. Combinaison des bribes d'information obtenues aux étapes 1 et 2 grâce à une analyse du discours qui permet la résolution des coréférences et l'inférence de faits moyennant des connaissances du monde.

## 3.4 Extraction des entités nommées

Le but de cette étape est l'annotation de groupes nominaux qui renvoient à des noms de compagnies, de lieux ou de personnes (entités ENAMEX selon la terminologie MUC), en plus des expressions numériques telles que les dates (entités TIMEX), monnaies et pourcentages (entités NUMEX). Un trait sémantique distingue les différentes catégories des entités

ENAMEX : ORGANIZATION, LOCATION et PERSON. Les entités NUMEX se déclinent aussi en catégories : DATE, PERCENT et MONEY. Le choix des classes sémantiques dépend des informations recherchées.

L'exemple (1) est un extrait de texte tiré des comptes rendus de MUC-6, annoté avec les étiquettes ENAMEX et NUMEX.

(1) `<ENAMEX type="ORGANIZATION"> Bridgestone Sports Co. </ENAMEX> said <NUMEX type="DATE"> Friday </NUMEX> it had set up a joint venture in <ENAMEX type="LOCATION"> Taiwan </ENAMEX> with a local concern and a Japanese trading house to produce golf clubs to be shipped to <ENAMEX type= "LOCATION"> Japan </ENAMEX>.`

### 3.4.1 Approches d'extraction des entités nommées

L'extraction des entités nommées est la tâche la plus maîtrisée en EI avec un F-score qui rejoint celui obtenu par les annotateurs humains (autour de 96 % contre 98 % pour les humains [Chinchor et Dungca, 1995]). L'annotation des entités ENAMEX est compliquée par le fait qu'une entité nommée peut être un nom propre mais aussi une collocation de plusieurs noms communs tels que `national museum`.

Les travaux en extraction d'entités nommées se basent sur des connaissances du domaine qui peuvent être classées en deux catégories :

**Les listes, dictionnaires et ontologies** sont des sources de connaissances plus ou moins organisées qui regroupent les noms de lieux, d'organismes et parfois de personnes. Ces approches sont très utilisées car elles sont faciles à réaliser grâce à la disponibilité de gazettes. Les meilleurs résultats pour cette approche ont été rapportés par les systèmes NYU [Grishman, 1995] avec un F-score de 96,4 % et LOUELLA [Childs *et al.*, 1995] avec 90,8 %. Le principal inconvénient de cette technique est la nécessité de tenir à jour ces listes toujours incomplètes, tandis qu'elle a pour avantage de diminuer l'ambiguïté sémantique, par exemple lorsqu'il s'agit de déterminer si un nom propre est le nom

d'une personne ou celui d'un organisme.

**Les règles ou expressions régulières** combinent des unités lexicales telles que `Mr.`, `CEO` ou `chairman` et des informations morpho-syntaxiques. Parmi les meilleurs systèmes ayant utilisé cette approche lors de MUC-6, nous retrouvons le système FASTUS avec un F-score de 94,0 % [Appelt *et al.*, 1995] et de 93,6 % [Weischedel, 1995] pour le système BBN.

D'autres systèmes ont utilisé des approches d'apprentissage, notamment l'apprentissage symbolique se fait sur un corpus où les entités nommées sont annotées avec leur catégorie syntaxique et des indicateurs de la présence d'une majuscule en début de mot, en plus de certaines unités lexicales particulières telles que `CEO`. Les meilleurs résultats ont été obtenus par les systèmes SRA [Krupka, 1995] avec un F-score de 95,6 % et de 91,2 % pour le système ALEMBIC [Aberdeen *et al.*, 1996]. D'autres techniques d'apprentissage ont été expérimentées, notamment par le système NYMBLE qui utilise un modèle de Markov [Bikel *et al.*, 1997]. Le résultat obtenu sur le corpus de MUC-6 est un F-score de 93 %. L'intérêt des techniques d'apprentissage réside dans la capacité d'adapter des systèmes rapidement : moyennant une quantité suffisante de données d'entraînement, il est possible d'obtenir des résultats proches de ceux obtenus avec des règles codées manuellement. Cependant, le principal inconvénient de ces approches est le besoin de disposer d'une taille importante de données annotées.

### 3.4.2 Des entités nommées vers les classes sémantiques

Le choix d'une technique d'extraction des entités nommées doit tenir compte des particularités des textes conversationnels spécialisés :

- Les irrégularités langagières de l'oral augmentent le nombre des erreurs d'analyse morpho-syntaxique lesquelles pénalisent les approches basées sur les expressions régulières (section 3.4.1).
- Certains domaines spécialisés tels que la biomédecine ou la recherche et sauvetage maritime présentent des termes spécialisés qui sont des noms propres tels que `Aurora` ou `King Air`, mais aussi des combinaisons de noms communs comme

l'expression `Emergency Locator Transmittor`. En biomédecine, les approches utilisées se basent sur des listes de termes disponibles via des ressources lexicales tels que EcoCyc [Karp *et al.*, 2002]. Toutefois, des travaux récents en EI montrent l'utilité des ontologies pour ces genre d'applications[5] [Boufaden, 2003; Aitken, 2002].

## 3.5 Extraction des faits individuels

Le but de cette étape est d'extraire les expressions pertinentes au domaine explicitement citées et qui ne sont pas des entités nommées. Un fait individuel est extrait par le biais d'un patron d'extraction qui le localise et le lie à l'entité nommée impliquée.

L'approche classique de conception des patrons d'extraction tient compte de la structure syntaxique des phrases pertinentes collectées à partir du corpus d'analyse, des verbes et de la classe sémantique de leurs arguments, c'est-à-dire les types d'entités nommées. Le filtrage des phrases pertinentes se fait selon un processus d'appariement qui vérifie si une phrase peut être unifiée avec un patron. Le cas échéant, les variables sont instanciées et l'information pertinente (le fait individuel) est extraite.

La figure 3.2 illustre un texte, un patron et le formulaire rempli après l'extraction des faits individuels relatifs à une co-entreprise.

De manière générale, l'identification des patrons s'effectue en trois étapes :

1. La collecte de certains mots clé, généralement le nom des champs des formulaires et un ensemble représentatif de leurs réponses.
2. Le repérage des phrases contenant un ou plusieurs de ces mots clé tels qu'une entité nommée ou un verbe informatif. Généralement, cette étape de repérage permet d'identifier un ensemble de verbes pertinents pour la tâche d'extraction.
3. La construction de patrons en se basant sur la structure syntaxique des phrases repérées.

[5] Conférence EUROLAN2003 sur le thème *Ontologies and Information Extraction*. `http://ic2.epfl.ch/~pallotta/ontoIE/`

(2) Bridgestone Sports Co. (ORGANISATION) said Friday (DATE) it had set up a joint venture in Taiwan (LOCATION) with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan (LOCATION).

Unification

⇓

(3) ORGANISATION had set up a joint venture with ORGANISATION

Extraction

⇓

| *Tie_up* | |
|---|---|
| 1 Org-name | Bridgestone Sport Co. |
| 2 Org-descriptor | - |
| 3 JV Company | a local concern, a Japanese trading house |
| 4 Capitalization | - |

Fig. 3.2 – Exemple de patron d'extraction pour les textes structurés et résultat de l'extraction de l'événement *Tie_up*.

Les approches d'apprentissage développées jusqu'à présent ne permettent pas l'automatisation complète des étapes 2 et 3. Toutefois, nous distinguons quelques travaux en apprentissage quasi non supervisé ; c'est-à dire qui utilisent un minimum d'information, par exemple un corpus partagé en une partie composée de textes pertinents et une partie composée de textes non pertinents [Riloff, 1996]. Dans ce qui suit, nous donnons des exemples de systèmes pour chacune des principales approches d'apprentissage de patrons d'extraction : l'approche symbolique et l'approche statistique.

### 3.5.1 Apprentissage symbolique des patrons d'extraction

Plusieurs systèmes d'extraction utilisent une approche symbolique pour l'apprentissage des patrons d'extraction tels que LIEP [Huffman, 1996], PALKA [Kim et Moldovan, 1995], WHISK [Soderland *et al.*, 1995] et AUTOSLOG [Riloff, 1993].

Ce dernier est le premier système à utiliser une approche d'apprentissage pour la génération de patrons d'extraction et est l'un des systèmes les plus performants, produisant 98 % des patrons d'extraction codés manuellement lors de MUC-4. Il se base sur la spécialisation d'un schéma d'extraction général. Il utilise un corpus où les réponses des champs de formulaires sont annotées sémantiquement. Le résultat de l'apprentissage est un ensemble de patrons qui décrivent des relations syntaxiques **sujet-verbe** ou **verbe-objet** représentées par des structures (*concept nodes*) rattachant chaque patron à un champ de formulaire que l'auteur appelle CONCEPT.

La construction d'un noeud concept suit quatre étapes. Étant donné un groupe nominal *réponse*, le système :

1. Trouve la phrase dans laquelle apparaît le groupe nominal *réponse*.
2. Fait une analyse syntaxique de la phrase afin d'identifier les différents regroupements syntaxiques, en particulier les groupes nominaux sujet et complément d'objet et le groupe verbal. Cela va permettre de situer le groupe nominal *réponse* par rapport au groupe verbal (position de sujet ou d'objet).
3. Applique les schémas linguistiques, tels que le schéma (4), afin d'identifier les rôles

thématiques dépendants du domaine. Le premier schéma qui s'applique permet d'instancier le champ POSITION du noeud concept. Le <target-np> représente le groupe nominal *réponse*.

(4) <active-voice-verb> followed by <target-np>=<direct object>

4. Génère une structure CONCEPT NODE contenant les attributs CONCEPT, TRIGGER, POSITION, CONSTRAINTS et ENABLING CONDITIONS représentant respectivement le champ d'un formulaire *Scenario*, le mot qui déclenche le patron, la position syntaxique (objet ou sujet) du groupe nominal *réponse*, sa classe sémantique et les conditions syntaxiques (voix active ou passive) que doit remplir la phrase contenant la *réponse*.

Ce même procédé se répète pour l'ensemble des groupes nominaux clé présents dans le corpus d'entraînement.

La figure 3.3 illustre le processus d'apprentissage des patrons illustré sur une phrase du domaine des catastrophes naturelles. Ce patron extrait l'information répondant au champ DAMAGE qui décrit les dégâts matériels causés par la catastrophe naturelle ; dans notre exemple, il s'agit d'une tornade.

### 3.5.2 Apprentissage statistique des patrons d'extraction

Les premiers systèmes employant une approche statistique pour l'apprentissage des patrons d'extraction ont été développés vers la fin des années 90 [Freitag et McCallum, 1999; Seymore *et al.*, 1999; Leek, 1997]. L'idée générale de l'approche est d'entraîner un modèle de Markov formé d'états qui représentent les champs du formulaire. Le corpus d'entraînement contient les réponses des champs annotées.

Pour illustrer ce type d'approche, nous présentons un système d'EI développé par Seymore *et al.* [1999] dans le cadre du projet CORA [McCallum *et al.*, 1999] : un moteur de recherche pour les articles scientifiques. Le but de ce système est d'identifier des informations telles que le titre d'un article, l'auteur, l'affiliation, l'adresse et le résumé de l'article, et ce à partir d'entêtes d'articles. Ces informations servent à peupler une base de données qui sera utilisée

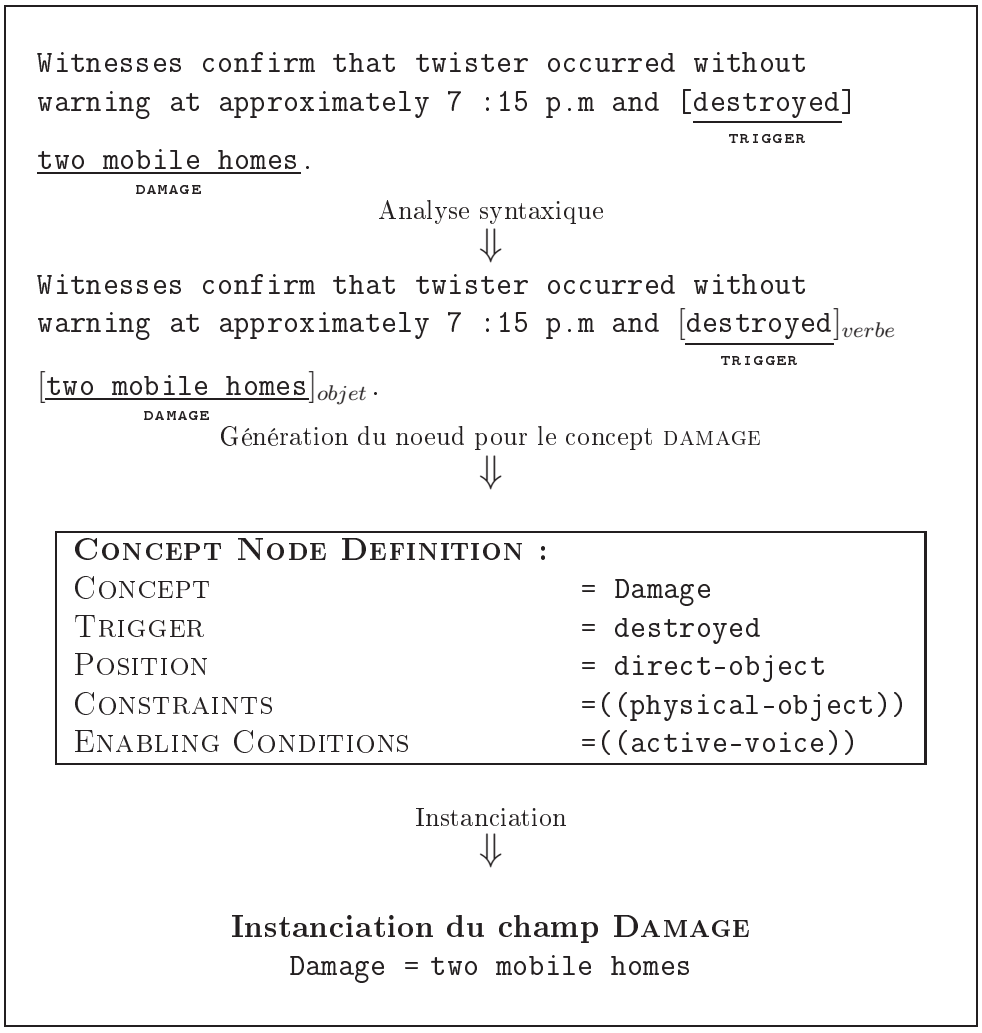


FIG. 3.3 – Patron généré à partir d'un exemple de phrase annotée sémantiquement et instanciation du patron pour l'extraction des dégâts matériels.

par le moteur de recherche CORA.

Les auteurs ont utilisé un modèle de Markov (figure 3.4) composé de 15 états, chacun représentant un champ de la base de données à peupler[6]. Le corpus d'entraînement est composé de 1000 entêtes d'articles dont les informations à extraire sont annotées par le champ correspondant, c'est-à-dire le titre (title), l'auteur (author), les adresses de courriel (email), les adresses http et civique (address), le numéro de publication (numpub), les remerciements

[6]Peupler une base de données revient à construire un formulaire *Scenario* qui contient tous les champs de la base de données.

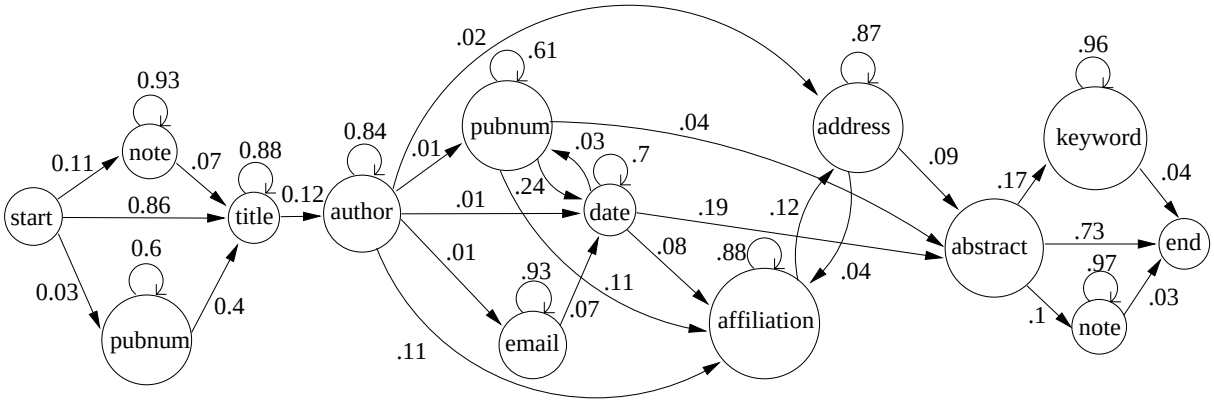


FIG. 3.4 – Modèle de Markov pour l'extraction d'information à partir d'entêtes d'articles.

(note), la date (date), l'affiliation (affiliation), les mots clé (keyword) et le résumé (abstract). Chaque état émet les mots qui remplissent le champ associé.

Les états sont reliés entre eux par des arcs avec une probabilité de transition égale à la fréquence relative avec laquelle deux champs se suivent dans un entête. Les résultats rapportés montrent un taux d'extraction correcte de 90,1 %.

### 3.5.3 Apprentissage de patrons à partir de textes conversationnels spécialisés

La plupart des systèmes reposant sur une approche d'apprentissage symbolique utilisent des connaissances générales, avec les règles de syntaxe. Dans le cas d'AUTOSLOG, ces règles définissent le noeud concept. Elles se basent sur la position des syntagmes nominaux clé comme le montre la règle linguistique (4).

La présence d'irrégularités langagières de l'oral modifie la structure syntaxique des énoncés, compliquant leur modélisation par des règles de manière exhaustive. Ainsi, les principaux obstacles à l'utilisation d'une approche symbolique sont les restrictions de contiguïté impo-

sées par les règles de grammaire.

Considérons l'énoncé (5) extrait de notre corpus : deux groupes nominaux, `an overdue vessel` et `missing boat on the south Town1`, se succèdent pour le même verbe `get`. Selon la règle linguistique (4), seul le premier groupe nominal est considéré lors de l'extraction ; pourtant, les deux groupes nominaux présentent des informations pertinentes.

(5) `We` $[get]_{verbe}$ $[$*an overdue vessel*$]_{gn}$`,` $[$*a missing boat*$]_{gn}$ `on the south east coast of Newfoundland .`

Par ailleurs, dans une évaluation de la difficulté de l'extraction d'information à partir de transcriptions de bulletins du journal télévisé, Grishman [1998] met l'emphase sur la robustesse des approches statistiques, dont les modèles de Markov, pour contrer le bruit introduit par les irrégularités de l'oral. Il propose un modèle de patrons statistiques qui se base sur deux éléments clés :

1. L'introduction d'états *joker* pour absorber les mots superflus tels que les répétitions apparaissant entre deux informations pertinentes. Ces états permettent de relaxer la contrainte de contiguïté des arguments d'une relation pertinente.
2. La gestion des cas d'omission à l'aide de transitions nulles. L'absence d'un argument ne doit pas entraîner l'échec de l'extraction des autres arguments présents dans l'énoncé.

Enfin, un avantage propre aux modèles statistiques est la possibilité d'obtenir une liste des réponses potentielles dans l'ordre décroissant de leur probabilité pour choisir une autre solution lorsque la première proposée ne satisfait pas certaines conditions, sémantiques par exemple.

## 3.6 Combinaison des bribes d'information pertinente

La combinaison des informations est une étape importante de l'EI et elle est la moins maîtrisée, comme en témoignent le F-score de 57 % lors de MUC-6 et de 51 % lors de MUC-7. C'est une étape qui fait partie de l'analyse du discours et qui se base sur les informations

extraites aux étapes précédentes. Deux méthodes sont utilisées à ce niveau : l'inférence et la résolution des coréférences.

La figure 3.5 illustre un exemple d'inférence qui utilise les différentes informations reliées à la co-entreprise "Joint venture". Dans la phrase (6), les partenaires de la co-entreprise sont identifiés grâce au patron (8), tandis que le patron (9) appliqué à la phrase (7) extrait le capital de la nouvelle société issue de la co-entreprise. L'inférence sur l'événement de co-entreprise permet de déduire que la compagnie `Bridgestone Sports Taiwan Co.` est issue de la co-entreprise impliquant `Bridgestone Sports Co.`, `a local concern` et `a Japanese trading house`, et de rattacher toutes les informations reliées à l'opération de co-entreprise.

La deuxième méthode, la résolution des coréférences, est la plus importante pour combiner les informations. Les meilleurs F-scores rapportés lors de MUC-6 et MUC-7 sont respectivement de 71 % et 62 %. La figure 3.6 illustre la résolution du lien de coréférence entre une anaphore[7] et son antécédent[8]. Les antécédents peuvent être des entités nommées ou des événements. Dans cet exemple, les relations sont établies entre les anaphores `Mrs. Fribble` et `Mr. Morton` et leurs antécédents respectifs, `Edna Fribble` et `Sam Morton`.

La majeure partie de la tâche de résolution des coréférences est consacrée au rattachement de noms et d'entités nommées, comme illustré à la figure 3.6. Les autres anaphores sont le pronoms `it` ainsi que les pronoms personnels.

Une des approches de résolution des coréférences les plus performantes a été proposée par Kameyema en atteignant un rappel de 59 % et une précision de 72 % lors de l'évaluation MUC-6 [Kameyama, 1997]. Elle se base sur une analyse syntaxique partielle du texte pour résoudre les liens de coréférence des pronoms et groupes nominaux définis et indéfinis. La sélection de l'antécédent s'effectue selon trois critères de préférence :

**L'espace de recherche** Il délimite une région du texte pour la collecte des antécédents. Cette région est de dix phrases pour les coréférences de type groupe nominal et de trois phrases pour les pronoms.

[7] Une anaphore est une expression (par exemple un pronom personnel ou un groupe nominal défini) qui fait référence à un élément lexical du texte.

[8] Un antécédent est l'élément lexical qui renvoie à un objet du monde réel.

**La compatibilité sémantique** Parmi les antécédents collectés à la première étape, seuls ceux qui sont compatibles sémantiquement avec le groupe nominal analysé sont retenus, c'est-à-dire que l'anaphore doit être égale à ou contenir l'antécédent. Par exemple, l'anaphore `cette maison` est un groupe nominal défini et peut avoir comme antécédent `la maison rouge`, mais pas l'inverse.

**La proéminence syntaxique** Elle donne la préséance aux antécédents inter-phrastiques sur les intra-phrastiques et selon leur ordre d'apparition. L'antécédent le plus récent dans un parcours de gauche à droite est favorisé par rapport au reste des candidats potentiels.

L'intérêt de cet algorithme est qu'il ne nécessite pas une analyse syntaxique complète comme c'est le cas des approches standard de résolution des coréférences. Ce choix théorique est un atout pour la résolution des coréférences des textes conversationnels, ceux-ci étant des textes bruités. Par ailleurs, la délimitation d'une région du texte pour la collecte des antécédents peut être appliquée aux textes conversationnels, par exemple en associant cette région à une unité thématique.

## 3.7 Synthèse : notre approche d'EI pour les textes conversationnels spécialisés

L'approche que nous proposons pour l'extraction de faits individuels et l'apprentissage de patrons d'extraction à partir de textes conversationnels spécialisés se base sur les conclusions tirées de l'analyse des caractéristiques de notre corpus (section 2.5) et l'évaluation de l'adéquation des techniques d'extraction des entités nommées (section 3.4.2) et d'apprentissage automatique des patrons d'extraction (section 3.5.3) pour les textes conversationnels spécialisés. Le tableau 3.4 récapitule ces conclusions. Ainsi, nous retenons trois notions clé :

1. La conception d'une ontologie du domaine de la recherche et sauvetage maritime pour rassembler les termes du domaine en classes sémantiques.
2. La segmentation des textes conversationnels en paires d'adjacence, qui sont des unités

| Synthèse | Analyse du corpus | Approches d'EI |
|---|---|---|
| Extraction des entités | • le vocabulaire est composé de termes spécialisés qui sont des noms propres, des noms communs, des adjectifs ou des verbes.<br>• le vocabulaire contient des variantes terminologiques qui sont des expressions sémantiquement similaires aux termes du vocabulaire. | • les approches d'extraction de termes utilisent un lexique du domaine ou une ontologie.<br>• l'extraction des termes est davantage une étape d'étiquetage sémantique.<br>• l'étiquetage sémantique des expressions sémantiquement similaires au vocabulaire du domaine permet une meilleure extraction des termes du domaine. |
| Apprentissage des patrons d'extraction | • l'unité linguistique qui tient compte du caractère interactif des textes conversationnels est la paire d'adjacence.<br>• les irrégularités langagières rendent l'analyse syntaxique complète d'un énoncé difficile et compliquent l'identification d'une relation sujet-verbe-objet.<br>• le phénomène de pronominalisation du thème est important dans les textes conversationnels. Cela nuit à l'étape d'apprentissage puisque moins de relations pertinentes sont détectables. | • pour pallier la sous-représentation de relations, il est d'usage d'utiliser des classes de mots pour définir les patrons d'extraction.<br>• une approche d'apprentissage basée sur les modèles de Markov permet de gérer les bruits introduits par les irrégularités langagières de l'oral.<br>• la résolution des pronoms en position de sujet avant l'apprentissage des patrons et l'extraction de faits permettent une meilleure couverture des relations pertinentes. |

TAB. 3.4 – Tableau récapitulatif des conclusions tirées après l'analyse des caractéristiques de notre corpus et de l'évaluation des techniques d'extraction des entités nommées et d'apprentissage des patrons pour les textes conversationnels spécialisés.

sémantiquement maximales pour l'apprentissage des patrons d'extraction, et en unités thématiques pour la résolution des coréférences résultant de la pronominalisation du thème.

3. L'apprentissage de relations prédicat-arguments modélisées par des modèles de Markov impliquant des classes de mots représentant les termes du domaine de la recherche et sauvetage maritime.

### 3.7.1 Étapes de notre approche

Nous utilisons deux sources de connaissances *a priori* : l'ontologie du domaine de la recherche et sauvetage et un dictionnaire thésaurus[9] qui fournit les connaissances du monde. L'ontologie joue un rôle fondamental dans notre approche car elle contient les classes de termes qui constituent les composantes de nos patrons d'extraction.

Les étapes de notre approche sont :

**La segmentation linguistique des conversations** Cette étape permet d'isoler les paires d'adjacence pour tenir compte des dépendances entre certains tours de parole. Le but de cette segmentation est de fournir au module d'apprentissage de patrons des unités sémantiquement maximales.

**La segmentation thématique** Elle partage une conversation en une suite de segments où chaque segment porte sur un même thème. Ce découpage est utilisé pour la détection des thèmes et facilitent la résolution des anaphores pronominales.

**La détection des thèmes** Elle est utilisée pour l'étiquetage des expressions sémantiquement similaires au vocabulaire du domaine et pour la définition de valeurs par défaut pour la résolution des anaphores pronominales.

**L'étiquetage sémantique robuste** des termes du domaine. Cette étape est plus générale que l'extraction des entités nommées à cause de la variété des termes à couvrir et des expressions sémantiquement similaires au vocabulaire du domaine.

[9]Un dictionnaire thésaurus est un dictionnaire qui intègre des informations sémantiques issuent d'un thésaurus.

**La résolution des anaphores pronominales** en position de sujet. Cette étape a pour but de faire émerger les relations **prédicat-arguments** masquées par la pronominalisation du thème.

**L'apprentissage des patrons d'extraction** à partir des unités linguistiques annotées sémantiquement.

### 3.7.2 Architecture de notre système d'apprentissage de patrons d'extraction

La figure 3.7 reprend les étapes de notre approche. Nous distinguons les principales étapes décrites pour l'apprentissage des patrons et l'extraction des faits individuels. Les étapes 1 à 5 sont les mêmes pour l'extraction des faits individuels et pour l'apprentissage des patrons.

Les principaux points qui différencient notre approche de celle préconisée pour les textes structurés non spécialisés (section 3.3) est :

1. Le remplacement de l'étape d'extraction des entités nommées par un étiquetage sémantique pour l'extraction des termes du domaine ainsi que de leurs variantes.
2. Le déplacement de l'étape de résolution des coréférences impliquant les termes du domaine avant l'étape d'extraction des faits et l'apprentissage des patrons.

## 3.8 Conclusion

Dans ce chapitre, nous avons présenté un bref état de l'art de la technologie de l'EI. En particulier, nous avons mis l'emphase sur les techniques les plus performantes pour l'extraction des entités nommées, l'apprentissage des patrons d'extraction et la résolution des coréférences. Ensuite, nous avons présenté une synthèse des observations tirées de l'analyse des caractéristiques de notre corpus et de l'évaluation de la adéquation des techniques d'extraction des entités nommées et de l'extraction des faits individuels pour les textes conversationnels spécialisés. Cette synthèse nous a permis de proposer une approche d'EI adaptée

à ces textes. Dans les prochains chapitres, nous explicitons chacun des modules de notre architecture (figure 3.7). Le chapitre 4 présente les approches de segmentation des textes conversationnels ainsi que le module de détection de thèmes.

Le chapitre 5 décrit la méthode utilisée pour la conception de l'ontologie du domaine et le chapitre 6 la modélisation de l'étiqueteur sémantique robuste.

Le chapitre 7 décrit la conception du module de résolution des anaphores pronominales, le modèle utilisé pour l'apprentissage des patrons et les résultats de ces deux modules obtenus avec notre corpus.

(6) Bridgestone Sports Co. (ORGANISATION) Friday (DATE) had set up a joint venture in Taiwan (LOCATION) with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan (LOCATION).

(7) The joint venture, Bridgestone Sports Taiwan Co. (ORGANISATION), capitalized at 20 million new Taiwan dollars (MONEY), will start production in January 1990 (DATE).

(8) ORGANISATION (ORG-NAME) had set up joint venture in LOCATION with ORGANISATION (JV COMPANY)

(9) ORGANISATION capitalized at money

| *Tie_up* | |
|---|---|
| 1 ORGA-NAME | Bridgestone Sport Co. |
| 2 ORG-DESCRIPTOR | a local concern, a Japenese trading house |
| 3 JV COMPANY | Bridgestone Sports taiwan Co. |
| 4 CAPITALIZATION | 20000000 TWD |

FIG. 3.5 – Exemple de combinaison de bribes d'information par inférence sur le nom de la compagnie "Bridgestone Sport Co."

(10) Edna Fribble and Sam Morton addressed the meeting yesterday. Ms. Fribble discussed coreference and Mr Morton discussed unnamed entities.

(11) <coref id="1"> Edna Fribble </coref> and <coref id="2"> Sam Morton </coref> addressed the meeting yesterday. <coref id="3" ref="1" type="ident" min="Fribble"> Ms.Fribble </coref> discussed coreference, and <coref id="4" ref="2" type="ident" min="Morton"> Mr. Morton </coref> discussed unnamed entities.

FIG. 3.6 – Exemple de combinaison de bribes d'information après résolution de la coréférence.

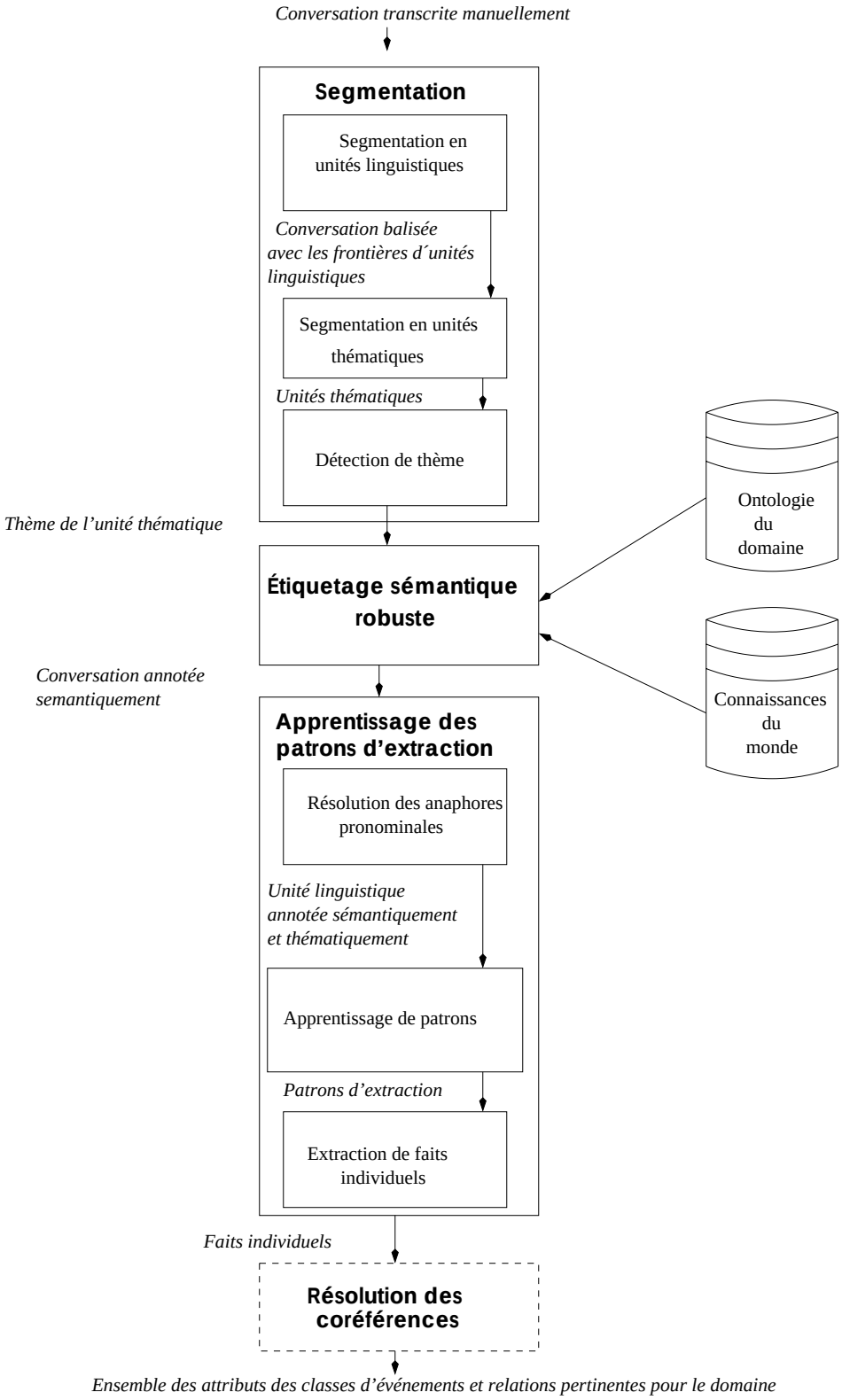


FIG. 3.7 – Étapes de notre approche d'EI à partir de textes conversationnels.

# Chapitre 4

# Segmentation des textes conversationnels

## 4.1 Introduction

La fragmentation de l'information est une des principales difficultés à gérer en EI à partir des textes conversationnels. Les paires question-réponse modifient la structure de l'information et le caractère spontané des conversations diminue la densité d'information dans un énoncé et augmente leur nombre. L'information est communiquée sur plusieurs énoncés ancrés au thème par le biais d'anaphores pronominales.

Ce chapitre présente deux approches de segmentation, linguistique et thématique, pour remédier à la fragmentation de l'information et faciliter le repérage de relations **prédicat-arguments**. La segmentation linguistique détecte les paires d'adjacence telles que les paires question-réponse qui nécessitent un traitement particulier. L'approche se base sur la présence de traits lexicaux discriminants modélisés par un modèle de Markov.

La segmentation thématique découpe les textes conversationnels en unités thématiques. L'approche que nous proposons a été présentée lors des conférences TALN [Boufaden *et al.*, 2004a] et NLPRS [Boufaden *et al.*, 2004b]. Elle se base sur la modélisation de marques lexicales, syntaxiques et discursives pour détecter les changements de thème. Nous présen-

tons également un module pour l'identification des thèmes, étape nécessaire pour l'étiquetage sémantique robuste et l'apprentissage de schémas d'extraction.

## 4.2 Segmentation en unités linguistiques

Le découpage linguistique des textes conversationnels a l'avantage de fournir des unités présentant des caractéristiques syntaxiques et sémantiques nécessaires aux applications de TAL. Dans le contexte de l'apprentissage de patrons, cette caractéristique est particulièrement importante puisque les patrons reposent sur la relation **sujet-verbe-objet** ou **prédicat-arguments**. Le découpage linguistique est essentiellement associé à la segmentation en paires d'adjacence [Stolcke et Shriberg, 1996], mais cela peut être limité à un tour de parole lorsque le style est narratif.

L'analyse de notre corpus (section 2.3.2) montre que les locuteurs n'observent pas toujours un style formel. Souvent, ils se connaissent et utilisent un style de discours qui varie entre narratif et conversationnel. En effet, 69 % des énoncés font partie d'échanges tandis que 31 % sont des contributions faites dans un style narratif. Le tableau 4.1 montre le type de segmentation linguistique que nous voulons effectuer de manière automatique. La conversation Overdue boat est segmentée manuellement par les frontières d'unités linguistiques représentées par les pointillés.

### 4.2.1 Définition d'une unité linguistique

Nous proposons une unité linguistique dont le contenu dépend du style de discours. Lorsque le style narratif est utilisé, l'unité linguistique correspond à un énoncé. Dans le cas d'un style conversationnel, une unité linguistique correspond à une paire d'adjacence.

Ainsi, le but du découpage en unités linguistiques est de distinguer les énoncés qui font partie des paires d'adjacence (style conversationnel) de ceux qui ne le sont pas (style narratif). Ce découpage permet notamment d'isoler les paires question-réponse qui nécessitent un traitement particulier pour la tâche d'EI. Dans cette thèse, nous soulevons le problème des

paires question-réponse, mais laissons l'étude d'un traitement pour reconstituer la structure argumentale de la paire pour des travaux futurs.

### 4.2.2 Marques utilisées pour la segmentation linguistique

Le tableau 4.1 illustre un découpage en unités linguistiques déterminé en fonction du style du discours. La première colonne du tableau correspond aux énoncés de la conversation Overdue boat et les colonnes 2 à 4 aux traits[1] utilisés pour la segmentation :

- Les conjonctions de coordination et les unités lexicales (colonne 2), telles que `and` et `well` en début d'énoncé, sont des particules de cohésion grâce auxquelles un locuteur ajoute de l'information [Halliday et Hassan, 1976]. Elles indiquent le début d'une unité linguistique. Les unités lexicales d'acquiescement, telles que `ok` et `yeah`, sont des marques lexicales qui maintiennent les échanges entre locuteurs [Maynard et Zimmerman, 1984]. Elles surviennent généralement à la fin d'une contribution ou complètent un segment thématique. Par exemple, 49,3 % des paires d'adjacence se terminent par une marque d'acquiescement et le point d'interrogation apparaît dans 66,3 % des paires d'adjacence.
- La longueur de l'énoncé (colonne 3) distingue les énoncés minimaux, tels que les formes d'acquiescement qui terminent les paires d'adjacence, de ceux qui contiennent un contenu propositionnel.
- Le locuteur (colonne 4) est utilisé pour distinguer le style narratif du style conversationnel. La présence de deux énoncés successifs générés par un même locuteur caractérise 62 % des énoncés narratifs du corpus.

Ces traits ne sont pas toujours discriminants, notamment dans le cas des marques d'acquiescement qui apparaissent au début d'un énoncé sans que ce dernier fasse partie d'une paire d'adjacence (énoncés 2-3, tableau 4.1). Selon les marques de l'énoncé précédent (en particulier le locuteur), il est possible de trancher sur la frontière du segment linguistique.

[1] La liste complète des traits utilisés pour la segmentation linguistique est présentée à l'annexe B.

| No | Loc | Énoncé | L | Lg | Lc |
|---|---|---|---|---|---|
| 1 | a | : Maritime operation centre, (INAUDIBLE) hello. | | l | a |
| | | ..... | .... | .. | .. |
| 2 | b | : hi, Mr. Wellington, it's captain Mr. VanHorn | hi | l | b |
| 3 | a | : yes. | yes | s | a |
| | | ..... | .... | .. | .. |
| 4 | b | : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the South Coast of Newfoundland, just in the area quite between Fortune Bay and Trepassey. | | l | b |
| | | ..... | .... | .. | .. |
| 5 | b | : it's on the south east coast of Newfoundland. | | l | b |
| | | ..... | .... | .. | .. |
| 6 | b | : this is been going on for, for 24 hours that the case has, or almost anyway, and we had an DFO King Air up flying this morning | | l | b |
| | | ..... | .... | .. | .. |
| 7 | b | : they did a radar search for us in that area. | | l | b |
| 8 | a | : yes. | yes | s | a |
| | | ..... | .... | .. | .. |
| 9 | b | : and their search turned up nothing. | and | l | b |
| 10 | a | : yeah. | yeah | s | a |
| | | ..... | .... | .. | .. |
| 14 | b | : so I'm wondering about the possibility of attempting it with a different platform perhaps someone with even other sensors other than the radar and, in fact, someone with a, with a radar that'll be a little more sensitive. | so | l | b |
| | | ..... | .... | .. | .. |
| 15 | b | : before I used the Challenger, I 'll use a Hurk. | | l | b |
| | | ..... | .... | .. | .. |
| 20 | a | : ...do you want this thing fired up now or you wanna wait till the Big Boys come in to work tomorrow morning? | ? | l | a |
| 21 | b | : well, I would like it if possible, I'd like them to, to be airborne at first light. | well | l | b |

TAB. 4.1 – Conversation Overdue boat de laquelle soient extraites des marques L (Lexicale), Lg (Longueur de l'énoncé) et Lc (Locuteur) pour la segmentation linguistique. Les pointillés représentent les frontières des unités linguistiques disposées manuellement.

Nous proposons d'utiliser des marques lexicales et syntaxiques pour prédire les frontières des unités linguistiques. Pour ce faire, nous utilisons un modèle de Markov qui permet de modéliser les dépendances temporelles observées entre les énoncés. Le choix des modèles de Markov est motivé par l'existence d'une relation temporelle entre les parties d'une paire d'adjacence : une question engendre une réponse, une affirmation engendre un acquiescement, une salutation engendre une salutation.

Dans la majorité des cas, c'est la deuxième partie de la paire qui est discriminante, à l'exception des cas de question-réponse où le point d'interrogation est un indicateur du début d'une paire question-réponse.

Ainsi, la segmentation en unités linguistiques revient à un problème de classification des énoncés en deux classes : la classe des énoncés constituant la deuxième partie des paires d'adjacence et qui, par conséquent, délimitent les paires (ADJ-P), et la classe des autres énoncés comprend ceux qui ne font pas partie d'une paire d'adjacence et ceux qui constituent la première partie d'une paire (NO-ADJ-P).

### 4.2.3 Modèle de langue

Nous supposons qu'entre deux énoncés il y a une frontière qui indique la présence d'une paire d'adjacence ADJ-P ou non NO-ADJ-P. Conséquemment, nous utilisons un modèle de Markov composé de deux états ADJ-P et NO-ADJ-P représentant deux classes d'énoncés :

- Les énoncés qui constituent la première partie d'une paire d'adjacence et ceux qui sont des énoncés individuels faisant partie d'un discours narratif. Ces énoncés ne présentent pas beaucoup de marques lexicales car ils sont en général des énoncés avec un contenu propositionnel.
- Les énoncés qui forment la deuxième partie d'une paire d'adjacence. Ces énoncés sont riches en marques lexicales d'acquièscement, par exemple tels que `ok`, `right` ou `well`.

La segmentation linguistique optimale est déterminée par l'équation 4.1 :

$$\hat{q} \quad = \quad \underset{q}{\operatorname{argmax}} \prod_{t=1}^{n} P(q^{t-1}|q^t) P(o^t|q^t) \tag{4.1}$$

$\hat{q}$ est la séquence optimale d'états qui reflète la meilleure segmentation linguistique d'une séquence de longueur $n$, $o^t$ est la $t^{\text{eme}}$ observation de la séquence d'observations de traits et les $q^i$ sont les états du modèle.

Les probabilités d'émission des observations sont les fréquences relatives des traits observés sur le corpus d'étude pour chaque état.

$$P(o^i|q^j) = \frac{\#(o^i, q^j)}{\sum_{q^k \in \{\text{ADJ-P,NO-ADJ-P}\}} \#(o^i, q^k)}$$

$o^i$ est la combinaison des marques syntaxiques et sémantiques extraites d'un énoncé $u$, $\#(o^i, q^j)$ le compte de l'observation $o^i$ émise par l'état $q^j$. Les probabilités de transition entre les états NO-ADJ-P et ADJ-P sont les fréquences relatives des transitions d'un état à un autre.

$$P(q^i|q^j) = \frac{\#(q^j, q^i)}{\sum_{q^k \in \{\text{ADJ-P,NO-ADJ-P}\}} \#(q^k, q^i)}$$

$q^i$ est un des états du modèle de Markov et $\#(q^i, q^j)$ représente le compte des transitions de l'état $q^j$ à $q^i$. Tous les comptes sont calculés sur le corpus d'entraînement.

La figure 4.1 illustre le modèle de Markov intégrant ces probabilités où $D$ et $F$ sont les états initial et final. Notons que l'état ADJ-P boucle sur lui-même car plusieurs énoncés composés uniquement de marques d'acquiescement peuvent se succéder pour marquer la clôture d'une unité thématique (section 4.3.3).

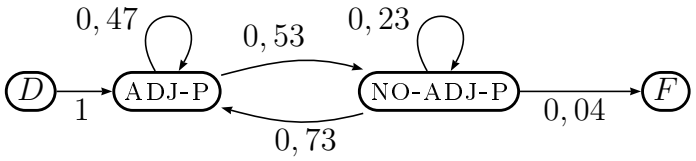


FIG. 4.1 – Modèle de Markov d'ordre 1 pour la segmentation en unités linguistiques.

| | Modèle de Markov | Arbre de décision | *Baseline* |
|---|---|---|---|
| Erreurs | 15,9 % | 21,7 % | 50,0 % |
| Précision | 89,5 % | 90,2 % | – |
| Rappel | 79,4 % | 73,3 % | – |
| F-score | 84,1 % | 80,9 % | – |

TAB. 4.2 – Résultats de la segmentation en unités linguistiques. Comparaison entre notre modèle, un arbre de décision et un *baseline*. La précision et le rappel sont calculés pour les frontières des paires d'adjacence.

### 4.2.4 Expériences et résultats

Afin d'illustrer la pertinence du choix d'un modèle de Markov pour la segmentation linguistique, nous avons effectué deux expériences. Dans la première, nous utilisons le modèle de Markov[2] d'ordre 1 décrit à la figure 4.1 et dans la deuxième, un arbre de décision généré par l'algorithme ID3 [Quinlan, 1986] (section 4.4) qui n'exploite pas la dépendance des énoncés d'une paire d'adjacence. Les deux modèles sont comparés à un *baseline* où la probabilité d'observer une paire d'adjacence est d'une chance sur deux (état ADJ-P ou état NO-ADJ-P).

Le corpus utilisé est composé de 64 conversations totalisant 3481 énoncés. Les traits décrits à la section 4.2.2 ont été extraits automatiquement. L'évaluation du modèle a été effectuée par 10 validations croisées[3] avec 80 % du corpus réservé à l'entraînement. Les résultats pour les deux modèles sont les taux d'erreur de classification rapportés au tableau 4.2. La précision et le rappel sont calculés pour les frontières des paires d'adjacence qui correspondent à l'état ADJ-P de la figure 4.1.

Le modèle de Markov donne un meilleur F-score que l'arbre de décision. En particulier, il reconnaît plus de frontières de paires d'adjacence avec un rappel de 79,4 %, comparativement à 73,3 % pour l'arbre de décision. Ce résultat confirme la nécessité de tenir compte du

[2]Le modèle de Markov utilisé dans nos expérimentations a été fourni par Philippe Langlais du laboratoire RALI. Dans cette implémentation, aucun lissage n'a été fait.

[3]La validation croisée consiste à (1) diviser les données d'apprentissage en $N$ (10 dans notre cas) échantillons de tailles égales ; (2) retenir l'un de ces échantillons pour le test et apprendre sur les $N-1$ autres ; (3) mesurer le taux d'erreur empirique pour l'échantillon test choisi ; (4) recommencer les étapes (2) et (3) $N-1$ fois en changeant à chaque fois l'échantillon de test. L'erreur estimée finale est la moyenne arithmétique des erreurs calculées sur chaque échantillon de test.

contexte d'énonciation lors de la détection des paires d'adjacence.

L'analyse des erreurs de classification de notre modèle montre que pour 47,2 % des erreurs de détection de paires d'adjacence, l'énoncé constituant la frontière d'une paire d'adjacence ne présente pas de marque lexicale. Dans ces cas, l'information prosodique permettrait de combler le manque d'information lexicale.

Un travail en segmentation linguistique de textes conversationnels similaire au nôtre a été effectué sur le corpus SWITCHBOARD par les chercheurs du groupe SRI [Stolcke et Shriberg, 1996; Stolcke, 1997]. Stolcke et al. utilisent les marques de début des tours de parole, la catégorie syntaxique des mots et d'autres marques lexicales telles que `okay`, `so` ou `well` pour entraîner un modèle de Markov. La présence d'une frontière d'unité linguistique est décidée sur la base de la fréquence des trigrammes observés. Les résultats obtenus sont de 85,2 % pour le rappel et 69,2 % pour la précision. À la différence de cette approche, nous n'utilisons pas la catégorie syntaxique des mots car celle-ci peut être erronée à cause des irrégularités langagières de l'oral. Dans notre corpus, le taux d'erreurs d'analyse morphosyntaxique est de l'ordre de 7 %[4].

## 4.3 Segmentation en unités thématiques

Nous proposons la segmentation des conversations en unités thématiques dans le but de faciliter la résolution des anaphores pronominales en position de sujet. Cette étape est pertinente à cause du taux important de pronominalisation des thèmes (section 2.4.2) et parce qu'il est plus facile de localiser l'information pertinente lorsque les arguments des relations **prédicat-arguments** sont identifiés.

La segmentation par thème a surtout été étudiée pour les textes écrits tels que les articles scientifiques. Différentes approches ont été proposées, Morris et Hirst [1991] et Hearst [1994] utilisent les répétitions de mots comme marque de cohésion dans des contextes adjacents. L'approche décrite par Hearst a été évaluée sur des articles scientifiques avec le système TEXTTILING. Ce dernier détecte les sous-thèmes avec une précision de 66 % et un rappel de

[4]Ce taux est obtenu avec l'étiqueteur de Brill entraîné sur le corpus Brown (LDC).

61 %. Youmans [1990] suggère que la première utilisation d'un nouveau mot dans un document indique un changement de thème. Enfin, Reynar [1999] propose une méthode qui utilise notamment la répétition des entités nommées. Plus récemment, des travaux en segmentation ont été consacrés à la segmentation de transcriptions automatiques de l'oral non conversationnel. En particulier, les campagnes d'évaluation TDT (Topic Detection and Tracking)[5] consacrées à la détection de thèmes à partir de transcriptions du bulletin du journal télévisé ont permis le développement d'algorithmes utilisant l'information prosodique tels que les changements de ton [Hirschberg et Nakatani, 1996] ou la durée des pauses [Passonneau et Litman, 1993].

La problématique de la segmentation en unités thématiques de textes conversationnels est différente de celles étudiées jusqu'à présent pour deux raisons :

1. Nos textes sont bruités à cause des irrégularités langagières de l'oral. Une approche basée sur les répétitions de mots pourrait souffrir d'irrégularités telles que les répétitions. Par ailleurs, la composante interactive présente dans nos conversations introduit une autre dynamique dans les changements de thème.
2. Le corpus n'est pas annoté prosodiquement, ce qui rend l'exploitation des durées de pause ou des changements de ton impossible.

Ainsi, l'étude que nous proposons repose sur les marques de cohésion utilisées en segmentation de textes, mais aussi sur les résultats d'études en analyse des conversations.

### 4.3.1 Définition d'une unité thématique

> *Le thème d'un discours, d'un fragment de discours ou d'une phrase est un syntagme nominal qui exprime de quoi il s'agit dans le discours, le fragment de discours ou la phrase. De manière générale, le thème peut être un objet, un événement, une idée ou une activité.* [Renkema, 1993]

Une unité thématique se caractérise par un thème unique autour duquel s'articule des énoncés qui contiennent des commentaires sur ce thème. Dans notre contexte, un thème

[5] `http://www.nist.gov/speech/tests/tdt`

correspond à un événement. Il s'agit de grouper les unités linguistiques adjacentes qui font référence à un même événement. Les événements considérés dans notre problématique sont définis par l'information recherchée et sont déduits des formulaires d'extraction. Les thèmes considérés font référence à :

- L'objet en difficulté (*Missing object*), c'est-à-dire sa description, le nom de son propriétaire.
- L'incident (*Incident*), c'est-à-dire le type d'incident, la cause, le type d'appel de détresse.
- L'unité de recherche (*Search-unit*), c'est-à-dire le nom de l'unité, la ou les ressources qu'elle utilise.
- La mission (*Mission*), c'est-à-dire le lieu de la mission, les conditions météorologiques, la date.
- La coordination (*Controller*), c'est-à-dire les organismes intervenant dans les missions et les personnes responsables de la mission.

Un autre type de thème a été ajouté (*Other*) pour réunir les sujets non pertinents à la tâche d'extraction.

Le tableau 4.3 montre un extrait de conversation découpé en unités thématiques. Afin d'évaluer la clarté du concept thème tel que nous l'avons défini, deux annotateurs différents ont annoté 12 conversations afin de calculer le coefficient kappa [Cohen, 1960] qui est un indice d'accord entre les annotateurs. Lorsque le score obtenu est supérieur à 0,8, cela indique que les annotateurs s'entendent sur la segmentation, signifiant que la tâche est bien définie. Dans notre cas nous avons obtenu un score de 0,84.

Nous traitons la segmentation en unités thématiques en étudiant deux aspects complémentaires de cette problématique, à savoir :

1. La cohésion des énoncés dans une unité thématique.
2. Les changements de thème.

### 4.3.2 Éléments caractéristiques de la cohésion

La cohésion est un critère important en analyse de discours pour qualifier un ensemble de phrases comme constituant un texte. Elle fait référence au lien qui existe entre les éléments d'un texte [Renkema, 1993]. Halliday et Hassan [Halliday et Hassan, 1976] identifient plusieurs marques de cohésion que l'on retrouve dans une unité thématique de textes conversationnels. Parmi celles-ci, nous retenons les marques lexicales facilement identifiables suivantes :

1. La conjonction qui est une particule liant un énoncé aux énoncés précédents ou suivants.
2. La répétition qui est la reproduction d'un mot ou d'une partie de l'énoncé précédent dans l'énoncé courant et qui ancre un énoncé à une unité thématique.

### 4.3.3 Caractéristiques des changements de thème

Sacks *et al.* [1974] et Maynard et Zimmerman [1984], parmi d'autres, ont décrit les règles qui régissent les changements de thème dans les dialogues. Ils soulignent l'importance du rôle du locuteur pendant le processus de développement d'une unité thématique.

> *Often one person at a time is given the responsability for developping a topic. Nevertheless, topical talk is a collaborative phenomenon in that while one person does topic developmental utterances, the other may produce questions, invitations, continuers, and so forth, to keep the line of talk going.* [Maynard et Zimmerman, 1984]

Dans une conversation entre deux locuteurs, chaque locuteur montre son intérêt et sa compréhension de ce qui est communiqué grâce à des réponses typiques telles que `ok`, `yeah` et `right`. En fonction du rôle du locuteur dans le processus développemental du thème, ces réponses peuvent être perçues comme un incitateur à continuer le thème ou, au contraire, comme un inhibiteur dans le but d'interrompre le thème.

Selon le rôle du locuteur dans le développement du thème, certaines conditions prédisent un changement de thème, dont celles que nous exploitons dans notre modèle :

1. La réponse à une question est brève, comme dans le cas d'un acquiescement (`yeah`, `right`, `ok`).
2. L'interlocuteur répond à la question d'un locuteur par une autre question.
3. L'introduction d'un nouvel objet qui n'est pas celui du thème développé.

### 4.3.4 Marques utilisées pour la segmentation par thème

Le tableau 4.3 reprend la conversation Overdue boat à laquelle sont ajoutées les frontières des unités thématiques et les traits[6] utilisés pour la segmentation. Les frontières de thèmes coïncident avec des frontières d'unités linguistiques. Les colonnes 2-4 du tableau 4.1 donnent des exemples de marques de cohésion et de changements de thème utilisées pour la segmentation. Nous distinguons quatre dimensions :

**Discursive** Elle met l'emphase sur l'organisation des tours de parole et spécifie le locuteur.

**Sémantique** Elle met l'emphase sur les objets mentionnés dans un tour de parole. De manière générale, cela revient à relever les syntagmes nominaux.

**Syntaxique** Elle met en évidence les marques telles que les conjonctions et les adverbes de lien qui jouent un rôle dans la cohésion thématique [Renkema, 1993].

**Lexicale** Elle met l'emphase sur certains mots reconnus pour être des marques d'acquiescement, de confirmation ou qui indiquent une clôture de thème [Maynard et Zimmerman, 1984], le début d'une conversation ou la fin [Sacks et Schegloff, 1973].

### 4.3.5 Modèle de langue

Le problème de la détection des changements de thème peut être transposé en un problème de classification de frontières. Pour ce faire, nous émettons l'hypothèse qu'entre chaque ensemble de marques extraites d'un énoncé se situe probablement une frontière qui permet de délimiter deux classes d'énoncés. La segmentation optimale en unités thématiques est

[6] La liste complète des traits utilisés pour la segmentation thématique est présentée à l'annexe B.

| No | Loc | Énoncé | Disc. | Sém. | Synt. | Lex. |
|---|---|---|---|---|---|---|
| 1 | a | : Maritime operation centre, (INAUDIBLE) hello. | a | Organisation | | hello |
| 2 | b | : hi, Mr. Wellington, it's captain Mr. VanHorn | b | Person | | hi |
| 3 | a | : yes. | a | | | yes |
| | | ..... | ..... | ..... | ..... | ..... |
| 4 | b | : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat on the South Coast of Newfoundland, just in the area quite between Fortune Bay and Trepassey. | b | Boat, Location | | ha |
| 5 | b | : it's on the south east coast of Newfoundland. | b | location | | |
| | | ..... | ..... | ..... | ..... | ..... |
| 6 | b | : this is been going on for, for 24 hours that the case has, or almost anyway, and we had an DFO King Air up flying this morning | b | Time, Aircraft | | |
| 7 | b | : they did a radar search for us in that area. | b | MeansOfDetection | | |
| 8 | a | : yes. | a | | | yes |
| 9 | b | : and their search turned up nothing. | b | | and | |
| 10 | a | : yeah. | a | | | yeah |
| | | ..... | ..... | ..... | ..... | ..... |
| 14 | b | : so I'm wondering about the possibility of attempting it with a different platform perhaps someone with even other sensors other than the radar and, in fact, someone with a, with a radar that'll be a little more sensitive. | b | | so | |
| 15 | b | : before I used the Challenger, I 'll use a Hurk. | b | Aircraft, Aircraft | before | |
| | | ..... | ..... | ..... | ..... | ..... |
| 20 | a | : ...do you want this thing fired up now or you wanna wait till the Big Boys come in to work tomorrow morning? | a | Time | ? | |
| 21 | b | : well, I would like it if possible, I'd like them to, to be airborne at first light. | b | | well | |
| 22 | a | : ok. | a | | | ok |

TAB. 4.3 – Marques discursives (Disc.), sémantiques (Sém.), syntaxiques (Synt.) et lexicales (Lex.) utilisées pour la segmentation en thèmes. Les lignes en pointillés représentent les frontières des unités thématiques.

déterminée par l'équation 4.2 similaire à l'équation 4.1 de la section 4.2.3 :

$$\hat{q} \quad = \quad \underset{q}{\operatorname{argmax}} \prod_{t=1}^{n} P(q^t|q^{t-1}) P(o^t|q^t) \tag{4.2}$$

$\hat{q}$ est la séquence optimale d'états qui reflète la meilleure segmentation en unités thématiques d'une séquence d'énoncés de longueur $n$, $o^t$ est la $t^{\text{eme}}$ combinaison de marques observées et les $q^i$ sont les états du modèle.

Notre modèle de Markov d'ordre 1 est composé de cinq états (figure 4.2). Chacun des états représente une classe d'énoncé :

- Les énoncés qui indiquent un début de conversation contenant, généralement, des salutations ainsi que l'identification des locuteurs sont représentés par la classe `BC` (Begin Conversation).
- Les énoncés qui clôturent une conversation sont représentés par la classe `EC` (End Conversation). Ces énoncés contiennent, généralement, des expressions typées telles que `talk to you later`, `bye`, `have a good day`.
- Les énoncés qui débutent un nouveau thème forment la classe `TC` (Topical Change).
- Les énoncés qui font partie du corps d'un thème sont représentés par la classe `No-TC` (No Topical Change).
- Les énoncés qui terminent un thème sont représentés par la classe `ET` (End of Topic). Ces énoncés sont souvent composés d'unités lexicales telles que `ok`, `right`, `well`.

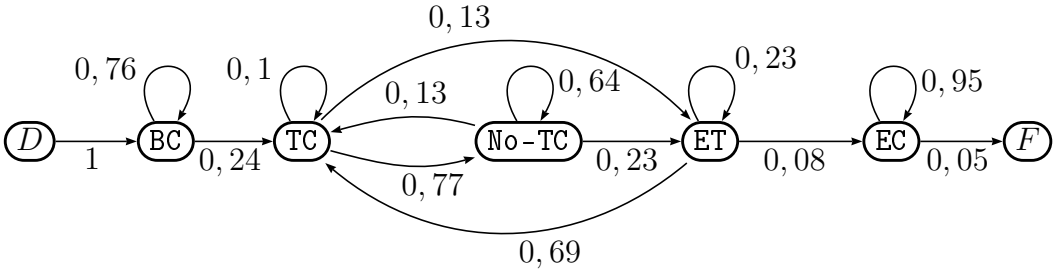


FIG. 4.2 – Modèle de Markov d'ordre 1 pour la segmentation par thème.

La figure 4.2 illustre notre modèle de langue. Les valeurs au-dessus des arcs sont les probabilités de transition entre les états `BC`, `TC`, `No-TC`, `ET`, `EC` et sont calculées par l'équation 4.3 :

$$P(q^i|q^j) = \frac{\#(q^i, q^j)}{\sum_{q^k \in Q} \#(q^i, q^k)} \tag{4.3}$$

$\#(q^i, q^j)$ représente le nombre de fois que les états $q^i$ et $q^j$ se succèdent. $\#(q^j)$ représente le nombre de fois que l'état $q^j$ apparaît dans le corpus d'étude. $Q$ est l'ensemble des états du modèle.

Les probabilités d'émission des observations sont les fréquences relatives des vecteurs de traits observés sur le corpus d'étude pour chaque état.

$$P(o^i|q^j) = \frac{\#(o^i, q^j)}{\sum_{o^i \in \Sigma} \#(o^i, q^k)}$$

$\Sigma$ représente l'ensemble des symboles émis, $o^i$ est la combinaison des marques syntaxiques et sémantiques extraites d'un énoncé $u$ et $\#(o^i, q^j)$ le compte de l'observation $o^i$ émise par l'état $q^j$. Les probabilités de transition entre les états du modèle sont les fréquences relatives des transitions d'un état à un autre.

Notons que les états `ET` et `No-TC` font référence aux énoncés contenus dans une même unité thématique. Cependant, ces deux classes se distinguent par des marques propres à chaque type : la première contient des marques de clôture d'une unité thématique, telles que les marques d'acquiescement, tandis que la deuxième contient des marques de cohésion. En conséquence, les erreurs de classification entre ces deux classes ne constituent pas des erreurs de détection de changement de thème.

### 4.3.6 Expériences et résultats

Le corpus utilisé pour la segmentation est composé de 65 conversations totalisant 3700 énoncés annotés avec les différentes classes d'énoncés. Nos résultats sont obtenus sur 10 validations croisées avec 85 % du corpus utilisé pour l'apprentissage et 15 % pour le test.

| | BC | EC | TC | No-TC |
|---|---|---|---|---|
| Markov d'ordre 1 | 24,0 % | 12,1 % | 38,6 % | 9,9 % |
| Baseline | 75,0 % | 75,0 % | 75,0 % | 75,0 % |

TAB. 4.4 – Taux d'erreurs de classification avec un modèle de Markov d'ordre 1.

| Classe d'énoncé | Rappel | Précision |
|---|---|---|
| BC | 83.0 % | 76.0 % |
| EC | 82.6 % | 87.9 % |
| TC | 67.3 % | 61.4 % |
| No-TC | 72.6 % | 75.1 % |

TAB. 4.5 – Pourcentage des rappel, précision et F-score pour les différentes classe d'énoncé de notre corpus.

Les taux d'erreurs de classification sont inférieurs à ceux du *baseline* obtenus en calculant la probabilité d'avoir un changement de thème au hasard (un sur quatre) (tableau 4.4). Les taux d'erreur les plus élevés sont observés pour les états modélisant les énoncés qui débutent la conversation (`BC`) et pour les changements de thème (`TC`) (tableau 4.5). L'analyse de ces erreurs montre que, pour la classe `BC`, 16,5 % des énoncés concernés contiennent uniquement la marque du locuteur, tandis que pour la classe `TC`, ce taux augmente à 35,9 %.

L'importance de la segmentation par thèmes pour la tâche d'EI a déjà été soulignée [Reynar, 1999; Manning, 1998], mais très peu expérimentée, à l'exception du système développé par Manning [Manning, 1998] sur des textes semi-structurés. Dans son approche, il utilise une segmentation par thèmes dans le but de définir une structure hiérarchique thématique d'annonces publicitaires pour la vente de maisons. Il utilise un algorithme basé sur des marques lexicales dépendantes du domaine, telles que `garage` et `bedrooms`, qui identifient deux niveaux de hiérarchie. Le segmenteur s'exécute avec 53 % de précision et 45 % de rappel. Malgré le taux d'erreur important de notre segmenteur, ce dernier donne de bons résultats étant donné la difficulté de la tâche et les résultats des systèmes obtenus sur d'autres types de textes moins complexes.

## 4.4 Identification des thèmes

Cette étape détermine le sujet ou thème de chaque unité thématique d'une conversation. Nous avions défini un thème comme un événement pour l'EI dans le domaine de la recherche et sauvetage. Cette information intervient dans trois étapes du processus d'EI :

1. L'étiquetage sémantique robuste des textes conversationnels présenté dans le chapitre 5.
2. La résolution d'anaphores pronominales en position de sujet présentée dans le chapitre 6.
3. L'apprentissage des patrons d'EI aussi présenté dans le chapitre 6.

### 4.4.1 Traits utilisés

Nous considérons que les éléments déterminants pour l'identification des thèmes sont les entités apparaissant dans une unité thématique. Ces entités sont les concepts de l'ontologie du domaine que nous avons développée (chapitre 5). Elles sont fournies par l'étiqueteur sémantique grâce à un automate à états finis qui reconnaît les termes du domaine répertoriés dans l'ontologie. Le tableau 4.6 donne des exemple des concepts utilisés pour l'identification des thèmes et ceux où ils peuvent apparaître.

Les entités ne sont pas individuellement discriminantes. Les expériences décrites à la section 4.4.2 montreront que les entités n'ont pas la même importance pour l'identification d'un thème.

### 4.4.2 Expériences et résultats

Nous avons expérimenté deux approches de classification :

1. L'algorithme ID3 [Quinlan, 1986] pour l'induction d'un arbre de décision qui suppose la présence d'attributs plus discriminants que d'autres. L'algorithme ID3 a été développé par Quinlan [1986] et repose sur le principe *diviser pour régner*. C'est une procédure récursive qui, à chaque étape, sélectionne un attribut maximisant une fonction de gain

| Entité | Thème |
|---|---|
| INCIDENT | *Incident* |
| AIRCRAFT-SAR | *Search-unit* |
| COLOR | *Missing-object* |
| VESSEL | *Missing-object*, *Search-unit*, *Incident* |
| TIME | *Search-unit*, *Incident* |
| DISTANCE | *Incident*, *Missing-object* |
| MEANSOFDETECTION | *Search-unit* |
| PERSON-SAR | *Controller*, *Search-unit* |
| TASK-RESULT | *Search-unit* |
| LOCATION | *Incident*, *Search-unit* |
| WEATHER | *Incident*, *Search-mission* |
| ORGANISATION | *Controller* |
| INITIAL-ALERT | *Incident*, *Missing-object* |

TAB. 4.6 – Exemples d'entités utilisées pour l'identification des thèmes et les thèmes où ils peuvent apparaître.

d'information définie par la différence entre l'entropie du noeud parent et la somme pondérée des entropies des noeuds fils (équation 4.4) :

$$\begin{aligned} G(a,y) &= H(l) - (p_L H(l_L) + p_R H(l_R)) \\ H(l) &= -\sum_{l_i \in t} p(l_i) \log p(l_i) \end{aligned} \tag{4.4}$$

$a$ est un concept des classes principales sur lequel s'effectue la division, $y$ prend une valeur binaire {yes, no} pour indiquer la présence du concept dans l'unité thématique, $l$ est la distribution du noeud que l'on veut partitionner, $l_i$ sont les exemples sous le noeud $l$, $p_L$ et $p_R$ sont les proportions des exemples passés dans le sous-arbre gauche et dans le sous-arbre droit et $l_L$ et $l_R$ sont les distributions des exemples du sous-arbre gauche et du sous-arbre droit.

2. Le classificateur bayésien naïf (Naive Bayes) qui, au contraire, suppose la même importance pour chaque attribut. Dans ce cas, le choix du thème étant donné le vecteur d'entités $P(t_i|v_j)$ est déterminé par l'équation 4.5 :

| Modèle | Taux d'erreur |
|---|---|
| ID3 | 19,3 % |
| Bayésien naïf | 35,3 % |

TAB. 4.7 – Taux d'erreurs de classification obtenus avec l'algorithme ID3 et le classificateur bayésien naïf.

$$\begin{aligned} t' &= \underset{t_i}{\operatorname{argmax}}\, P(t_i|v_j) \\ &= \underset{t_i}{\operatorname{argmax}}\, \frac{P(v_j|t_i)}{P(v_j)} P(t_i) \\ &= \underset{t_i}{\operatorname{argmax}} [\log P(v_j|t_i) + \log P(t_i)] \end{aligned} \tag{4.5}$$

Avec :

- $v_j = (v_j^1, \ldots, v_j^K)$ avec $K$ le nombre d'entités, et

$$v_j^i = \begin{cases} 1 & \text{si la } j^{\text{eme}} \text{ entité apparaît dans l'unité thématique} \\ 0 & \textit{sinon.} \end{cases}$$

- $t_i \in \{\textit{Missing-object}, \textit{Incident}, \textit{Search-Unit}, \textit{Mission}, \textit{Controller}, \textit{Other}\}$

Les expériences ont été effectuées dans l'environnement Weka[7] (Waikato Environment for Knowledge Analysis) [Witten et Frank, 2000] sur un corpus de 300 unités thématiques et les résultats correspondent à la moyenne de 10 validations croisées. Les tableaux 4.7 et 4.8 montrent les résultats obtenus pour ces deux modèles.

L'algorithme ID3 montre une meilleure performance que le classifieur bayésien naïf car certaines entités sont plus discriminantes que d'autres pour la classification des thèmes. Par exemple, 78 % des occurrences de l'entité MEANSOFDETECTION apparaissent dans le thème *Search-unit*. La source majeure des erreurs se situe au niveau de la détection du thème *Incident* avec un rappel de 56,6 % car 34,5 % des unités thématiques de type *Incident* ont été

[7] `http://www.cs.waikato.ac.nz/~ml/weka/`

| Thème | Précision | Rappel |
|---|---|---|
| *Incident* | 72,0 % | 56,6 % |
| *Mission* | 75,8 % | 75,8 % |
| *Search-unit* | 68,0 % | 75,1 % |
| *Missing-object* | 84,2 % | 70,9 % |
| *Controller* | 78,0 % | 76,2 % |
| *Other* | 81,2 % | 91,6 % |

TAB. 4.8 – Pourcentage de la précision et du rappel par classe de thèmes avec l'algorithme ID3.

classées *Other*. D'une part, la longueur des unités thématiques portant sur l'incident (2 énoncés) est plus petite que la moyenne des longueurs des autres unités thématiques (3,4 énoncés). D'autre part, moins d'entités discriminantes apparaissent avec le thème *Incident*, car les incidents sont souvent décrits par des adjectifs ou des verbes (`overdue`, `missing`) qui ne sont pas très représentés dans l'ontologie : la majorité des informations représentées dans l'ontologie sont des noms. Enfin, le meilleur résultat est obtenu avec la détection des thèmes non pertinents pour le domaine (classe *Other*) car le nombre de ses exemples (38 % du corpus) dépasse celle de toutes les autres classes.

## 4.5 Conclusion

Dans ce chapitre, nous avons présenté deux approches de segmentation. La première détermine les segments linguistiques qui contiennent une relation **prédicat-arguments**. Essentiellement, cette tâche consiste à distinguer les paires d'adjacence.

La deuxième approche de segmentation proposée effectue un découpage par thème des textes conversationnels. Nous avons défini le thème comme étant un événement pertinent au domaine auquel est associé un formulaire d'extraction. Les thèmes établissent une relation entre les entités d'une unité thématique et un formulaire d'extraction.

Les résultats obtenus pour ces tâches nous permettent d'intégrer ces modules en amont de l'étiquetage sémantique et de l'apprentissage de schémas d'extraction. Dans le prochain

chapitre, nous présentons l'ontologie du domaine utilisée pour l'étiquetage sémantique.

# Chapitre 5

# Conception de l'ontologie du domaine de la recherche et sauvetage

## 5.1 Introduction

Dans ce chapitre, nous décrivons les étapes de la conception de notre ontologie du domaine de la recherche et sauvetage. Le but de cette ontologie est de fournir une représentation structurée des termes du domaine pour l'étiquetage sémantique des expressions pertinentes au domaine.

Le choix d'une ontologie est motivé par la nature des tâches d'étiquetage sémantique et d'apprentissage des patrons d'extraction auxquelles elle servira. La première nécessite le calcul de scores de similarité entre les expressions et les concepts de l'ontologie pour identifier celles sémantiquement similaires aux termes du domaine, tandis que la seconde nécessite le repérage des arguments d'un prédicat.

Également, les relations hiérarchiques telles que *is-a* et *part-of* permettent la généralisation de l'utilisation de classes de mots dans les patrons d'extraction pour contrer la sous-représentation des relations pertinentes ; dans les approches standard d'EI, l'utilisation

des classes de mots est généralement limitée aux entités nommées.

Par ailleurs, l'ontologie établit une relation non ambiguë entre un sens particulier d'un terme et un objet du domaine. Cette caractéristique, importante pour le calcul des scores de similarité, réduit le problème de désambiguïsation des sens des mots [Graeme, 2004].

Nous proposons ici un bref état de l'art des approches de conception d'ontologie et nous décrivons le processus que nous avons suivi pour en construire une pour le domaine de la recherche et sauvetage.

## 5.2 Définition de la notion d'ontologie

Une ontologie est une spécification formelle de la conceptualisation d'une réalité. Pour une réalité donnée, les éléments importants et les relations existant entre eux sont modélisés. Elle comprend un ensemble de termes regroupés en concepts ou catégories, leurs définitions et les relations qui les lient.

Les termes peuvent être regroupés de différentes façons selon la taxonomie utilisée.

Les relations *is-a* et *part-of* sont deux exemples de taxonomie couramment employéees. La première est une relation hiérarchique permettant de réorganiser les termes qui sont des instances de concept, ces termes peuvent être des synonymes ou non. La seconde décrit une relation partie-tout.

Ces deux relations permettent une organisation structurelle des connaissances, tandis que d'autres relations peuvent être utilisées pour faire une classification fonctionnelle des connaissances.

Par exemple l'ontologie UMLS [Humphreys et Lindberg, 1993] développée pour le domaine médical (Unified Medical Language System) utilise la relation *is-a* pour regrouper les connaissances en concepts tels que SYMPTÔMES ou MALADIES, tandis qu'une deuxième relation de type *functionally-related-to* modélise la relation de cause à effet qui existe entre certains termes de ces concepts. La figure 5.1 présente une partie de la hiérarchie des relations *functionally-related-to* de UMLS.

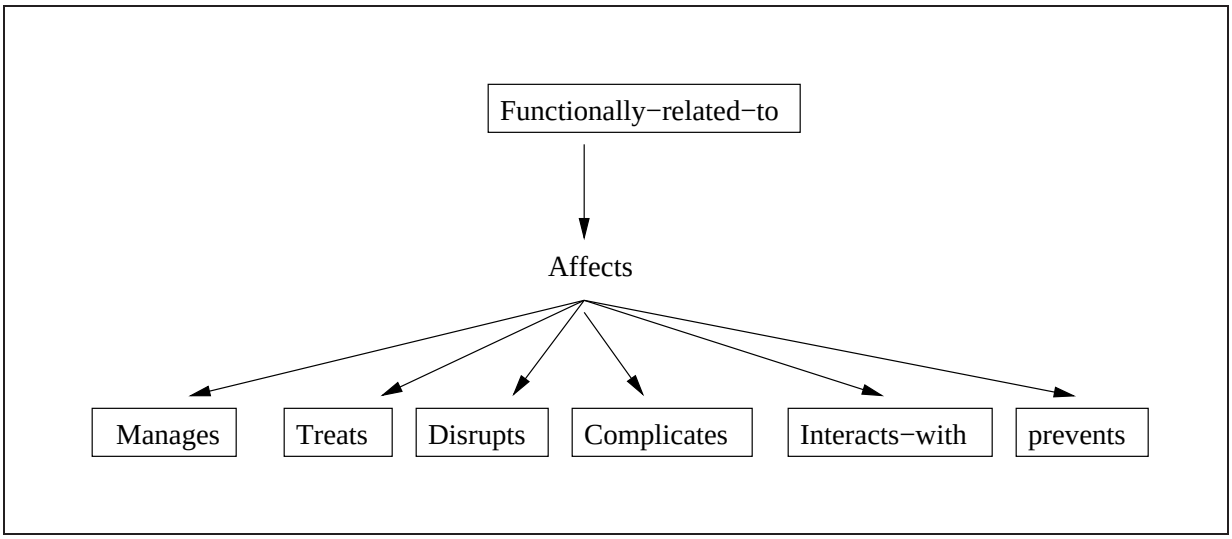


FIG. 5.1 – Extrait de la hiérarchie des relations de l'ontologie UMLS.

## 5.3 Approches de conception

Une des difficultés de la conception d'une ontologie est d'établir une correspondance entre la représentation structurelle des données et leur représentation fonctionnelle, c'est-à-dire déterminer *à quoi vont servir les connaissances*. Selon la prise en compte de cet aspect, différentes approches de conception peuvent être utilisées. Nous distinguons trois approches de conception d'ontologie [Noy et Hafner, 1997] :

- L'approche descendante consiste à raffiner les concepts les plus généraux vers les plus spécifiques. Chaque étape de spécialisation se base sur un critère qui divise les classes en sous-classes disjointes. Cette approche a été utilisée par Sowa[1] pour définir une ontologie générale. Certes, elle permet de couvrir de manière exhaustive les termes du domaine, cependant le choix des critères pour diviser les classes générales ne tient pas compte de l'objectif pour lequel l'ontologie a été créée (le côté fonctionnel). Ainsi, les critères peuvent ne pas garantir une division en sous-classes disjointes et augmenter l'ambiguïté des termes, ce qui est problématique pour la tâche d'étiquetage sémantique.
- L'approche ascendante a été utilisée pour la conception d'un bon nombre d'ontolo-

[1] `http://www.jfsowa.com/ontology/`

gies telles que Wordnet [Miller, 1990]. Dans cette approche, le point de départ est un dictionnaire ou un corpus à partir duquel sont relevés les termes importants et les relations entre ces termes. Les termes sont organisés en concepts spécifiques qui à leur tour sont regroupés en concepts plus généraux selon une taxonomie choisie. Bien que cette approche soit très utilisée, elle présente le même inconvénient que l'approche descendante, toutefois avec un risque de ne couvrir qu'une partie des termes pertinents pour la tâche cible.

- La troisième approche combine les deux approches précédentes et a été utilisée pour la conception de l'ontologie TOVE[2] dans le domaine des entreprises commerciales et publiques.

  L'étape descendante sert à définir le rôle de l'ontologie dans la réalisation d'une tâche donnée. Cela peut se faire en établissant une liste de questions que doit couvrir l'ontologie pour accomplir la tâche cible. Les questions définissent les concepts des niveaux supérieurs de l'ontologie.

  L'étape ascendante détermine les réponses possibles aux questions et entraîne la définition des concepts du niveau inférieur de l'ontologie. Enfin, un processus itératif de regroupement de ces concepts permet de converger vers les concepts généraux identifiés lors la première étape. Cette approche peut être qualifiée d'orientée-tâche puisque tout le processus de développement de l'ontologie est lié aux besoins de la tâche. L'approche descendante garantit l'exhaustivité de la couverture des concepts pertinents du domaine, tandis que l'approche ascendante oriente la couverture des termes du domaine vers ceux qui sont utilisés pour la tâche cible.

[2] http://www.eil.utoronto.ca/enterprise-modelling/tove/

## 5.4 Approche utilisée pour la conception de l'ontologie du domaine de la recherche et sauvetage

Ainsi, des trois approches de conception d'ontologie présentées, nous retenons la dernière car le développement de l'ontologie est dirigé par les besoins réels de l'application. Aussi, nous pouvons exploiter deux sources d'information existantes pour définir la portée de l'ontologie et le contenu des classes :

- Les formulaires d'extraction peuvent être utilisés pour définir la couverture de l'ontologie et construire une liste de questions qui délimitent la portée de l'ontologie.
- Les réponses de champs de formulaires sont les informations que nous voulons trouver dans les textes conversationnels et elles peuvent être utilisées pour définir les termes à collecter et à organiser dans l'ontologie. Une partie importante de ces termes est décrite dans les manuels du domaine de la recherche et sauvetage maritime [Pêches et Océans Canada, 2000].

La figure 5.2 montre les concepts regroupant les types de réponses pour chaque champ considéré. Les flèches indiquent la relation entre les champs de formulaires d'extraction et les concepts de l'ontologie.

### 5.4.1 Étapes de la conception

L'approche que nous utilisons pour la conception de notre ontologie entre dans le cadre d'une approche combinée descendante et ascendante et s'inspire de l'ouvrage de Noy et McGuinness [2001]. C'est un processus itératif qui englobe les étapes suivantes :

**Déterminer le domaine de l'ontologie** Le domaine couvert par l'ontologie est celui de la recherche et sauvetage maritime, plus particulièrement les aspects opérationnels des missions de recherche et sauvetage tels que les ressources disponibles pour la recherche (avion, bateau, hélicoptère), les lieux concernés, les organisations mises à contribution et les types d'incident.

**Déterminer la portée de l'ontologie** Il s'agit de définir une liste des termes importants

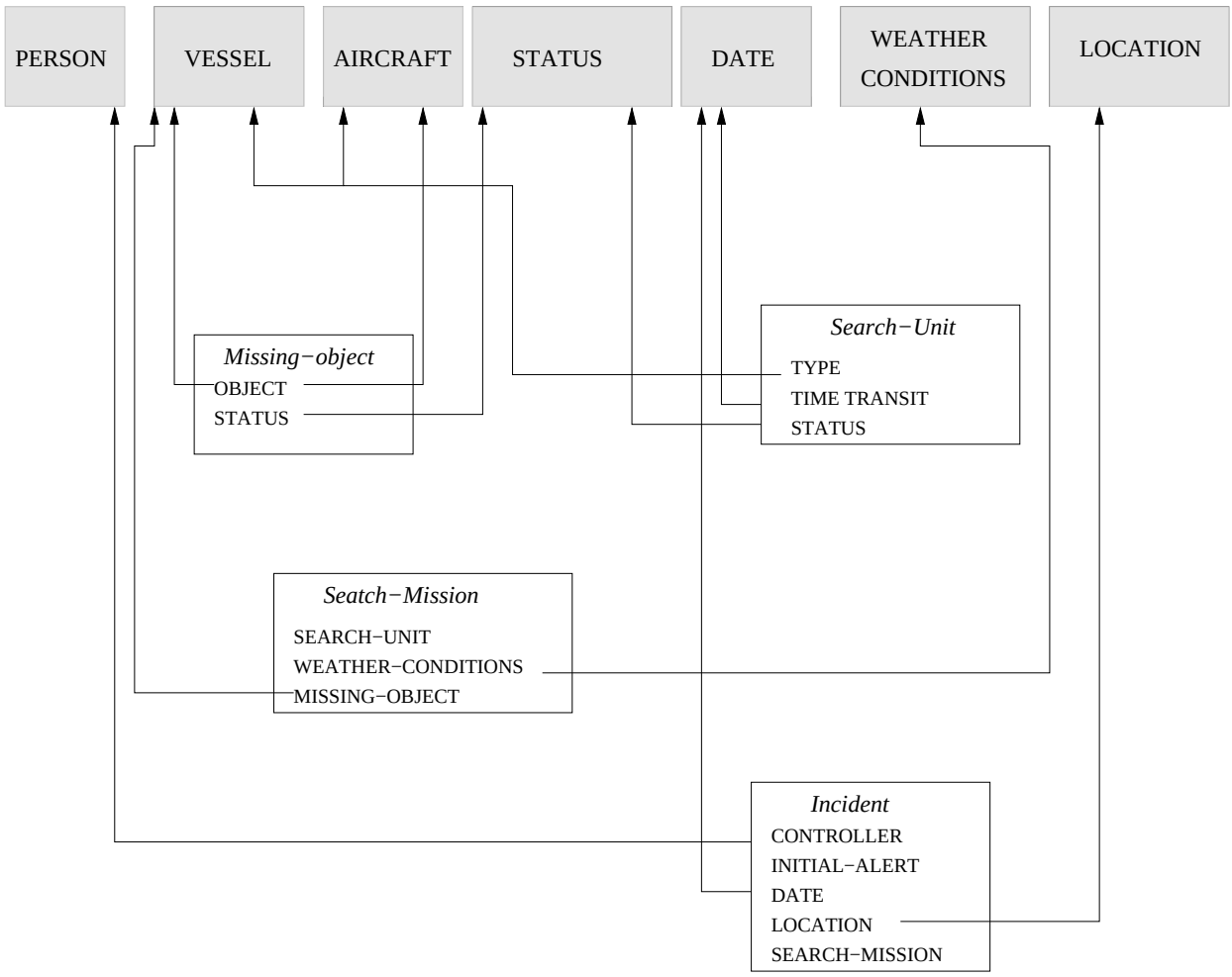


FIG. 5.2 – Exemples des relations entre des concepts de l'ontologie et les champs de formulaires d'extraction.

pour la tâche cible : l'extraction d'information. L'ontologie contient un ensemble représentatif des réponses de champs de formulaires d'extraction. Elle permet de répondre à des questions telles que :

- Quels sont les types de bateaux utilisés pour une mission ?
  - *Zodiacs*, *SAR auxiliary* ...
- Quels sont les types d'avions utilisés pour une mission ?
  - *Aurora*, *King Air*, *Hercules* ...
- Quelles sont les organisations qui font partie du département de la recherche et sauvetage ?
  - *Rescue Coordination Centre*, *Maritime Operation Centre* ...
- Quels sont les critères des conditions météorologiques ?
  - *Wind*, *Sea*, *Precipitation* ...
- Quelles sont les conditions météorologiques ?
  - *High sea*, *Fog*, *Haze* ...
- Quels sont les exemples d'incident ?
  - *Overdue*, *Missing*, *Fire* ...
- Quels sont les exemples de description des bateaux ?
  - *Red*, *Trim*, *Open boat* ...

**Définir la hiérarchie de classes** Nous avons utilisé deux types de relations pour définir la taxonomie de notre ontologie. La relation *is-a* groupe les termes qui sont des exemples d'un concept, par exemple les termes `fog`, `rain`, `wind` et `visibility` sont des exemples du concept WEATHER-CONDITIONS, tandis que les termes `broken`, `disabled` et `fire` sont des exemples du concept INCIDENT.

La relation *part-of* décrit les fonctions de certains concepts par rapport à d'autres. Par exemple, nous avons utilisé la relation *part-of* pour établir un lien entre les composantes d'un bateau et le bateau lui-même, car souvent les composantes de bateaux sont utilisées pour indiquer la cause d'un incident, comme dans `fire in the captain's cabin`.

**Définir les classes de l'ontologie** Les termes qui composent les classes de l'ontologie ont été collectés à partir d'un manuel fourni par le Secrétariat de la Recherche et Sauvetage National [Pêches et Océans Canada, 2000] et d'un échantillon de dix conversations choisies au hasard.

Ces termes ont été groupés selon les relations décrites à l'étape précédente. Ce sont surtout des noms propres tels que les noms d'avions (`King Air`), mais aussi des collocations de noms communs (`Emergency Locator Transmitter`), des expressions adjectivales (`overdue`) ou verbales (`drifting`).

Par ailleurs, dans la section 2.3.1, nous avons montré que le vocabulaire de nos conversations est varié, ce qui implique que la couverture de tous les termes du domaine est difficile à réaliser, en particulier pour les verbes. Pour pallier ce problème, nous avons proposé de calculer des scores de similarité pour identifier les expressions sémantiquement similaires aux termes de l'ontologie. Cependant, comme les instances des concepts ne sont pas des synonymes, nous avons enrichi chaque instance de l'ontologie avec une liste d'au plus 3 synonymes ou mots similaires extraits du dictionnaire thésaurus Wordsmyth (section 6.3). L'étape d'enrichissement a été effectuée de manière semi-automatique. Nous avons développé un script qui extrait toutes les définitions d'une instance à partir de la version Prolog de Wordsmyth[3]. Ensuite, une étape manuelle de révision des définitions ajoutées a permis de garder uniquement les définitions propres au domaine de la recherche et sauvetage. Dans la section 6.5, nous montrons que cette procédure permet le lissage des scores de similarité afin de réduire les chances d'obtenir des scores nuls entre deux expressions sémantiquement similaires.

[3]La version originale de Wordsmyth que nous avons obtenue est en format texte. Toutefois, pour les besoins de notre projet, nous avons développé un programme pour construire de manière automatique une version de Wordsmyth en format Sicstus Prolog.

| Super-classe | Classe | Exemples de termes |
|---|---|---|
| **Physical Entity** | AIRCRAFT | `Aurora`, `King Air` |
| | VESSEL | `Zodiac`, `Doray` |
| | LOCATION | `Halifax`, `St-Johns` |
| | ORGANISATION | `MOC`, `RCC` |
| | MEANSOFDETECTION | `Radar`, `Satellite` |
| **Conceptual Entity** | INCIDENT | `Fire`, `Overdue` |
| | PROPERTIES | `Completed`, `red` |
| | SEARCH-UNIT | `Rescue 117`, `Rescue 306` |
| | WEATHER | `Haze`, `Fog` |
| | INITIAL-ALERT | `emergency`, `difficulty` |
| | LANGUAGE | `Alpha`, `Wetecker` |
| | TIME | `morning`, `today` |
| | TASK | `Search`, `tow` |

TAB. 5.1 – Liste des concepts du niveau supérieur de la hiérarchie *is-a* de notre ontologie du domaine de la recherche et sauvetage et quelques exemples de leurs instances.

## 5.5 Implémentation

Nous avons identifié 13 classes principales, groupées sous les super-classes **Physical Entity** et **Conceptual Entity**. Ces classes sont listées dans le tableau 5.1 avec des exemples d'instances.

Les termes sont organisés en 51 classes selon une hiérarchie *is-a* (43 classes) et une autre *part-of* (8 classes) incluant les classes principales présentées dans le tableau 5.1. Une liste complète des classes de la hiérarchie *is-a* et *part-of* est donnée dans les figures 5.3 et 5.4. L'ontologie a une profondeur de 4 niveaux et contient 783 feuilles[4] Chaque entrée de l'ontologie contient un terme, la définition textuelle du sens propre au domaine et une liste exhaustive de synonymes et de mots similaires, tous extraits de Wordsmyth. Les entrées du terme `land` sont données en exemple à la figure 5.5.

Le développement de l'ontologie a été élaboré de manière à couvrir les noms propres contenus dans toutes les conversations (nom d'avions, bateaux), la liste des noms de lieux au Canada, ainsi que les termes décrit dans les manuels du domaine de la recherche et sauvetage

[4]La liste des termes du domaine définis dans l'ontologie est présentée à l'annexe A.

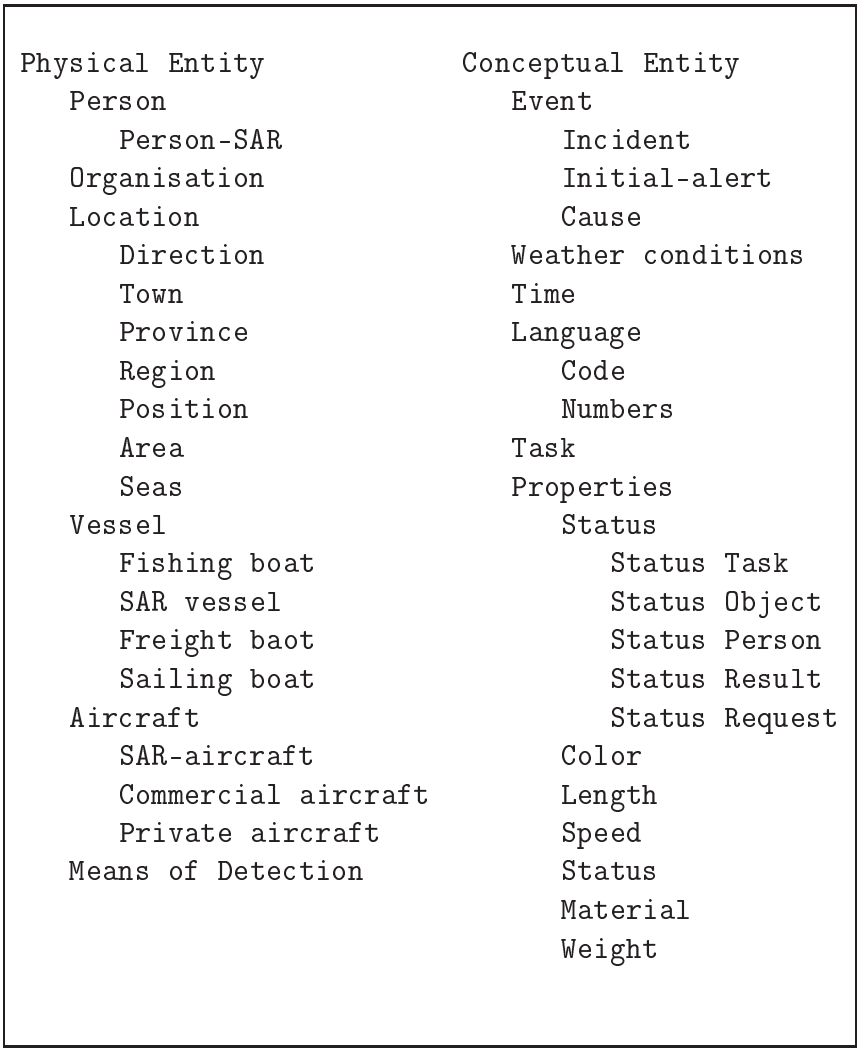


FIG. 5.3 – Hiérarchie *is-a* de l'ontologie du domaine de la recherche et sauvetage maritime. La première colonne représente les classes des différents niveaux sous la classe Physical entity, tandis que la deuxième colonne contient les classes sous le concept Conceptual Entity.

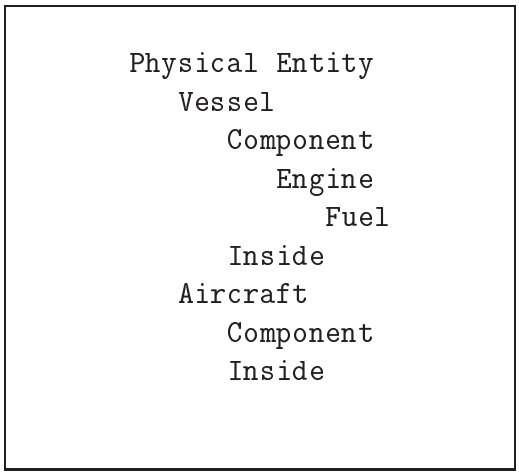


FIG. 5.4 – Hiérarchie *part-of* de l'ontologie du domaine de la recherche et sauvetage maritime.

maritime [Pêches et Océans Canada, 2000] tels que les types d'incident.

## 5.6 Conclusion

Nous avons développé une ontologie pour le domaine de la recherche et sauvetage maritime pour l'étiquetage sémantique robuste des expressions pertinentes au domaine. L'approche que nous avons utilisée a permis de couvrir de manière assez exhaustive l'ensemble des termes pertinents dans le domaine de la recherche et sauvetage.

Toutefois, plusieurs améliorations peuvent être apportées à la version actuelle de l'ontologie. Parmi celles-ci, nous retenons l'utilisation d'un logiciel d'édition d'ontologie tel que Ontolingua[5] pour permettre la réutilisation de l'ontologie par d'autres applications. Également, cela permet d'exploiter les ontologies déjà développées sous cette plate-forme, par exemple une ontologie du temps.

Une autre avenue à explorer, est l'ajout d'axiomes : qui sont des contraintes, telles qu'une valeur de vérité sur les propriétés et le rôle d'un concept pour permettre l'inférence d'information. Cet ajout pourrait être utile pour l'étape de combinaison des informations extraites pour générer des faits plus complexes (section 3.6).

[5] `http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/index.html`

```
status(['land'],'verb',['STATUS-OBJECT']).
hash(arrive,[node(land,verb,1,land,status)]).
hash(secure,[node(land,verb,1,land,status)]).
defonto('land','verb',['cause','touch','down','surface']).
defonto('land','verb',['come','shore']).
```

FIG. 5.5 – Entrées de l'ontologie pour le verbe `land`. La première ligne représente l'instance de la classe `status`, la deuxième et la troisième spécifient ses synonymes : `arrive` et `secure`. Le premier champ indique la racine du synonyme et le deuxième champ représente les informations relatives à l'instance considérée pour la relation de synonymie : sa catégorie syntaxique, son niveau dans la hiérarchie et le concept impliqué. Les deux dernières lignes sont les mots pleins (verbes, adjectifs et noms) contenus dans les définitions de l'instance `land`. Ces définitions sont extraites de Wordsmyth de manière semi-automatique.

# Chapitre 6

# Étiquetage sémantique robuste de textes conversationnels

## 6.1 Introduction

Les approches standard d'EI identifient les faits individuels en procédant à une analyse du contexte (la phrase) et non en analysant le contenu des mots. Nous avons montré que l'utilisation de patrons basés sur la structure syntaxique d'une phrase est inappropriée à cause des irrégularités langagières de l'oral, car celles-ci modifient le contexte. Afin de contourner ce problème, nous avons proposé d'utiliser des patrons basés sur la relation **prédicat-arguments**. Dans ce chapitre, nous définissons une approche d'étiquetage sémantique des expressions pertinentes au domaine composant les prédicats et arguments de relations que nous voulons apprendre de manière automatique. Ce chapitre reprend les résultats de nos travaux présentés à LREC'04 [Boufaden *et al.*, 2004b], TALN'04 [Boufaden *et al.*, 2004a] et ACL'03 [Boufaden, 2003]. Toutefois, l'architecture de l'analyseur sémantique est légèrement différente de celle présentée à ACL 2003, car nous avons inclu un système de détection des thèmes (voir section 6.5.2).

L'étiquetage sémantique est une étape fondamentale de notre processus d'EI, plus générale que celle de l'extraction des entités nommées dans la mesure où elle couvre plus de classes sémantiques et détecte les expressions sémantiquement similaires à celles couvertes par l'ontologie du domaine.

Notre approche repose sur l'utilisation d'une ontologie du domaine et d'un dictionnaire-thésaurus qui représente les connaissances du monde. Le résultat de cette étape est une conversation où les termes et leurs expressions sémantiquement similaires sont identifiés et classés. Le tableau 6.1 illustre une sortie produite par notre étiqueteur.

## 6.2 Architecture de l'étiqueteur

L'étiqueteur sémantique est formé de deux modules complémentaires :

1. Le premier reconnaît les termes couverts par notre ontologie. L'approche utilisée s'inscrit dans le cadre des approches basées sur les connaissances du domaine (section 3.4.1).
2. Le second effectue un travail de repêchage en sélectionnant et filtrant les expressions sémantiquement similaires aux termes couverts par notre ontologie. Ce module gère les variations langagières des conversations et garantit la robustesse de cette étape. Il utilise trois sources d'information :
   - L'ontologie du domaine.
   - Le dictionnaire-thésaurus Wordsmyth.
   - Le thème de l'unité thématique contenant l'expression sémantiquement similaire à un terme du domaine.

La figure 6.1 montre l'organisation des différentes composantes de l'étiqueteur sémantique.

| No | Loc | Énoncé |
|---|---|---|
| 1 | a | : Maritime operation centre (ORGANISATION), (INAUDIBLE) hello. |
| 2 | b | : hi, Mr. Wellington (PERSON), it's captain Mr. VanHorn (PERSON) |
| 3 | a | : yes. |
| | | ..... |
| 4 | b | : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat (VESSEL) on the South Coast of Newfoundland (LOCATION), just in the area quite between Fortune Bay and Trepassey (LOCATION). |
| 5 | b | : it's on the south east coast of Newfoundland (LOCATION). |
| | | ..... |
| 6 | b | : this is been going on for, for 24 hours (TIME) that the case has, or almost anyway, and we had an DFO King Air (AIRCRAFT) up flying (STATUS) this morning (TIME). |
| 7 | b | : they did (STATUS) a radar search (MEANSOFDETECTION) for us in that area (LOCATION). |
| 8 | a | : yes. |
| 9 | b | : and their search (TASK) turned up (STATUS) nothing (RESULT). |
| 10 | a | : yeah. |
| | | ..... |
| 14 | b | : so I'm wondering about the possibility of attempting it with a different platform perhaps someone with even other sensors other than the radar and, in fact, someone with a, with a radar that'll be a little more sensitive. |
| 15 | b | : before I used (TASK) the Challenger (AIRCRAFT), I 'll use (TASK) a Hurk (AIRCRAFT). |
| | | ..... |
| 20 | a | : do you want this thing fired up (STATUS) now or you wanna wait (STATUS) till the Big Boys come in to work tomorrow morning (TIME)? |
| 21 | b | : well, I would like it if possible, I'd like them to, to be airborne (STATUS) at first light (TIME). |
| 22 | a | : ok. |

TAB. 6.1 – Extrait de la conversation Overdue boat où les termes du domaine sont annotés sémantiquement. L'étiquette sémantique en dessous des barres de soulignement sont des concepts de l'ontologie du domaine de la recherche et sauvetage. Les lignes horizontales sont les frontières des unités thématiques (ajoutées manuellement).

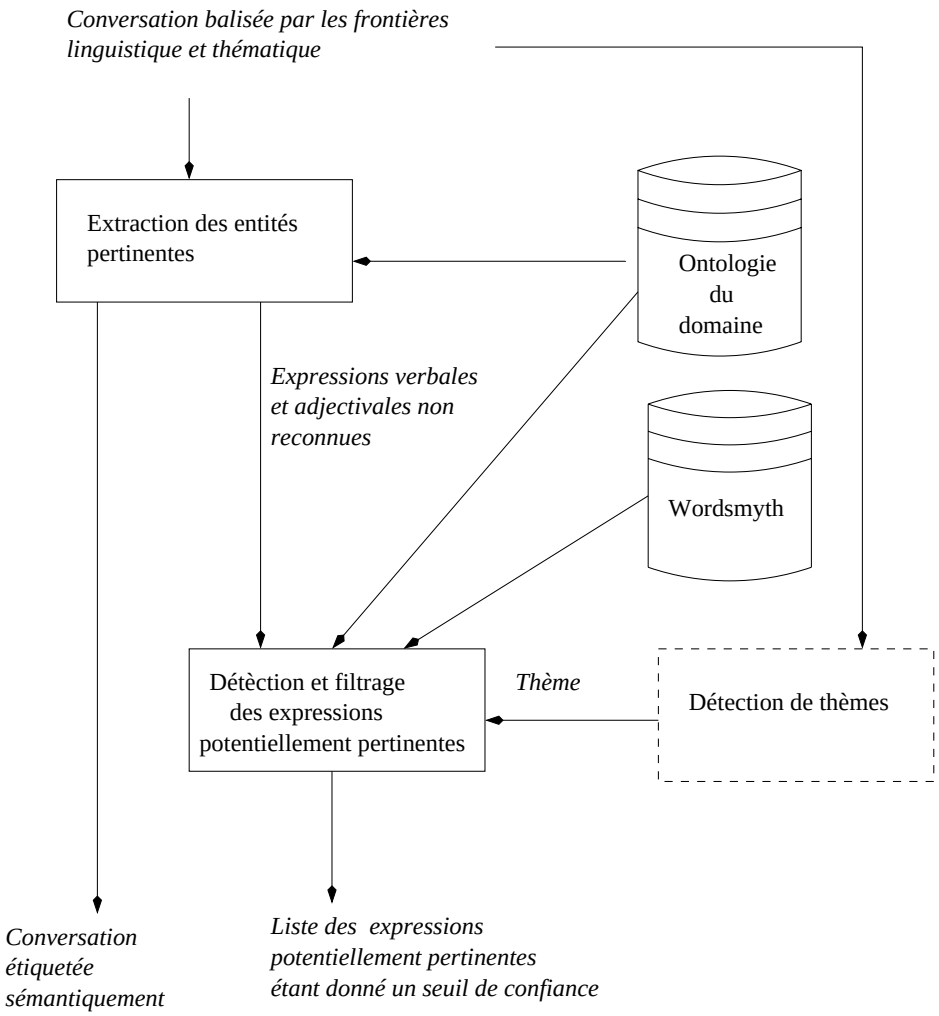


FIG. 6.1 – Architecture de l'étiqueteur sémantique.

## 6.3 Connaissances du monde : dictionnaire-thésaurus Wordsmyth

La ressource représentant les connaissances du monde joue un rôle important dans la détection et l'étiquetage des expressions sémantiquement similaires aux termes du domaine. Trois types d'information sont extraites de cette ressource :

- La définition textuelle d'un mot.
- La liste des synonymes par sens du mot.
- La liste des mots similaires qui sont des hyponymes, des hyperonymes ou mots connexes, pour chaque sens du mot.

Ces informations sont utilisées pour calculer un score de similarité qui permet de détecter des expressions nominales, adjectivales ou verbales potentiellement pertinentes pour le domaine. Ainsi, le choix de cette ressource est conditionné par la couverture et la complétude des informations pour ces trois types d'expression.

Très peu de ressources sont disponibles publiquement comme WordNet [Miller *et al.*, 1990]. Cependant, nous avons renoncé à l'utilisation de WordNet car il ne couvre pas toutes les définitions des mots, mais seulement celles des sens les plus fréquents dans le corpus Brown [Morato *et al.*, 2004].

Bien que Wordsmyth ne soit pas disponible gratuitement, notre choix s'est posé sur ce dictionnaire-thésaurus[1] [2] pour les raisons suivantes :

1. Contrairement à WordNet, il couvre à peu près tous les sens d'un mot de manière exhaustive sans se limiter aux mots les plus fréquents.
2. Il intègre les informations d'un dictionnaire et d'un thésaurus dans une même entrée, ce qui permet d'avoir accès aux définitions textuelles d'un mot, à une liste de mots synonymes ainsi qu'a une liste de mots similaires ou connexes. Ces listes sont importantes car elles sont utilisées dans le calcul des scores de similarité (section 6.5).

[1] Wordsmyth nous a été prêté gracieusement pour la durée de la thèse.
[2] `http://www.wordsmyth.net`

| | Ressource | Noms | Verbes | Adjectifs |
|---|---|---|---|---|
| Nombre d'entrées | Wordsmyth | 34291 | 10046 | 9879 |
| | Wordnet | 114648 | 11306 | 21436 |
| Taux de polysémie | Wordsmyth | 1,63 | 1,84 | 1,76 |
| | Wordnet | 1,23 | 2,17 | 1,44 |

TAB. 6.2 – Tableau comparatif du nombre d'entrées et du taux de polysémie respectifs de Wordsmyth et Wordnet.

Pour illustrer la différence des couvertures de sens entre Wordsmyth et Wordnet, nous avons regardé le nombre de noms, des adjectifs et des verbes couverts par Wordsmyth et WordNet ainsi que leurs degrés de polysémie. Les statistiques concernant WordNet ont été prises à partir de leur site[3], tandis que les statistiques obtenues pour Wordsmyth ont été calculées à partir de la version dont nous disposons. Le tableau 6.2 montre les scores obtenus. La couverture de mots (nom, verbe et adjectif) dans Wordnet est nettement supérieure à celle de Wordsmyth notamment à cause de l'ajout d'expressions composées et de noms propres. Par contre, nous remarquons que le taux de polysémie pour les adjectifs et les noms est supérieur dans Wordsmyth.

À titre d'exemple, nous avons regardé les entrées respectives de l'adjectif `overdue` dans WordNet (figure 6.2) et dans Wordsmyth (figure 6.3). Ce mot clé indique un incident comme le montre l'énoncé `4` de la conversation Overdue boat (tableau 6.1).

Nous remarquons que dans WordNet, un seul sens est spécifié et il ne correspond pas au sens propre au domaine de la recherche et sauvetage, tandis que ce sens est couvert par la deuxième définition proposée par Wordsmyth.

## 6.4 Étiquetage des expressions couvertes par l'ontologie

Les expressions considérées dans cette étape de l'étiquetage sémantique sont les groupes nominaux, les verbes ainsi que les adjectifs prédicatifs couverts par l'ontologie.

Pour l'étiquetage morphosyntaxique, nous avons utilisé le système de Brill [1992] en-

[3] `http://wordnet.princeton.edu/`

```
> wn overdue
No information available for noun overdue
No information available for verb overdue
Information available for adj overdue
        -antsa          Antonyms
        -synsa          Synonyms (ordered by frequency)
        -famla          Familiarity & Polysemy Count
        -grepa          List of Compound Words
        -over           Overview of Senses
No information available for adv overdue

> wn overdue -synsa
Similarity of adj overdue
1 sense of overdue
Sense 1
delinquent, overdue
       => due (vs. undue), owed

> wn overdue -over
Overview of adj overdue
The adj overdue has 1 sense (no senses from tagged texts)
1. delinquent, overdue -- (past due; not paid at the scheduled time;
"an overdue installment"; "a delinquent account")
```

FIG. 6.2 – Sortie de WordNet 1.6 pour l'adjectif `overdue`.

traîné sur le corpus Brown. L'étiquetage syntaxique a été réalisé avec le système d'analyse syntaxique partielle CASS de Abney [1996]. Cependant, à cause des irrégularités langagières de l'oral, nous avons réduit la couverture grammaticale des constructions complexes, par exemple celle impliquant une conjonction NP → NP CONJ NP ou une collocation NP → NP NP. Le résultat de l'étiquetage syntaxique d'un énoncé est une forêt d'arbres syntaxiques.

### 6.4.1 Approche

L'étiqueteur sémantique est similaire à un système d'extraction d'entités nommées (section 3.3). Il reconnaît des entités nommées telles que les lieux, les personnes, les organisations

```
ENT :  overdue
SYL :  o-ver-due
PRO :  o vEr du
POS :  adjective
DEF :  1. not paid, delivered, or returned by the due date.
SYN :  late (1), outstanding (3), due (1), unpaid pay (vt
       1,2), owed owe (vt 1), owing
SIM :  tardy, behindhand
DEF :  2. expected or needed for a long time ; long-awaited.
SIM :  awaited await (vt), expected expect (vt), anticipated
       anticipate, promised promise (vt), prospective, due
```

FIG. 6.3 – Description d'une entrée du dictionnaire-thésaurus Wordsmyth pour l'adjectif `overdue` utilisé pour indiquer un bateau en retard et faisant partie des expressions pertinentes qui font référence à un incident. Les acronymes `ENT`, `SYL`, `PRO`, `POS`, `INF`, `DEF`, `EXA`, `SYN`, `SIM` sont respectivement l'entrée, la syllabisation, la prononciation, la catégorie syntaxique, les formes fléchies, la définition textuelle, un exemple, les mots synonymes et les mots similaires. Les chiffres entre parenthèses à côté des mots synonymes (`SYN`) et similaires (`SIM`) indiquent le sens considéré dans la relation de synonymie ou de similarité.

et les noms d'avions, de bateaux et de matériel de détection. Il est basé sur un automate à états finis qui effectue l'étiquetage en trois étapes.

1. La recherche d'un appariement entre la tête d'un syntagme et les instances des concepts de l'ontologie. Les critères d'appariement sont la racine du mot ainsi que sa catégorie syntaxique. Lorsqu'un appariement réussit, la tête est annotée par le concept approprié.
2. La récupération des collocations et conjonctions de syntagmes nominaux écartés à cause de la restriction imposée sur la couverture grammaticale. Elle concerne en particulier les constructions qui contiennent une conjonction telles que l'organisation `Search and Rescue` ou les entités qui sont des collocations de syntagmes nominaux telles que `Rescue Coordination Centre`. Également, cette étape reconnaît les expressions qui font référence à des dates ou des périodes de temps telles que `24 hours ago` et des positions telles que `54 degree`.
3. La propagation de l'étiquette sémantique de la tête du syntagme à tout le syntagme. La figure 6.4 donne un exemple de propagation du trait sémantique WEATHER-

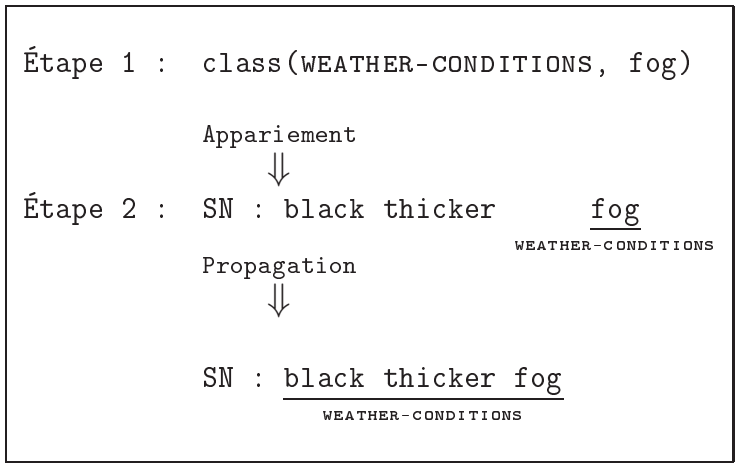


FIG. 6.4 – Le syntagme nominal SN : `black thicker fog` est étiqueté avec le concept WEATHER-CONDITIONS. La première étape de l'analyse sémantique reconnaît la tête `fog` comme un type de conditions climatiques. La deuxième étape propage le concept à tout le syntagme nominal.

CONDITIONS.

### 6.4.2 Expérience et résultat

Nous avons évalué notre système d'étiquetage des expressions sur 64 conversations totalisant 1021 expressions qui sont des termes de l'ontologie. Nous avons basé l'évaluation sur la présence d'un terme de l'ontologie dans l'expression étiquetée, qui est généralement un syntagme (étape de propagation, figure 6.4), et en vérifiant si le concept qui lui est associé est correct. Le rappel correspond au nombre d'expressions étiquetées correctement par le système divisé par le nombre d'expressions étiquetées manuellement, tandis que la précision correspond au nombre d'expressions étiquetées correctement divisé par le nombre total d'expressions étiquetées par le système. Nous avons obtenu un rappel de 85,3 % et une précision de 94,8 % pour un F-score de 89,8 %.

Le premier constat est que le F-score obtenu est relativement bas, comparativement au meilleur score d'extraction d'entités nommées de 96,4 % (section 3.3).

L'analyse des erreurs montre que les mots sémantiquement ambigus ont causé des erreurs dans le choix du concept, influençant donc la précision du système. Par exemple, le mot

`Kilo` fait référence à la mesure de poids mais aussi à la lettre `K` dans l'alphabet militaire. Ainsi, comme nous n'avons pas de mécanisme de désambiguisation du sens à ce niveau, le concept retourné par l'étape d'appariement est le premier rencontré dans notre ontologie : la lettre de l'alphabet. Un autre exemple est celui des noms propres transcrits de manières différentes. Cela arrive surtout avec les noms de lieux tels que `St-Johns` et `Saint Johns` ou avec les noms de bateaux ou d'avions tels que `Hurk` et `Hercules`. Les autres erreurs relevées ont plutôt influencé le rappel, lequel est relativement bas. Elles découlent pour la plupart des erreurs d'étiquetage morphosyntaxique causées par les irrégularités langagières, ces dernières faussent le processus d'appariement basé sur la racine du mot et sa catégorie morphosyntaxique. D'autres erreurs moins fréquentes sont causées par des erreurs d'analyse syntaxique, en particulier pour les collocations. Parmi les exemples de collocations qui ont été mal étiquetées, nous avons trouvé les expressions `Fox island` (lieu), `heavy jacks` (type d'avion) ou `maritime rescue sub-center` (organisation).

## 6.5 Étiquetage des expressions non couvertes par l'ontologie

Cette étape étiquette les expressions sémantiquement similaires aux termes du domaine couverts par l'ontologie. Les expressions visées par ce processus sont les noms communs, les adjectifs et les verbes qui sont pertinents au domaine. Nous voulons étiqueter ces expressions avec les concepts de l'ontologie les plus vraisemblables étant donné le contexte. Une façon d'aborder cette problématique est de maximiser la vraisemblance de la distribution de probabilité :

$$P(C^{1,\ldots,n}|w^{1,\ldots,n}, T^{1,\ldots,n})$$

Où $C^{1,\ldots,n}$ est la séquence des concepts qui étiquettent les mots $w^{1,\ldots,n}$ et $T^{1,\ldots,n}$ la séquence des thèmes des contextes d'énonciation des $w^{1,\ldots,n}$.

L'identification des concepts $k$ est conditionnelle à deux sources d'information complémentaires : le mot $w$, qui apparaît dans un énoncé, et le thème $T$, de cet énoncé (défini à la

section 4.3.1). Un modèle performant doit exploiter le lien qui peut exister entre un mot et un concept ainsi qu'entre un concept et un thème.

Nous proposons d'exploiter le lien de similarité entre un mot $w$ et un concept $C$ et la fréquence des concepts $C$ étant donné un thème $T$. Notre problématique devient donc celle de combiner de manière optimale :

1. La distribution de probabilité de similarité entre les mots $w$ par rapport aux concepts $C$ de l'ontologie : $P(C|w)$.
2. La distribution de probabilité d'observer les concepts $C$ dans les différents contextes d'énonciation $T$ : $P(C|T)$.

Une manière simple de combiner ces deux sources d'information est une combinaison linéaire de la probabilité de similarité avec les concepts et la fréquence de ces concepts étant donné les thèmes :

$$P(C^t = k|w^t = i, T^t = j) = \lambda P(C^t = k|w^t = 1) + (1 - \lambda)P(C^t = k|T^t = j) \qquad (6.1)$$

$\lambda \in [0, 1]$ est un coefficient qui représente la contribution de chaque composante dans le choix du concept.

L'inconvénient de cette modélisation est qu'elle n'est pas assez discriminante car le choix du concept se fait par une des composantes : c'est une disjonction des composantes. Dans notre cas, nous voulons que le choix du concept soit un vote des deux composantes, c'est-à-dire une conjonction des composantes.

Une approche statistique récente [Hinton, 2002] qui formaliserait cette contrainte est celle basée sur le produit de modèles probabilistes. L'avantage de ce modèle est qu'il permet d'exploiter l'expertise de chaque composante dans une problématique plus simple et de la combiner de manière à ce que seules les données qui maximisent les probabilités générées par chaque composante soient retenues. Pour notre problématique, les modèles experts doivent calculer la probabilité de similarité, $P(C|w)$, d'un concept étant donné un mot et la probabi-

lité, $P(C|T)$ d'observer un concept étant donné un thème. Ainsi, la formulation du produit d'experts qui correspond à notre problématique est la suivante :

$$P(C^t = k|w^t, T^t) = \frac{P(C^t = k|w^t)^{\beta_1} P(C^t = k|T^t)^{\beta_2}}{\Sigma_{l=1}^{K} P(C^t = l|w^t)^{\beta_1} P(C^t = l|T^t)^{\beta_2}} \quad (6.2)$$

$k$ est un des concepts de l'ontologie, $P(C^t = k|w^t)$ représente la probabilité d'observer le concept $k$ étant donné le mot $w^t$, que nous estimons à la section 6.5.1, et $P(C^t = k|T^t)$ est la probabilité d'observer le concept $k$ étant donné le thème $T^t$ que nous estimons à la section 6.5.2. $K$ est le nombre de concepts dans l'ontologie. Les coefficients $\beta_1$ et $\beta_2$ sont des poids indépendants des concepts qui reflètent la contribution de chaque expert et $\Sigma_{l=1}^{K} P(C^t = l|w^t)^{\beta_1} P(C^t = l|T^t)^{\beta_2}$ est utilisé pour normaliser le produit des probabilités.

Le modèle décrit dans l'équation 6.2 permet d'attribuer un concept à un mot étant donné un thème. Cependant, dans la mesure où nous voulons étiqueter uniquement les mots qui sont pertinents pour le domaine, seuls les mots $w^t$ associés à un concept $C^*$ maximisant $P(C^*|w^t, T^t)$ avec une probabilité supérieure à un seuil de confiance $\delta$ sont étiquetés. Ainsi, une seconde condition s'ajoute à notre système d'étiquetage robuste :

$$P(C = k^*|w^t, T^t) > \delta, \text{ avec} \quad (6.3)$$
$$k^* = \underset{k}{\operatorname{argmax}}\, P(C^t = k|w_t, T^t)$$

La détermination du modèle défini par les équations 6.2 et 6.3 revient à estimer les modèles $P(C^t|w^t)$, $P(C^t|T^t)$, les paramètres $\beta_1$ et $\beta_2$ indépendants des concepts et le seuil de confiance $\delta$.

Les paramètres $\beta_1$ et $\beta_2$ sont estimés avec l'algorithme de Newton-Raphson [Press *et al.*, 1988] que nous détaillons dans l'annexe C.1 sur un corpus d'entraînement contenant des observations $o^t = (w^t, C^t, T^t)$ où $w^t$ est le mot que nous voulons étiqueter, $C^t$ est le concept correct attribué à $w^t$ et $T^t$ est le thème du contexte de $w^i$ (l'unité thématique).

Dans ce qui suit, nous décrivons les approches utilisées pour la modélisation des probabilités $P(C^t|w^t)$ et $P(C^t|T^t)$.

### 6.5.1 Distribution des concepts étant donné les mots

Pour calculer la probabilité $P(C|w)$, nous avons utilisé une mesure de similarité inspirée de la méthode de Lesk [Lesk, 1986] qui est un baseline pour les approches de désambiguisation de sens utilisant un dictionnaire [Kilgarriff et Palmer, 2000]. Il s'agit de calculer le nombre de mots[4] lemmatisés en commun et contenus dans la définition textuelle de deux mots $w$ et $u$ pour un sens donné, tel que décrit par l'équation 6.4.

$$sim(w,u) = \frac{\mid D_w \mid \cap \mid D_u \mid}{min(\mid D_w \mid, \mid D_u \mid)} \tag{6.4}$$

$w$ et $u$ sont deux mots, $D_w$ et $D_u$ sont respectivement les ensembles de mots lemmatisés extraits de la définition textuelle de $w$ et $u$ pour un sens donné. Les définitions textuelles sont extraites de Wordsmyth.

La probabilité $P(C|w)$ est obtenue en normalisant et lissant les scores de similarité entre le mot $w$ pour un sens donné $s$ et chaque concept $C$ de l'ontologie calculés selon l'algorithme 1.

Pour tenir compte des scores de similarité par sens de mot, nous supposons que les sens d'un mot sont indépendants les uns des autres, tel que la formule par l'équation 6.5.

$$\begin{aligned} P(C|w) &= P(C|s(w) = s_1, s_i, \ldots, s_l) \\ &= \sum_{s_l \in S(w)} P(C|w)P(s_l|w) \end{aligned} \tag{6.5}$$

$s_l \in S(w)$ sont les différents sens du mot $w$ et $P(C|s_l)$ est la probabilité d'obtenir le concept $C$ étant donné le mot $s_l$.

L'algorithme 1 se base sur deux étapes que nous décrivons dans ce qui suit.

Soit :

[4]Seuls les mots de classe ouverte (adjectifs, verbes, noms) sont retenus pour le calcul du score de similarité.

- $C = \{c_1, \ldots, c_t, \ldots, c_K\}$ l'ensemble des concepts des classes principales de l'ontologie et K le nombre de ces concepts.
- $I = \{i_1, \ldots, i_l, \ldots, i_N\}$ l'ensemble des instances d'un concept $c_t$ et $N$ le nombre d'instances du concept $c_t$.
- $M = \{m_1, m_2, \ldots, m_T\}$ l'ensemble des mots synonymes et des mots similaires d'une instance $i_l$ et $T$ le nombre de ces mots.
- $S(w) = \{s_1, s_2, \ldots, s_P\}$ l'ensemble des sens d'un mot $w$ et $P$ leur nombre.

La boucle la plus imbriquée calcule le score de similarité $sim(m, s)$ entre un synonyme $m \in M$ et un sens $s \in S(w)$ selon l'équation 6.4. Le calcul des scores de similarité $sim(m, s)$ pour chaque $m \in M$ définit un vecteur de scores de similarité $v_{M,s}^{\rightarrow}$ que nous utilisons pour calculer le score de similarité $sim(i_l, s)$ entre l'instance $i_l \in I$ et $s$. Ce dernier est obtenu en prenant la médiane de $v_{M,s}^{\rightarrow}$.

La seconde étape consiste à calculer le score de similarité $sim(i_l, s)$ pour chaque instance $i_l \in I$ pour obtenir un vecteur de similarité $v_{I,s}^{\rightarrow}$. La moyenne des scores de ce vecteur détermine le score de similarité $sim(c_t, s)$ entre le concept $c_t$ et $s$.

**Algorithme 1** Algorithme pour le calcul de similarité

**Entrée:** $s \in S(w)$ un sens du mot $w$, $I$ l'ensemble des instances du concept $c_t$ et $M$ l'ensemble des mots synonymes et mots similaires d'une instance $i_l$.

1: **Pour tout** instances $i_l \in I$ **Faire**
2: **Pour tout** synonymes $m \in M$ **Faire**
3: $sim(m, s) = \frac{|D_m| \cap |D_s|}{min(|D_m|, |D_s|)}$
4: **Fin Pour**
5: $\vec{v}_{M,s} \stackrel{\text{def}}{=} (sim(m_1, s), \ldots, sim(m_T, s))$
6: $sim(s, i_l) = médiane(v_{M,s}^{\rightarrow})$
7: **Fin Pour**
8: $v_{I,s}^{\rightarrow} \stackrel{\text{def}}{=} (sim(i_1, s), \ldots, sim(i_N, s))$
9: $sim(s, C) = max(v_{I,s}^{\rightarrow})$

Le choix de la médiane pour l'identification du score de similarité entre $w$ et une instance $i$ repose sur les résultats obtenus à partir de deux expériences que nous avons effectuées : une utilisant la moyenne et la seconde utilisant la médiane. Les résultats obtenus avec la médiane étaient supérieurs à ceux obtenus avec la moyenne. Enfin, le choix du meilleur score

pour l'identification du score de similarité entre $w$ et le concept $C$ repose sur une heuristique qui donne la priorité au score de similarité le plus significatif entre $w$ et une instance de ce concept. L'algorithme 1 calcul les scores de similarité entre un mot (pour un sens donné) et les différents concepts de l'ontologie.

L'application de cet algorithme aux différents sens du mot $w$ génère un vecteur de scores de similarité par sens de mot. Afin de simplifier notre problématique, nous avons fait abstraction du problème de désambiguisation de sens de $w$ qui est une problématique à part entière ([Ide et Véronis, 1998] pour une revue de l'état de l'art).

Ainsi, nous supposons que les sens d'un mot sont équiprobables dans le contexte du domaine de la recherche et sauvetage. Ce qui se traduit par l'équation suivante :

$$P(s|w) = \frac{1}{\mid S(w) \mid}$$

Dans l'algorithme que nous avons présenté, nous calculons le score de similarité entre un mot et une instance $i_t$ d'un concept en tenant compte de l'instance et de la liste de mots synonymes et mots similaires utilisés pour enrichir l'ontologie. Nous avons proposé cet ajout à l'algorithme initial de Lesk afin de diminuer les chances d'obtenir un score nul pour deux mots sémantiquement similaires.

Toutefois, dans les cas où les mots ne sont pas sémantiquement similaires, nous obtenons un score nul et par conséquent une probabilité $P(C|w)$ nulle. Pour éviter les probabilités nulles, nous avons effectué un lissage des scores de similarité en attribuant une petite valeur aux scores nuls et avons normalisé les probabilités pour qu'elles somment à 1. La formule pour le lissage est définie par l'équation 6.6.

$$P(C|w_i^l) = \frac{sim(w_i^l, C) + \epsilon}{\sum_{t=1}^{K} sim(w_i^l, C^t) + K\epsilon} \tag{6.6}$$

$\epsilon$ est une petite valeur que nous avons pris égale à $0,01$ et $K$ est le nombre de concepts.

Ces approximations modifient l'équation 6.5 comme suit :

$$\begin{aligned} P(C|w) &= \sum_{w_i \in S(w)} P(C|w_i) P(w_i|w) \\ &= \frac{1}{\mid S(w) \mid} \sum_{w_i \in S(w)} \frac{sim(w_i, C) + \epsilon}{\sum_{t=1}^{K} sim(w_i, C^t) + K\epsilon} \end{aligned} \tag{6.7}$$

### 6.5.2 Distribution des concepts étant donné les thèmes

L'idée motivant la modélisation de la distribution des concepts étant donné un thème est de fournir une mesure de confiance sur la pertinence des concepts associés aux mots évalués. Cette étape ajoute une condition de pertinence au mot à étiqueter en supposant que le mot n'est pertinent que si le concept qui lui est associé est fréquemment observé pour le thème de son contexte.

Ainsi, le modèle $P(C|T)$ filtre les faux positifs, c'est-à-dire les mots sémantiquement similaires à un terme de l'ontologie mais qui, étant donné le thème, ne constituent pas une information pertinente.

Le tableau 6.3 illustre un exemple de faux positif (*) écarté grâce au thème. Par exemple, le verbe `land` est sémantiquement similaire au concept STATUS et constitue le prédicat de la relation pertinente (1) extraite de l'énoncé `7`. Dans l'énoncé `42`, le thème est quelconque et de ce fait, ce verbe ne constitue pas une expression pertinente.

(1) `Relation` ***`Incident`*** `: AIRCRAFT STATUS INITIAL-ALERT LOCATION`

Nous avons traduit la condition de pertinence $P(C|T)$ par la fréquence relative des concepts étant donné les thèmes. Toutefois cette association peut être problématique lorsqu'un concept n'a jamais été observé pour un thème donné. Pour contourner ce problème, nous proposons de remplacer cette probabilité par une moyenne arithmétique pondérée $P_\alpha(C|T)$ de la fréquence relative d'un concept étant donné un thème et de sa fréquence relative dans le corpus. Cette modification permet d'attribuer la fréquence du concept dans le corpus lorsque $P(C|T) = 0$. Ainsi le modèle obtenu est une combinaison linéaire des fréquences relatives tel que décrit par l'équation 6.8.

| *Thème* | No Loc Énoncé |
|---|---|
| *Incident* | 7 a : On the way to go, he (AIRCRAFT) had to land (STATUS) in emergency (INITIAL-ALERT) south east coast of Newfoundland (LOCATION) . |
| | ... |
| *Other* | 42 b : Now, the question he had was is there some place for a small helicopter (AIRCRAFT) to land (*) there |

TAB. 6.3 – Exemple de faux positif pour l'étape d'analyse sémantique. Dans l'énoncé 7, le verbe land est pertinent considérant le thème *Incident*. Dans le cas de l'énoncé 42, le thème est quelconque (*Other*). Dans ce dernier contexte, le verbe land n'est pas un mot pertinent au domaine et ne sera pas retenu pour l'étiquetage sémantique.

$$P_\alpha(C^t = k|T^t = j) = \alpha P(C^t = k) + (1 - \alpha)P(C^t = k|T^t = j) \tag{6.8}$$

$\alpha$ est le coefficient de pondération estimé par l'algorithme EM [Dempster *et al.*, 1977] que nous détaillons en annexe C.2 et les fréquences relatives sont définies par :

- $P(C = k)$ est la fréquence relative du concept $k$ dans le corpus d'entraînement :

$$P(C = k) = \frac{\#(C = k)}{\sum_{l=1}^{K} \#(C = l)}$$

  $\#(C = k)$ représente le compte du concept $k$ dans le corpus d'entraînement et $K$ le nombre de concepts issus des classes principales de l'ontologie (tableau 5.1).

- $P(C = k|T = j)$ est la fréquence relative du concept $k$ sachant le thème $j$ :

$$P(C = i|T = j) = \frac{\#(C = i, T = j)}{\sum_{l=1}^{K} \#(C = k, T = j)}$$

  $\#(C = i, T = j)$ représente le nombre de fois que le concept $i$ et le thème $j$ sont observés simultanément dans le corpus d'entraînement.

### 6.5.3 Expériences et résultats

Nous avons entraîné notre modèle d'étiquetage des expressions pertinentes sur un corpus d'entraînement constitué de 3413 mots extraits de 65 % des 64 conversations annotées manuellement avec les concepts de l'ontologie et automatiquement avec les thèmes. Les expressions annotées manuellement sont les réponses des champs de formulaires fournies par le CRDV et qui ne sont pas dans l'ontologie. Ce choix a eu pour d'obtenir un corpus constitué en majorité d'expressions adjectivales et verbales et moins de noms communs, la majorité des expressions nominales étant des noms propres couverts par l'ontologie.

Le seuil $\delta$ a été évalué sur le corpus de test à cause de la taille modeste de notre corpus.

L'évaluation des expressions étiquetées a été faite uniquement pour les unités thématiques pertinentes, c'est-à-dire celles ayant un thème différent de *Other*, puisque ce sont uniquement les unités thématiques pertinentes qui sont considérées pour l'extraction d'information et l'apprentissage des patrons d'extraction. Les mots étiquetés manuellement ont comme étiquette un concept de l'ontologie de la classe supérieure (tableau 5.1) s'ils sont proches du domaine ou un concept OTHER qui rassemble toutes les expressions qui ne sont pas pertinentes au domaine. Cela signifie qu'un même mot peut selon le thème et la similarité sémantique être étiqueté par un concept de l'ontologie ou par l'étiquette OTHER.

Nous avons évalué trois modèles : la probabilité, $P(C^t|T^t)$, des concepts étant donné les thèmes, la probabilité $P(C^t|w^t)$ de similarité et notre système d'étiquetage sémantique $P(C^t|T^t, w^t)$ sur un corpus de test composé de 1138 mots, dont 282 mots sont des réponses de champs de formulaires.

Nous avons pris le modèle $P(C^t|T^t)$ comme *baseline* pour comparer la performance du modèle basé uniquement sur les scores de similarité avec celle de notre modèle d'étiquetage.

Le tableau 6.4 donne le rappel et la précision obtenus pour le seuil $\delta$=0,35. Ce seuil est calculé de manière empirique sur le corpus de test.

Les résultats de la classification des concepts par thème $P(C^t|T^t)$ présente un taux d'erreurs de classification de 48,9 %. Ce taux important est en partie dû à la disparité de la distribution des concepts, une grand partie du corpus étant partagée entre les concepts

| | $P(C^t\|T^t)$ | | $P(C^t\|w^t)$ | | $P(C^t\|T^t, w^t)$ | |
|---|---|---|---|---|---|---|
| | Précision | Rappel | Précision | Rappel | Précision | Rappel |
| Expressions pertinentes | 51,1 % | 51,1 % | 70,6 % | 70,6 % | 86,0 % | 67,4% |

TAB. 6.4 – Classification des mots $w^t$ par rapport aux 24 concepts $C^t$ de l'ontologie. Les mots informatifs sont les réponses aux champs des formulaires d'extraction. Le résultat du modèle $P(C^t|T^t, w^t)$ est valable pour le seuil de confiance optimale $\delta = 0.35$

OTHER et STATUS.

Nous remarquons que le modèle exponentiel a une meilleure performance, avec un F-score de 75,3 %, que celle de notre *baseline* et le modèle basé uniquement sur la similarité, qui a un F-score de 70 %. Bien que le modèle basé sur les thèmes ait une faible performance de 52,1 %, il a permis d'augmenter le F-score du produit d'experts de 7,1 %.

La performance moyenne du modèle basé sur la similarité est probablement due à l'approximation faite pour passer du vecteur de scores de similarité à $P(C^t|w^t)$. Une amélioration possible est de représenter $P(C^t|w^t)$ comme une mixture de gaussiennes où chaque gaussienne est une fonction de la similarité par rapport à un concept donné. Les erreurs du modèle d'extraction des expressions pertinentes sont en partie dues aux erreurs d'étiquetage lexical causées par les irrégularités langagières de l'oral et influençant le score de similarité. Par ailleurs, à cause de la taille modeste de notre corpus d'entraînement, nous avons opté pour des paramètres $\beta_1$ et $\beta_2$ indépendants du concept. Cependant, la disparité de la distribution des concepts dans le corpus d'entraînement (le concept STATUS à lui seul représente 29,5 % du corpus d'entraînement) fait que les coefficients $\beta_1$ et $\beta_2$ sont influencés de manière à favoriser une meilleure classification du concept prédominant. Le passage vers un modèle où les paramètres dépendent du concept permettrait une meilleure performance.

Le F-score moyen incluant les résultats des deux modules d'étiquetage sémantique (section 6.4 et 6.5) est de 82,2 %. Le résultat obtenu est inférieur aux meilleurs résultats obtenus pour l'extraction des entités nommées lors de MUC 6 avec un F-score de 97 % et de MUC 7 avec un F-score de 94 %, cependant notre tâche est plus ambitieuse car elle s'attaque aux expressions sémantiquement similaires aux termes du domaine et couvre plus de classes sé-

mantiques.

## 6.6 Conclusion

Nous avons proposé une approche pour l'étiquetage des termes et expressions sémantiquement similaires représentant des arguments et prédicats potentiels pour les relations pertinentes au domaine. L'originalité de notre approche réside dans la détection et l'étiquetage sémantique des expressions pertinentes au domaine et à l'étape d'apprentissage des relations **prédicat-arguments** pertinentes. Elle présente l'avantage de diminuer la sous-représentation des données en remplaçant les mots par les classes de mots auxquelles ils sont associés.

Toutefois, cette approche présente un inconvénient lié à la granularité de notre étiquetage sémantique. En particulier, à cause de la petite taille de notre corpus, nous avons effectué un étiquetage en ne considérant que les classes principales de l'ontologie. Certaines classes sont assez spécifiques comme la classe AIRCRAFT ; par contre, d'autres classes sont trop générales, notamment la classe STATUS qui comprend les sous-classes STATUS-TASK, PROPERTIES, STATUS-TASK-RESULT, ce qui ne permet pas une évaluation précise de la performance de notre étiqueteur.

Notre approche d'étiquetage sémantique a été proposée pour les textes bruités, mais elle peut être utilisée pour les textes structurés. En particulier, elle peut être utilisée pour réduire l'effet des variations langagières sur le processus d'apprentissage des patrons pour ces textes. Également, c'est un moyen de pallier le problème des informations distantes engendré, par exemple, par des constructions présentant une apposition. Ces deux problèmes constituent des limites aux approches standard d'extraction (section 3.2.3).

# Chapitre 7

# Apprentissage de patrons d'extraction d'information

## 7.1 Introduction

Les systèmes d'EI développés pour les textes écrits utilisent des patrons d'extraction basés sur la relation **sujet-verbe-objet** adaptée à la structure des phrases. Cependant, si cela est possible pour des textes écrits où les événements et faits sont exprimés dans des formes grammaticales bien définies, cela ne l'est pas pour les textes en langage oral spontané. Deux conditions nécessaires pour ces systèmes sont violées : la grammaticalité et la proximité de l'information. Dans ce chapitre, nous présentons la dernière étape de notre approche pour l'apprentissage des schémas d'extraction. La première partie de notre approche concernait l'étape de segmentation des textes conversationnels dans le but de distinguer les unités linguistiques pour l'apprentissage des schémas d'extraction, et les unités thématiques pour identifier leur thème et établir un lien entre l'information contenue dans ces unités et les formulaires d'extraction. La deuxième étape du processus a permis l'identification des expressions pertinentes pour la tâche d'apprentissage.

Ainsi, dans cette dernière étape, nous apprenons les schémas d'extraction qui permettent d'associer une expression pertinente à un champ particulier d'un formulaire d'extraction. Plus particulièrement, nous abordons deux problématiques. La résolution des anaphores pronominales pour faire émerger une partie des relations pertinentes et garantir une meilleure couverture de ces dernières lors de l'EI (section 2.4.1). La deuxième problématique concerne l'apprentissage de schémas d'extraction représentés par des modèles de Markov à partir de séquences de concepts étiquetant les expressions pertinentes extraites contenues dans une unité linguistique.

## 7.2 Résolution des anaphores pronominales en position de sujet

L'état de l'art en résolution des anaphores pour les textes conversationnels [Byron et Stent, 1998; Eckert et Strube, 1999] tels que les dialogues fait référence à la théorie du centrage développée par Grosz *et al.* [1995]. Cette théorie établit un lien entre la cohérence d'un énoncé (*local coherence*) et le choix des référents.

Grosz a appliqué la théorie du centrage à la résolution des anaphores non pronominales telles que les groupes nominaux définis (par exemple, `ce livre`).

Nous retenons cette théorie car elle a été développée pour les textes conversationnels orientés tâche. L'approche utilise la structure hiérarchique propre aux tâches décrites dans les dialogues et associe un *centre d'attention explicite* (*explicit focus of attention*) et un *centre d'attention implicite* (*implicit focus of attention*) pour chaque segment du discours qui correspond à un sous-dialogue (partie du dialogue portant sur une sous-tâche). Chaque centre d'attention explicite est un ensemble d'entités contenues dans les énoncés du sous-dialogue, tandis que le centre d'attention implicite contient des entités sémantiquement proches de celles du centre d'attention explicite. Ainsi, parallèlement à la structure du discours est bâtie une hiérarchie de centres d'attention explicites et implicites. Cette hiérarchie est utilisée dans la recherche des antécédents pour les anaphores non pronominales. Grosz propose de

commencer la recherche de l'antécédent d'abord au niveau de l'énoncé pour traiter le cas des anaphores intra-phrastiques. Sinon, la recherche commence en sélectionnant le premier antécédent compatible avec l'anaphore parmi ceux du centre d'attention explicite ou implicite actif, c'est-à-dire du sous-dialogue courant. En cas d'échec de cette recherche, le processus est répété en parcourant successivement les centres d'attention de la hiérarchie jusqu'à la racine de l'arbre, en commençant par le centre d'attention de la partie du dialogue contenant le sous-dialogue courant.

L'approche que nous proposons est une version simplifiée de cette théorie mais qui présente des ajouts pour traiter le cas des anaphores pronominales, en particulier la définition de valeurs par défaut.

### 7.2.1 Approche

Nous nous intéressons aux anaphores pronominales `they`, `we`, `she`, `he` et `it` en position de sujet. Les antécédents recherchés sont les concepts qui représentent les principaux actants retrouvés dans les différents thèmes : AIRCRAFT, SAR-AIRCRAFT, VESSEL, SAR-VESSEL, SEARCH-UNIT-TEAM, PERSON et SAR-PERSON. Les antécédents ne sont pas les expressions d'un énoncé mais plutôt les concepts de l'ontologie générés par l'étiqueteur sémantique puisque le but est d'annoter sémantiquement les unités linguistiques pour l'étape d'apprentissage des schémas d'extraction.

L'algorithme que nous présentons s'inspire de celui de Grosz. Il utilise la structure thématique d'un texte conversationnel (les sous-dialogues sont des unités thématiques), ainsi qu'une liste contenant les concepts des expressions pertinentes d'une unité thématique (centre d'attention explicite).

À l'instar de Kameyama [1997], nous n'effectuons pas d'analyse syntaxique et exploitons uniquement la position de l'anaphore, le nombre de l'anaphore, le thème et les classes sémantiques des mots. Les simplifications et ajouts apportés dans notre algorithme sont les suivants :

1. La structure du discours est linéaire et non hiérarchique comme c'est le cas dans l'al-

gorithme présenté par Grosz. Cette simplification a été effectuée pour l'automatisation de la segmentation en unités thématiques. La conséquence de cette simplification est au niveau de l'espace de recherche de l'antécédent. Dans le cas de l'algorithme proposé par Grosz, il s'étend à une partie de l'arbre des centres d'attention qui peut inclure des segments de discours dont les thèmes sont variés. Dans notre cas, l'absence de lien hiérarchique nous limite à considérer uniquement le centre d'attention de l'unité thématique portant sur le même thème.

2. Il n'y a pas de centre d'attention implicite car nous limitons les concepts considérés pour la recherche d'un antécédent.
3. L'algorithme proposé par Grosz ne traite pas les anaphores pronominales, mais plutôt les anaphores définies (groupes nominaux définis tels que `ce bateau` ou non définis tels que `le bateau`). Donc, d'une part, nous avons ajouté des contraintes sémantiques de compatibilité avec les différents concepts candidats (tableau 7.1). D'autre part, nous avons défini des valeurs par défaut pour les anaphores qui n'ont pas d'antécédent (tableau 7.2).

La structure thématique est utilisée pour identifier des antécédents par défaut et établir un ordre de préférence (tableau 7.2) entre les candidats ambigus (tableau 7.1). Ces deux tableaux ont été remplis à partir du corpus d'étude. Le choix d'un antécédent est dirigé par deux contraintes :

**Sémantique** La distinction majeure est au niveau de la compatibilité sémantique. Par exemple, le pronom `they` ne peut faire référence à des conditions météorologiques. Par contre, en dépit de l'incompatibilité syntaxique, le pronom `she` peut faire référence à un bateau comme le montre l'extrait d'une conversation de notre corpus au tableau 7.3.

**Thématique** Cette information est destinée à fournir un antécédent par défaut, lorsqu'aucun concept compatible avec l'anaphore n'a été détecté dans les énoncés précédents de l'unité thématique courante ou de la précédente portant sur le même thème.

La procédure de résolution des anaphores pronominales en position de sujet utilise le

| Thème | Classe sémantique | she | he | they | we | it |
|---|---|---|---|---|---|---|
| *Missing-object* | AIRCRAFT | x | x | | | x |
| | SAR-AIRCRAFT | | | | x | |
| | VESSEL | x | x | | | x |
| | SAR-VESSEL | | | | x | |
| | SEARCH-UNIT-TEAM | | | x | x | |
| | PERSON | | x | | | |
| | SAR-PERSON | | | | | |
| *Incident* | AIRCRAFT | x | | | | x |
| | SAR-AIRCRAFT | | x | x | x | |
| | VESSEL | x | | | | x |
| | SAR-VESSEL | | x | x | x | |
| | SEARCH-UNIT-TEAM | | x | x | x | |
| | PERSON | | x | | | |
| | SAR-PERSON | | x | | x | |
| *Mission* | AIRCRAFT | x | | | | x |
| | SAR-AIRCRAFT | | x | x | x | |
| | VESSEL | x | | | | x |
| | SAR-VESSEL | | x | x | x | |
| | SEARCH-UNIT-TEAM | | | x | x | |
| | PERSON | | x | | | |
| | SAR-PERSON | | x | | x | |
| *Search-Unit* | AIRCRAFT | x | | | | x |
| | SAR-AIRCRAFT | | x | x | x | |
| | VESSEL | x | | | | x |
| | SAR-VESSEL | | x | x | x | |
| | SEARCH-UNIT-TEAM | | x | x | x | |
| | PERSON | | x | | | |
| | SAR-PERSON | | x | | x | |

TAB. 7.1 – Table de compatibilité des pronoms étant donné un thème $T$ et une classe sémantique $C$. Les croix indiquent les combinaisons $(C, T)$ compatibles pour la résolution des anaphores pronominales. Les combinaisons compatibles sont déduites à partir de l'analyse de 31 conversations.

| Thème | *Missing-Object* | *Incident* | *Mission* | *Search-Unit* |
|---|---|---|---|---|
| she | VESSEL | VESSEL | VESSEL | SEARCH-UNIT-TEAM |
| he | VESSEL | SEARCH-UNIT-TEAM | PERSON | SEARCH-UNIT-TEAM |
| they | SEARCH-UNIT-TEAM | SEARCH-UNIT-TEAM | SEARCH-UNIT-TEAM | SEARCH-UNIT-TEAM |
| it | VESSEL | VESSEL | - | - |
| we | SEARCH-UNIT-TEAM | SEARCH-UNIT-TEAM | SEARCH-UNIT-TEAM | SEARCH-UNIT-TEAM |

TAB. 7.2 – Table des classes sémantiques par défaut étant donné le thème et un pronom. Le "-" indique un pronom indéfini. Les valeurs par défaut sont obtenues à partir des concepts les plus fréquents pour chaque thème.

| No | Loc | Énoncé |
|---|---|---|
| 1 | a : | And the name of the boat is Jocelyn Boy. |
| 2 | b : | Jocelyn Boy. |
| 3 | a : | And she's 43-foot, blue ( INAUDIBLE ), and there's 2 people on board. |

TAB. 7.3 – Exemple d'anaphore pronominale qui viole la contrainte de compatibilité de genre. L'antécédent est un bateau comme l'indique l'énoncé 1.

tableau 7.1 contenant les combinaisons (pronom, thème, concept) compatibles et le tableau 7.2 contenant les antécédents par défaut pour un pronom sachant le thème du contexte. Les valeurs contenues dans les tableaux ont été obtenues à partir d'une analyse de 31 conversations de notre corpus. Les valeurs par défaut sont les concepts les plus fréquents pour chaque thème. Par exemple, VESSEL est le concept le plus fréquent dans les unités thématiques ayant pour thème *Missing-object*. Les étapes de la résolution sont décrites par les algorithmes 2, 3, 4 et 5 :

1. *Résolution-anaphore* (algorithme 2) teste si un mot est une anaphore en position de sujet. Le cas échéant, un antécédent (une classe sémantique) est fourni par la fonction *Chercher-antécédent*. Par contre, si le mot $w$ n'est pas une anaphore, alors le concept qui l'annote est comparé avec la liste des concepts retenus pour la tâche de résolution d'anaphores. Le cas échéant, il est ajouté à la liste des concepts du thème courant c'est-à-dire le centre d'attention explicite.
2. *Chercher-antécédent* (algorithme 3) teste la compatibilité de l'anaphore avec chaque concept du centre d'attention explicite d'un thème $T$. La recherche s'arrête lorsqu'un

concept est compatible et le retourne comme antécédent. Si la recherche échoue, un antécédent par défaut est retourné.

3. *Compatibilité* (algorithme 4) vérifie la compatibilité du concept proposé pour une anaphore. La compatibilité dépend du thème courant. La table de compatibilité est fournie au tableau 7.1.

4. *Valeur-défaut* (algorithme 5) retourne le concept par défaut lorsqu'aucun antécédent du centre d'attention explicite n'est compatible avec l'anaphore. La valeur par défaut est le concept le plus fréquent pour cette anaphore dans toutes les unités thématiques portant sur un même thème des 31 conversations de notre corpus d'étude. Les concepts par défaut sont fournis au tableau 7.2.

**Algorithme 2** *Résolution-anaphore*

**Entrée:** Le centre d'attention *explicite* $L_T$ pour chaque thème $T$, le mot $w$, et son concept $C_w$.

1: **Si** $w \in \{he, she, it, they, we\}$ et $w$ en début d'énoncé **Alors**
2: $C \leftarrow$ *chercher-antécédent(w,t,*$L_T$*)*
3: **Si** $C <>$ *null* **Alors**
4: $C_w \leftarrow C$
5: **Fin Si**
6: **Sinon Si** $C_w \in \{$AIRCRAFT, VESSEL, SEARCH-UNIT-TEAM, ...$\}$ **Alors**
7: $L_T = L_T \cup \{C_w\}$
8: **Fin Si**

**Algorithme 3** *Chercher-antécédent*

**Entrée:** Le thème $T$, le pronom $w$ et le centre d'attention explicite, $C_T$.

**Si** $L_T \neq \varnothing$ **Alors**
  **Pour tout** $C \in L_T$ **Faire**
    **Si** *Compatible(C,w,T)* **Alors**
      retourner $C$
    **Fin Si**
  **Fin Pour**
**Fin Si**
retourner *Valeur-défaut(w,t)*

**Algorithme 4** *Compatible*

**Entrée:** Le concept $C$, le pronom $w$, le thème $T$ et la table de compatibilité $Table_{\text{Compatibilité}}$ (tableau 7.1).
retourner $Table_{\text{Compatibilité}}(C, w, T)$

**Algorithme 5** *Valeur-défaut*

**Entrée:** Le pronom $w$, le thème $T$ et la table des valeurs par défaut $Table_{\text{défaut}}$ (tableau 7.2).
**Si** $Table_{\text{défaut}}(w, T)$ défini **Alors**
$C \leftarrow Table_{\text{défaut}}(w, T)$
retourner $C$
**Sinon**
retourner *null*
**Fin Si**

### 7.2.2 Expériences et résultats

L'évaluation de notre approche a été effectuée sur 31 conversations de notre corpus, soit 161 anaphores pronominales en position de sujet. Les taux d'erreurs d'annotation sont présentés au tableau 7.4.

Habituellement, les résultats de la résolution des anaphores sont calculés en terme de précision et de rappel. Toutefois, comme notre système propose toujours un antécédent (même en cas d'échec de la recherche, le module propose une valeur par défaut), les taux de précision et de rappel sont identiques et ils correspondent à un taux de résolution.

Bien que notre tâche soit plus limitée que celle de la résolution de coréférences en général, nous avons rencontré plusieurs difficultés qui expliquent notamment les performances du module pour la résolution de la coréférence du pronom `he`. Dans le cas du pronom `it`, les erreurs ont lieu avec les pronoms indéfinis.

Un problème majeur des conversations spontanées est la confusion du genre et du nombre pour faire référence aux principaux actants. Par exemple, le pronom `she` est exclusivement réservé au bateau ou à l'objet de la recherche, tandis que le pronom `he` est utilisé aussi bien pour désigner une personne qu'un bateau ou une unité de recherche. Cette confusion au niveau du genre et du nombre n'est pas problématique pour les locuteurs car ils font usage de connaissances partagées qui permettent de désambiguïser les antécédents. Cette

| | Taux de résolution |
|---|---|
| `she` | 100,0 % |
| `he` | 40,7 % |
| `they` | 95,2 % |
| `we` | 84,9 % |
| `it` | 76,8 % |
| `Moyenne` | 79,5 % |

TAB. 7.4 – Taux de résolution des anaphores pronominales en position de sujet.

connaissance mutuelle est contenue en partie dans l'information thématique qui donne un ordre de préférence aux antécédents possibles. Dans notre approche, nous nous sommes inspirés de ce mécanisme pour suggérer des antécédents en fonction du thème et des concepts mentionnés dans une unité thématique. Les résultats obtenus sont intéressants, bien que dans le cas du pronom `he` le taux d'erreur soit particulièrement élevé. Nous voyons deux raisons qui expliquent ce taux d'erreurs.

**La structure du discours** Les travaux en analyse du discours, notamment ceux de Grosz, suggèrent une structure hiérarchique du discours. Cela permet de remonter dans la hiérarchie des centres d'attention pour chercher un antécédent (section 7.2.1). Dans notre approche, nous avons supposé une segmentation linéaire du discours. Cette simplification retire une partie de l'information discursive importante (lien hiérarchique) qui permet de relier les anaphores thématiquement éloignées. Un exemple concret de ce problème tiré de notre analyse des erreurs du système est celui du pronom `he` qui apparaît dans une unité thématique *Search-unit*. Dans cette unité, le pronom fait référence à une personne (concept PERSON-TYPE). L'unité thématique suivante qui a pour thème *Missing-object* contient aussi le pronom `he` en début d'unité, cependant, n'ayant pas d'objet dans la liste des concepts de l'unité thématique, le système suggère une valeur par défaut qui dans ce cas est VESSEL et cause une erreur car le pronom `he` fait encore référence à une personne. La structure hiérarchique aurait permis de donner préséance de la dernière valeur rencontrée sur la valeur par défaut afin de garder le même antécédent, c'est-à-dire PERSON-TYPE.

**Les connaissances mutuelles** D'une part, elles ne sont pas toutes explicitées dans chaque conversation. Dans quelques conversations, les locuteurs apportent de l'information sur un événement qui a été mentionné dans une conversation précédente (section 2.2). D'autre part, les locuteurs ne s'identifient pas toujours, ce qui complique la résolution de l'anaphore pronominale `we`.

L'approche proposée présente certaines similarités avec celle présentée par Kameyama [1997]. En effet, nous avons utilisé l'unité thématique pour limiter l'espace de recherche des antécédents, tandis que Kameyama a limité l'espace de recherche aux trois phrases précédentes. Notre approche a l'avantage de définir un espace de recherche thématiquement cohérent. À l'instar de l'approche proposée par Kameyama et contrairement à l'approche proposée par Grosz, nous n'effectuons pas d'analyse syntaxique. Enfin dans les trois approches (Grosz, Kameyama et la nôtre), le critère de préférence sémantique joue un rôle important dans la sélection des antécédents.

Les résultats obtenus par l'approche de Kameyama pour la résolution des anaphores pronominales sont de 71 % de détection correcte, tandis que pour notre approche elle est de 79,5 %. Toutefois, notre problématique est beaucoup moins générale que celle de la résolution des coréférences sur des textes non spécialisés. Peu de travaux ont été effectués sur des textes conversationnels spécialisés à part ceux de Byron sur la résolution de pronoms pour des textes extraits du corpus TRAIN [Byron, 2002]. Bien que sa problématique soit plus générale que la nôtre (elle traite les anaphores pronominales et les démonstratives), nous pouvons analyser certains aspects reliés à notre approche. En particulier, elle utilise les connaissances du domaines telles que le type de l'entité et les actes de dialogues **Request**, **WH-question** et **Confirm** pour définir des contraintes sémantiques sur le choix des antécédents. Son approche résout correctement 72 % des pronoms personnels et démonstratifs.

## 7.3 Apprentissage de schémas d'extraction

Le but de cette étape est de générer automatiquement les schémas d'extraction qui seront utilisés pour la tâche d'EI. Ces schémas extraient les faits individuels à partir d'un texte

conversationnel.

Dans le reste de ce chapitre, les champs traités sont ceux prenant des faits individuels en réponse.

Les noms des champs des formulaires sont les rôles que nous voulons associer de manière automatique aux expressions pertinentes extraites à l'étape d'étiquetage sémantique. Le choix de modèles de Markov pour représenter les schémas d'extraction se justifie par :

**La polyvalence d'un concept** Un concept peut avoir des rôles différents dans un même formulaire. Bien que le thème soit un indicateur direct du type de formulaire visé, il n'y a pas toujours un lien univoque entre une entité et un rôle pour un formulaire donné. Par exemple, l'entité NUMBER peut avoir le rôle LOCATION, lorsqu'il s'agit de coordonnées (latitude, longitude), ou le rôle DATE comme dans le formulaire <*Incident*> illustré au tableau 7.9. Cette situation souligne l'importance du contexte dans les textes conversationnels.

**Les irrégularités langagières de l'oral** Elles modifient la structure d'un énoncé en introduisant du bruit qui rend l'utilisation de règles d'extraction inadéquate (section 3.5.3).

### 7.3.1 Approche

Nous proposons d'apprendre un modèle de Markov pour chaque type de formulaire. Le but est de générer des rôles étant donné les concepts observés. Ainsi, les états des modèles de Markov sont les champs définis dans chaque formulaire. La séquence optimale de rôles est déterminée par l'équation 7.1.

$$\hat{q}^{1,\ldots,n} = \underset{q^{1,\ldots,n}}{\operatorname{argmax}} \prod_{t=1}^{n} P(q^{t-1}|q^t)P(C^t|q^t) \tag{7.1}$$

$\hat{q}^{1,\ldots,n}$ est la séquence de rôles optimale étant donné une séquence de concepts $C^{1,\ldots,n}$ extraites à partir d'une unité linguistique, $n$ est la longueur de la séquence et les $q^t$ sont les états du modèle.

Les probabilités d'émission des rôles sont les fréquences relatives des entités observées sur le corpus d'étude pour chaque état tel que défini dans l'équation 7.2.

$$P(C = j|q^i) = \frac{\#(C = j, q^i)}{\sum_{l=1}^{K} \#(C = l, q^i)} \tag{7.2}$$

$C$ représente un concept extrait par l'analyseur sémantique robuste à partir d'une unité linguistique, $\#(C = j, q^i)$ le compte du concept $j$ émis par l'état $q^i$ et $K$ le vocabulaire des concepts émis.

### 7.3.2 Données en entrée

Le tableau 7.5 contient des exemples de séquences de concepts extraits à partir des unités linguistiques de la conversation Overdue boat (tableau 7.6). Les séquences représentent les entrées des modèles de Markov pour chaque formulaire.

Un premier constat se dégage à partir du tableau 7.5 : un même rôle peut être attribué plusieurs fois à différentes entités. C'est le cas du rôle LOCATION qui est attribué à trois concepts LOCATION. Deux phénomènes en sont la cause :

1. Les irrégularités langagières de l'oral telles que les répétitions et reprises, comme `on the south coast of Newfoundland`, `on the south east coast of Newfoundland`.
2. L'analyse syntaxique partielle effectuée en amont (chapitre 5). L'approche que nous avons préconisée pour contrer le problème des irrégularités langagières au niveau de l'analyse syntaxique était de limiter la couverture grammaticale à des unités minimales. Les effets secondaires de ce choix théorique se voient au niveau de l'expression `between Fortune Bay and Trepassey` pour laquelle l'analyseur syntaxique a généré deux syntagmes nominaux `between Fortune Bay` et `Trepassey`. L'analyse sémantique de ces deux syntagmes génère des concepts LOCATION qui à leur tour se voient attribués le rôle de LOCATION. La même situation se répète avec l'expression `one person on-board`, pour laquelle l'analyse syntaxique a donné deux syntagmes nominaux : `one person` et `on-board`. L'analyse sémantique de ces deux syntagmes a produit le concept PERSON

| Formulaire | Unité lexicale | Entité générée par l'étiqueteur sémantique | Rôle généré par les modèles de Markov |
|---|---|---|---|
| *Incident* | do not know | STATUS | - |
| | was handled | STATUS | - |
| | have got | STATUS | - |
| | an overdue boat | VESSEL | VESSEL |
| | on the South Coast of Newfoundland | LOCATION | LOCATION |
| | in the area | LOCATION | LOCATION |
| | between Fortune Bay | LOCATION | LOCATION |
| | Trepassey | LOCATION | LOCATION |
| *Incident* | an overdue boat | VESSEL | VESSEL |
| | the South East Coast of Newfoundland | LOCATION | LOCATION |
| *Mission* | | SEARCH-UNIT-TEAMthey | - |
| | did | STATUS | - |
| | a radar search | SEARCH-UNIT-TASK | TASK |
| | in that area | LOCATION | LOCATION |
| *Mission* | their search | SEARCH-UNIT-TASK | TASK |
| | turned | STATUS | - |
| | nothing | STATUS-TASK-RESULT | RESULT |
| *Missing-object* | an overdue vessel | VESSEL | VESSEL |
| | overdue | INCIDENT | VESSEL |
| | 20 foot open boat | PROPERTIES-DISTANCE | TYPE |
| | a Doray | VESSEL | VESSEL |
| | one person | PERSON | PERSON |
| | on-board | PART-OF-PHYSICAL-OBJECT | PERSON |
| *Incident* | started | STATUS-TASK | - |
| | 18 Zulu | TIME | DATE |
| | on, 8 | NUMBER | DATE |
| | 24 hours ago | TIME | DATE |

TAB. 7.5 – Exemple d'entrées et de sorties des modèles de Markov. La colonne 3 représente les classes sémantiques attribuées par l'étiqueteur sémantique aux unités lexicales de la colonne 1 tirée de la conversation Overdue boat. La colonne 4 représente les rôles attribués par le modèle de Markov entraîné pour un formulaire donné.

| No | Loc | Énoncé |
|---|---|---|
| 1 | a | : Maritime operation centre (ORGANISATION), (INAUDIBLE) hello. |
| 2 | b | : hi, Mr. Wellington (PERSON), it's captain Mr. VanHorn (PERSON) |
| 3 | a | : yes. |
| 4 | b | : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat (VESSEL) on the South Coast of Newfoundland (LOCATION), just in the area quite between Fortune Bay and Trepassey (LOCATION). |
| 5 | b | : it's on the south east coast of Newfoundland (LOCATION). |
| 6 | b | : this is been going on for, for 24 hours (TIME) that the case has, or almost anyway, and we had an DFO King Air (AIRCRAFT) up flying (STATUS) this morning (TIME). |
| 7 | b | : they did (STATUS) a radar search (MEANSOFDETECTION) for us in that area (LOCATION). |
| 8 | a | : yes. |
| 9 | b | : and their search (TASK) turned up (STATUS) nothing (RESULT). |
| 10 | a | : yeah. |
| 14 | b | : so I'm wondering about the possibility of attempting it with a different platform perhaps someone with even other sensors other than the radar and, in fact, someone with a, with a radar that'll be a little more sensitive. |
| 15 | b | : before I used (TASK) the Challenger (AIRCRAFT), I 'll use (TASK) a Hurk (AIRCRAFT). |
| 20 | a | : ...do you want this thing fired up (STATUS) now or you wanna wait (STATUS) till the Big Boys come in to work tomorrow morning (TIME)? |
| 21 | b | : well, I would like it if possible, I'd like them to, to be airborne (STATUS) at first light (TIME). |
| 22 | a | : ok. |

TAB. 7.6 – Conversation Overdue boat où les mots soulignés sont des réponses aux champs de formulaires. Les étiquettes sous les barres en soulignés sont des concepts de l'ontologie. Les pointillés sont les frontières des unités linguistiques générées automatiquement.

et le concept PART-OF-PHYSICAL-OBJECT qui est un concept dans la hiérarchie *part-of* de notre ontologie.

L'attribution d'un même rôle plus d'une fois a pour conséquence de générer plus d'une réponse à un champ de formulaire. La problématique de surgénération de réponses est une problématique similaire à celle étudiée en analyse syntaxique de textes de l'oral [Boufaden *et al.*, 1998; Hindle, 1983; Heeman *et al.*, 1996; Bear *et al.*, 1992]. Dans les deux cas, il s'agit de reconnaître et écarter les syntagmes ou parties de syntagme superflues introduites par les irrégularités langagières de l'oral. Cette problématique est un domaine actif de la recherche que nous retenons pour des travaux futurs.

## 7.4 Expériences et résultats

Les corpus d'entraînement sont générés de manière semi-automatique. Ce sont des séquences de concepts générés par l'étiqueteur sémantique et pour lesquels l'étape de résolution d'anaphores a été effectuée. Les balises d'unités linguistiques sont celles que nous avons manuellement ajoutées dans le corpus d'entraînement du module de segmentation linguistique (chapitre 4). Nous n'avons pas pris la sortie du segmenteur afin de mieux évaluer notre modèle sans que les erreurs des étapes précédentes interfèrent.

Les concepts d'une unité linguistique ont été manuellement annotés par le rôle correspondant dans un formulaire. Les concepts qui n'ont aucun rôle dans un formulaire donné se voient attribués une étiquette spéciale : dans le tableau 7.5, cette étiquette est représentée par le symbole "-". Nous avons effectué deux expériences afin de déterminer l'ordre du modèle de Markov qui donne les meilleures performances pour chaque schéma d'extraction. Nous avons testé un modèle de Markov d'ordre 1 et un modèle d'ordre 2. Dans le premier cas, les probabilités de transition sont données par l'équation 7.3, tandis que dans le second modèle, les probabilités de transition sont données par l'équation 7.4.

| Schémas d'extraction | Modèle de Markov | Rappel | Précision | F-score |
|---|---|---|---|---|
| *Incident* | Ordre 1 | 59,0 % | 79,6 % | |
| | Ordre 2 | 63,8 % | 85,0 % | 72,9 % |
| *Mission* | Ordre 1 | 79,0 % | 89,5 % | 83,9 % |
| | Ordre 2 | 70,7 % | 81,4 % | |
| *Search-unit* | Ordre 1 | 53,3 % | 75,2 % | |
| | Ordre 2 | 52,9 % | 76,9 % | 62.7 % |
| *Missing-object* | Ordre 1 | 54,4 % | 71,7 % | |
| | Ordre 2 | 70,8 % | 80,8 % | 75,5 % |

TAB. 7.7 – Rappel, précision et F-score de l'apprentissage des schémas d'extraction pour les formulaires *Incident*, *Mission*, *Search-unit* et *Missing-object*. Le rappel et la précision sont obtenus par la méthode de validation croisée "Leaving one out" pour les deux modèles de Markov. Le F-score est la moyenne des F-scores du meilleur modèle.

$$P(q_k|q_i,q_j) = \frac{\#(q_k,q_i,q_j)}{\sum_{q\in Q}\#(q,q_i,q_j)} \tag{7.3}$$

$$P(q_i|q_j) = \frac{\#(q_i,q_j)}{\sum_{q\in Q}\#(q,q_i)} \tag{7.4}$$

$\#(q,q_i,q_j)$ représente le nombre de fois que les états $q_i$, $q_j$ et $q_k$ se succèdent, tandis que $\#(q,q_i)$ représente le nombre de fois que les états $q_i$ et $q$ se suivent.

Étant donné la taille modeste des corpus d'entraînement (<100) pour les différents schémas d'extraction, nous avons opté pour une validation croisée avec l'approche "Leaving one out" [Ney *et al.*, 1995]. Cette approche consiste à entraîner $N$ fois le modèle sur un corpus de taille $N-1$, où à chaque étape, un exemple différent est enlevé du corpus pour former le corpus de test.

Le rappel correspond au nombre de rôles corrects générés par le système sur le nombre de rôles dans le corpus de test, tandis que la précision est le nombre de rôles corrects générés par le système sur le nombre de rôles qu'il fournit. Les F-scores des meilleures performances sont indiqués au tableau 7.7.

Le schéma d'extraction du formulaire *Mission* présente une meilleure performance avec

| Numéro | Schémas d'extraction | Probabilité |
|---|---|---|
| 1 | START-TEAM-STATUS-LOCATION-END | 0,003 |
| 2 | START-WEATHER-STATUS-END | 0,05 |
| 3 | START-TEAM-STATUS-DATE-END | 0,003 |

TAB. 7.8 – Exemples de schémas d'extraction appris avec un modèle de Markov d'ordre 1. Les états correspondent aux champs du formulaire *Search-Mission*.

le modèle de Markov d'ordre 1, tandis que les autres schémas d'extraction *Missing-object*, *Incident* et *Search-Unit* performent mieux avec les modèles d'ordre 2.

La figure 7.1 montre la topologie et les probabilités de transition entre les différents rôles du formulaire *Mission*. Le tableau 7.8 montre des exemples de patrons d'extraction appris avec leur probabilité d'observation. Les probabilités d'observer les séquences de rôles (patrons) sont obtenues en multipliant les probabilités de transitions d'un état à l'autre de la séquence.

Nous remarquons que tous les états bouclent sur eux-mêmes, ce qui cause une surgénération de réponses (section 7.3.2). Aussi, nous pouvons dégager un ensemble de dépendances fortes entre les rôles qui reflètent les types de relations pertinentes dans le corpus d'étude. Parmi celles-ci figure la relation **prédicat-arguments** TEAM-STATUS et WEATHER-STATUS. Cette dernière est utilisée pour communiquer les conditions météorologiques (section 2.3).

Le choix de l'ordre du modèle dépend essentiellement du taux de concepts polyvalents et du bruit introduit par les irrégularités langagières telles que les répétitions et reprises. D'une part, les concepts polyvalents peuvent avoir différents rôles et la désambiguïsation du rôle dépend alors du contexte représenté par les concepts précédents. Des exemples de ces entités sont NUMEX (pour les nombres), STATUS qui regroupe les verbes et adjectifs pertinents tels que `secure` et `available`, et PROPERTIES-TYPE qui regroupe des descriptions telles que les longueurs et les couleurs. Quant aux concepts non polyvalents, ils sont par exemple WEATHER-CONDITIONS et SAR-AIRCRAFT.

Dans l'unité thématique *Mission*, le concept le plus fréquent est WEATHER-CONDITIONS avec une fréquence relative de 37,7 %. Ce concept étant non polyvalent, il ne nécessite pas le

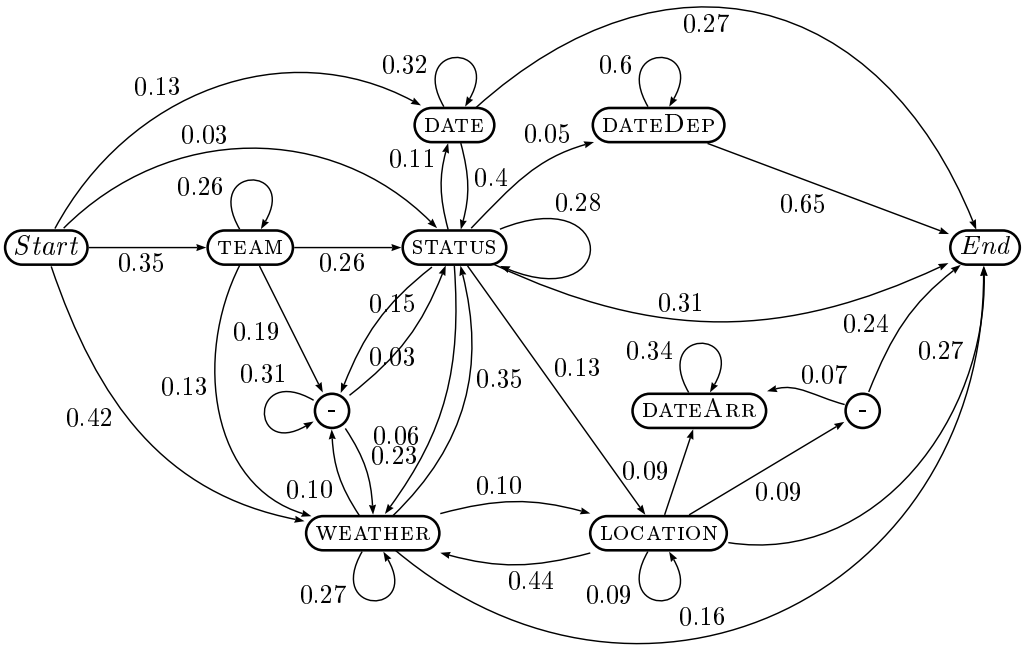


FIG. 7.1 – Modèle de Markov pour l'extraction des informations pour le formulaire Mission. Les états sont les rôles du formulaire et le symbole "-" représente un état pour les entités qui n'ont pas de rôle.

contexte pour désambiguïser son rôle dans le formulaire *Mission*. Par contre, pour les unités thématiques *Incident* et *Missing-object*, les concepts les plus fréquents sont NUMEX (14,4 %), CODE (13,4 %) et STATUS (16,5 %). Le concept NUMEX peut avoir les rôles LOCATION, CODE ou DATE. Par exemple, le nombre `8` présent dans l'expression `on 8` (énoncé `35`, tableau 1.1) peut être la réponse à un champ DATE (c'est le cas du champ *Incident*:DATE du tableau 7.9). D'autre part, à cause du bruit introduit par les irrégularités langagières, notamment les répétitions, il arrive souvent qu'un contexte de longueur 1 (un concept précédent) ne soit pas suffisant pour désambiguïser son rôle, ce qui explique la meilleure performance des modèles de Markov d'ordre 2.

## 7.5 Extraction d'information à partir d'un texte conversationnel

Afin de valider notre approche, nous avons procédé à l'extraction d'information à partir d'une conversation et avons rempli les formulaires *Missing-object*, *Search-unit* et certains champs du formulaire *Incident* qui sont des faits individuels. L'évaluation a consisté à vérifier si les réponses fournies par les modèles étaient présentes dans les réponses fournies par le CRDV. Nous avons fait abstraction des réponses en trop causées par les répétitions entre autres.

Le tableau 7.9 permet de comparer les performances des modèles de Markov par rapport à celles d'un humain. Nous avons rapporté les résultats de l'attribution des rôles obtenus par les différents modèles à un extrait de notre conversation témoin. Les formulaires concernés sont *Incident*, *Missing-object* et *Search-unit*. Les formulaires remplis ont été fournis par le CRDV. Les champs ID sont des identificateurs générés automatiquement et donc ne sont pas considérés dans l'évaluation, de même que pour le champ *Incident*:OBJECT-DESCRIPTION rempli après l'étape de coréférence et de combinaison de bribes d'information. Enfin, les champs *Search-unit*:STATUS, *Search-unit*:DETECTION-EQUIPEMENT, *Search-unit*:AVAILABILITY et *Missing-object*:CATEGORY sont des champs inférés. Concrètement, l'évaluation porte sur les champs suivants :

- *Incident*:CONTROLLER, *Incident*:LOCATION et *Incident*:TIME-OCCURRENCE.
- *Missing-object*:TYPE, *Missing-object*:ENGINE-TYPE et *Missing-object*:PERSON-ON-BOARD.
- *Search-unit*:TYPE, *Search-unit*:REGION, *Search-unit*:TYPE et *Search-unit*:RESULT.

Sur les 10 champs considérés pour l'évaluation, les modèles de Markov en ont annoté correctement 8 et fourni 9. Le rappel est de 80,0 % et la précision de 88,9 %. Les deux erreurs répertoriées sont d'origines différentes. La première est une erreur de classification : INCIDENT a été classé comme VESSEL, tandis que la deuxième erreur est une expression non

étiquetée par l'analyseur sémantique (`10-horse power upboard`).

Enfin, le champ CONTROLLER a deux réponses car à l'étape d'étiquetage sémantique, le module a repéré deux noms de personnes en début de conversation et le rôle de CONTROLLER leur a été attribué à cause de leur position en début de conversation. Ce problème de surgénération de faits individuels au niveau d'un champ est causé par diverses raisons. Par exemple, dans le cas des contrôleurs, le manque d'information sur le rôle dans le monde réel des personnes citées dans les conversations rend le choix de la personne difficile. La même situation de surgénération de faits est observée aussi pour le champ LOCATION. Cependant, la cause est reliée aux irrégularités langagières de l'oral, notamment les reprises comme dans les expressions `on the south coast of Newfoundland` et `the south east coast of Newfoundland`. L'évaluation de 11 conversations totalisant 44 champs recevant des faits individuels comme réponses a donné un rappel de 75,0 % et une précision de 83.3 %.

## 7.6 Conclusion

La conception de schémas d'extraction est une tâche cruciale pour l'EI car ce sont des filtres qui extraient les faits individuels ayant un rôle défini dans les formulaires d'extraction. Les évaluations effectuées durant les campagnes MUC font état de deux problèmes relativement à cette tâche [Sundheim, 1995] :

- La difficulté à couvrir l'ensemble des relations pertinentes, notamment à cause de la sous-représentation de certaines relations (la distribution des relations pertinentes suit la loi de Zipf).
- La limite des approches symboliques face aux cas de non-contiguïté des informations pertinentes, c'est-à-dire sujet-verbe ou verbe-objet éloignés.

Ces deux problématiques sont amplifiées dans les textes conversationnels. D'une part, le taux élevé de pronominalisation dans les unités thématiques réduit l'accessibilité aux relations pertinentes. D'autre part, la présence d'irrégularités langagières de l'oral introduit du bruit qui augmente les cas de non-contiguïté des informations pertinentes.

| Formulaires remplis par le CRDV | Formulaires remplis à partir des informations extraites par les modèles de Markov |
|---|---|
| *Incident* | *Incident* |
| ID : INCIDENT3<br>CONTROLLER : CAPITAIN VAN HORN<br><br>INITIAL ALERT : OVERDUE<br>LOCATION : SOUTH EAST COAST OF NEWFOUNDLAND,BETWEEN FORTUNE BAY AND TREPASSEY<br><br>TIME OCCURENCE : 18 ZULU, 24 HOURS AGO<br>OBJECT DESCRIPTION : OBJECT1 | ID :<br>CONTROLLER : Mr. Wellington, Mr. VanHorn Captain<br>INITIAL ALERT :<br>LOCATION : on the south coast of Newfoundland, in the area, between Fortune Bay, Trepassey, the south east coast of Newfoundland<br>TIME OCCURENCE : 18 Zulu, on 8, 24 hours ago |
| *Missing-object* | *Missing-object* |
| ID : OBJECT1<br>CATEGORY : BOAT<br>TYPE : 20-feet DORAY<br>ENGINE TYPE : 10-HORSE POWER UPBOARD<br>PERSONS ON BOARD : 1 | ID :<br>CATEGORY :<br>TYPE : 20 foot open boat, a Doray<br>ENGINE TYPE :<br><br>PERSONS ON BOARD : one person, on-board |
| *Search-Unit* | *Search-Unit* |
| ID : UNIT2<br>BRAND : KING AIR<br>ORGANISATION : DFO<br>REGION : SOUTH EAST COAST OF NEWFOUNDLAND<br>TASK : RADAR-SEARCH<br>STATUS : COMPLETED<br>RESULT : NONE<br>DETECTION EQUIPMENT : RADAR<br>AVAILABILITY : YES | ID :<br>BRAND :<br>ORGANISATION :<br>REGION : in that area<br><br>TASK : a radar search<br>STATUS :<br>RESULT : nothing<br>DETECTION EQUIPMENT :<br>AVAILABILITY : |

TAB. 7.9 – Tableau comparatif du contenu des formulaires *Incident*, *Missing-object* et *Search-Unit* remplis par un humain (colonne de droite) et à partir des informations extraites par nos schémas d'extraction (colonne de gauche).

| Schémas d'extraction | Sans résolution | Avec résolution des anaphores |
|---|---|---|
| *Incident* | 62,4 % | 72,9 % |
| *Mission* | 39,3 % | 83,9 % |
| *Missing-Object* | 62,3 % | 75,5 % |
| *Search-Unit* | 18,4 % | 75,2 % |

TAB. 7.10 – Comparaison des F-scores obtenus pour les différents schémas d'extraction appris sur un premier corpus sans résolution des anaphores pronominales en position de sujet et sur un deuxième avec résolution des anaphores pronominales.

Dans ce chapitre, nous avons proposé une approche qui tient compte de ces deux problématiques. Tout d'abord, nous avons conçu un module de résolution des anaphores pronominales en position de sujet afin de faire émerger plus de relations pertinentes. Ensuite, nous avons utilisé une approche d'apprentissage stochastique pour générer les schémas d'extraction. L'avantage des modèles de Markov réside dans la robustesse de ces modèles face au bruit introduit par les irrégularités langagières. Enfin, l'ajout d'une étape de résolution des anaphores pronominales a augmenté les performances des schémas d'extraction comme le montre le tableau 7.10.

Afin d'évaluer la performances des schémas d'extraction dans un processus d'EI, nous avons comparé les réponses fournies par nos schémas et celles annotées par un humain et fournies par le CRDV. L'évaluation a porté sur 11 conversations et a donné un F-score de 78,9 % comparativement au F-score maximal de 80 % rapporté lors de MUC-6 pour des textes structurés non spécialisés.

Notre analyse de ces résultats identifie deux sources d'erreurs. La première est une conséquence de notre approche d'analyse syntaxique partielle qui ne reconnaît pas les conjonctions de syntagmes nominaux. Cette limite a pour conséquence de favoriser la surgénération de réponses comme cela a été le cas pour l'expression `between Fortune Bay and Trepassey` pour laquelle l'analyseur syntaxique a généré deux syntagmes nominaux `between Fortune Bay` et `Trepassey` lesquels ont donné lieu à deux entités LOCATION au lieu d'une seule entité. La deuxième source d'erreurs est due à notre étiqueteur sémantique robuste et plus particulièrement le module qui reconnaît les expressions non couvertes par l'ontologie du

domaine (section 6.5). Nous avions défini un seuil de confiance au-delà duquel une expression était considérée comme sémantiquement similaire à une instance de l'ontologie et donc se voyait attribuer la classe de cette instance. Le seuil choisi a permis d'obtenir une combinaison rappel/précision optimale, cependant, lors de l'apprentissage de schémas, beaucoup d'expressions non pertinentes ont passé le seuil de confiance et ont par conséquent augmenté le nombre de concepts superflus dans nos corpus d'entraînement.

Bien que les textes soient différents et que la méthode d'évaluation que nous avons utilisée ne tienne pas compte des réponses excédentaires causées par les répétitions, nous remarquons que notre approche donne des résultats satisfaisants comparativement aux textes structurés non spécialisés. Le F-score obtenu témoigne de la faisabilité de notre approche d'EI pour les textes conversationnels spécialisés.

# Chapitre 8

# Conclusion

Nous avons proposé une approche d'EI à partir de textes conversationnels spécialisés. Dans ce cadre, nous avons divisé notre problématique en deux sous-problèmes qui touchent deux aspects différents de l'EI, à savoir :

1. L'EI à partir de textes spécialisés
2. L'EI à partir de textes conversationnels

Notre démarche a consisté en l'étude des caractéristiques linguistiques des textes spécialisés et des caractéristiques structurelles des textes conversationnels dans le but d'évaluer si l'approche standard d'EI à partir de textes structurés non spécialisés était appropriée aux conversations spécialisées qui nous intéressent.

De cette analyse, nous retenons trois particularités qui ont conduit à la proposition d'une nouvelle approche d'EI adaptée à ces textes :

1. La composante interactive. Ces textes sont des transcriptions manuelles de conversations téléphoniques spontanées (par opposition à préparées). Ils se composent de tours de parole qui peuvent être entrecoupés, engendrant une présentation fragmentaire de l'information pertinente.

De plus, ils présentent un taux importants de pronoms, notamment à cause de la pronominalisation des thèmes. Ce phénomène présent dans les textes écrits est amplifié dans les textes conversationnels.

Nous avons montré que ces deux caractéristiques ne permettaient pas l'utilisation de l'approche standard d'EI qui utilise une unité linguistique basée sur la phrase.

2. Les irrégularités langagières de l'oral modifient la structure syntaxique d'un énoncé et constituent un obstacle à la production d'une analyse syntaxique complète pour la détermination d'une relation **sujet-verbe-objet**. Nous avons montré qu'une analyse partielle était plus adéquate pour ces textes. Conséquemment, la conception des patrons d'extraction ne peut reposer uniquement sur la relation syntaxique **sujet-verbe-objet**. Une alternative au manque d'information syntaxique qui permet l'identification des rôles thématiques est l'étiquetage sémantique de ces composantes et l'identification des relations **prédicat-arguments**.
3. La présence d'expressions spécialisées du domaine qui ne sont pas uniquement des entités nommées mais aussi des noms communs, des adjectifs et des verbes. Cette particularité impose une méthode d'extraction d'entités plus générale que l'extraction des entités nommées et qui s'oriente davantage vers l'étiquetage sémantique.
4. Les expressions sémantiquement similaires aux expressions du domaine ne peuvent être prises en compte par les techniques d'extraction des entités nommées. Ces expressions nécessitent un traitement particulier qui étend l'étape d'étiquetage sémantique et constitue la composante garantissant la robustesse de cette étape.

## 8.1 Apports scientifiques et améliorations possibles

Nous avons proposé une approche d'extraction des faits individuels adaptée aux particularités des textes conversationnels spécialisés. Les différences entre notre approche et l'approche standard d'EI se situent au niveau des étapes précédant l'extraction des faits individuels : la segmentation linguistique et thématique, la détection des thèmes, l'étiquetage sémantique

robuste, la résolution des coréférences et l'apprentissage des relations **prédicat-arguments**.

### 8.1.1 Segmentation linguistique

Nous avons proposé les paires d'adjacence comme unité linguistique pour la tâche d'extraction. Cette unité remplace la phrase utilisée comme unité pour les textes structurés et présente des caractéristiques sémantiques qui tiennent compte des particularités de nos textes, telle que la dépendance des énoncés composant une paire question-réponse. Le module développé se base sur la modélisation de marques lexicales indiquant le début ou la fin d'une paire d'adjacence dans un modèle de Markov. Cette approche a permis une segmentation avec un taux de rappel de 79,4 % et une précision de 89,5 %, soit un F-score de 84,1 %.

D'autres travaux ont expérimenté la segmentation linguistique. En particulier, Stolcke et al. [Stolcke et Shriberg, 1996] ont étudié cette problématique pour faciliter l'élaboration de modèles de langue de reconnaissance automatique de la parole. Les résultats qu'ils ont obtenus sur le corpus Switchboard donnent un F-score de 76,4 %, un rappel de 85,2 % et une précision de 69,2 %. Toutefois, notons que les corpora ne sont pas les mêmes et que, dans notre cas, la segmentation fournie est prosodique, ce qui est plus avantageux que le découpage par tours de parole comme c'est le cas de l'expérimentation effectuée par Stolcke et al..

### 8.1.2 Segmentation thématique

L'approche standard d'EI établit une division claire entre les traitements indépendants du discours (extraction des faits individuels) et ceux dépendant du discours (résolution des coréférences et combinaison de l'information). Cela est moins évident pour les textes conversationnels spécialisés. Toutefois, l'intégration d'une étape de segmentation thématique dans le processus d'EI a déjà été soulevée dans quelques travaux [Manning, 1998; Crowe, 1995], cependant peu de travaux ont intégré cette étape dans le processus d'EI. En particulier, Manning a proposé d'effectuer une segmentation thématique hiérarchique des textes semi-structurés (des annonces de vente d'habitation) pour contrer la fragmentation

des informations due par exemple aux appositions. Nous avons montré que cette étape était fondamentale pour contrer la fragmentation de l'information causée par l'aspect interactif des conversations.

Nous avons développé une approche de segmentation des textes conversationnels se basant sur une combinaison des marques lexicales, syntaxiques et discursives qui caractérisent les changements de thème ou la cohésion thématique. Ces marques ont été utilisées comme observations pour l'entraînement d'un modèle de Markov qui détermine si un énoncé constitue un changement de thème ou non.

Les principaux apports de notre approche sont l'association de la notion de thème aux événements visés par le processus d'extraction et couverts par les formulaires d'extraction et l'intégration d'une étape de segmentation thématique dans le processus d'EI, ce qui, à notre connaissance, n'a pas été réalisé dans d'autres travaux portant sur les textes conversationnels. Les travaux en segmentation thématique ont surtout été consacrés aux textes structurés. Dans ce cadre, les meilleurs scores cités dans la littérature sont ceux obtenus par le système *TextTiling* avec un rappel de 61 % et une précision de 66 % obtenus sur des articles scientifiques. Avec notre approche, nous détectons les changements de thème (classes `TC`) avec un rappel de 61,4 % et une précision de 67,3 %.

L'étiquetage thématique pourrait être amélioré en intégrant les pronoms en position de sujet comme marques de cohésion. Ces derniers sont souvent le résultat de la pronominalisation du thème qui est une marque de cohésion.

### 8.1.3 Identification des thèmes

L'identification des thèmes est surtout associée à la classification de documents par thème comme dans le cadre des conférences Topic Tracking and Detection (TDT)[1]. Ce que nous avons proposé est une identification à une échelle réduite au niveau des unités thématiques d'un document. Cette tâche est plus difficile que celle de la classification des documents compte tenu de la granularité des unités considérées pour détecter les changements de thème.

[1] `http://www.nist.gov/speech/tests/tdt/`

Toutefois, dans la mesure où nous n'avons pas connaissance de travaux similaires au nôtre, nous ne pouvons que situer notre travail par rapport à ceux effectués en classification de documents. Bigi et al. [Bigi *et al.*, 2001] ont présenté une étude comparative de quelques approches utilisées en identification de thème pour les textes journalistiques et les courriels. Cette étude montre que le meilleur F-score obtenu pour les textes journalistiques est de 82 %, tandis que pour les courriels il est de 67,5 %. Dans les deux cas, les approches utilisent des mots clé représentatifs des thèmes. Dans notre modèle, nous avons exploité la cooccurrence des concepts étiquetant les termes de l'ontologie et nous avons obtenu un F-score de 81,7 %. Bien que les tâches soient relativement différentes, ce résultat montre l'intérêt de l'utilisation des classes de mots pour l'identification des thèmes.

### 8.1.4 Étiquetage sémantique robuste

Cette étape représente le coeur de notre approche et la différencie de celles développées lors des MUC. Nos contributions se situent à deux niveaux : la conception d'une ontologie du domaine et l'étiquetage sémantique des expressions pertinentes au domaine. L'avantage de l'intégration d'une ontologie du domaine dans le processus d'EI a été récemment souligné dans plusieurs travaux dans le cadre de la conférence EUROLAN'03[2]. Les ontologies fournissent une représentation structurée des connaissances du domaine qui facilite la tâche d'étiquetage sémantique, étape fondamentale pour une meilleure modélisation du contenu d'un texte. Nous avons utilisé l'ontologie du domaine pour l'étiquetage sémantique des termes couverts par l'ontologie, mais aussi pour calculer des probabilités de similarité entre des expressions et les concepts de l'ontologie. Cette procédure a permis l'étiquetage des expressions sémantiquement similaires aux termes du domaine pour une meilleure couverture des informations pertinentes.

Pour effectuer l'étiquetage sémantique des expressions pertinentes au domaine, nous avons divisé cette problématique en deux parties : l'étiquetage des termes du domaine couvert par l'ontologie et l'étiquetage des expressions pertinentes non couvertes par l'ontologie. Le

[2] `http://www.racai.ro/EUROLAN-2003/html/workshop/cfp-Onto-IE.pdf`

premier module similaire à la tâche d'extraction des entités nommées reconnaît les termes du domaine avec un F-score de 89,8 %. Le F-score obtenu est légèrement en dessous des meilleures performances obtenues lors de MUC-6, qui étaient de 96.4 %. Toutefois, considérant les erreurs morphosyntaxiques causées par les irrégularités langagières et la variété des classes sémantiques couvertes, ces résultats sont probants et montre la faisabilité de cette tâche pour les textes conversationnels.

Le F-score de 75,3 % obtenu pour l'étiquetage des expressions sémantiquement similaires aux termes du domaine est plus modeste à cause des simplifications que nous avons effectuées dans notre modèle. Pour calculer les probabilités de similarité, nous devions tenir compte du problème de désambiguïsation de sens. Toutefois, pour simplifier notre approche, nous avons supposé l'équiprobabilité des sens d'un mot. Ensuite, le passage des scores de similarité vers les probabilités de similarité a été fait de manière relativement *ad hoc* dans la mesure où ces probabilités n'ont pas été tirées d'une distribution calculée.

Cependant, quelques améliorations peuvent être apportées à notre modèle statistique. La première est l'utilisation de l'approche de Lesk [Lesk, 1986] pour la désambiguïsation de sens. Cette modification peut être simplement intégrée dans notre module puisque notre algorithme utilise une approche similaire pour le calcul de similarité. La deuxième plus laborieuse, consiste à estimer les probabilités de similarités avec un modèle statistique, par exemple une gaussienne modélisant la probabilité de similarité entre les mots et un concept donné.

### 8.1.5 Apprentissage des patrons d'extraction

Nous avons modélisé les patrons d'extraction par des modèles de Markov qui associent des rôles aux arguments des prédicats. La problématique de l'apprentissage de l'attribution de rôles sémantiques est récente et a fait l'objet de quelques travaux en rapport avec l'EI. De ces travaux, nous retenons ceux de Gildea et Palmer [2002] et de Surdeanu *et al.* [2003] effectués sur des textes journalistiques. Bien que ces textes soient moins complexes que les textes conversationnels, ces travaux situent les performances de notre modèle.

Gildea et Palmer utilisent une approche basée sur les modèles de Markov pour l'affectation de rôles sémantiques. Le modèle a été entraîné sur le corpus PropBank [Kingsbury et Palmer, 2002] qui est une collection de textes journalistiques annotés avec les rôles sémantiques des arguments de prédicat. Les résultats rapportés montrent un F-score de 82 % comparativement à 83,7 % obtenu par l'approche proposée par Surdeanu *et al.* et basée sur les arbres de décision.

Notre modèle statistique se compare à celui proposé par Gildea et Palmer avec un F-score de 78,9 %. Toutefois, ce F-score ne tient pas compte des informations en trop causées par les répétitions. La prise en considération de ces informations pour le calcul du F-score diminuerait considérablement la précision de notre modèle.

Plusieurs améliorations peuvent être apportées à notre modèle : une qui nous paraît importante est la gestion des paires question-réponse représentant 66,3 % des paires d'adjacence (section 4.2.1). Une étape de classification des questions (celles nécessitant une réponse oui/non, par opposition, à celles nécesitant une réponse avec contenu propositionnel), faciliterait la réorganisation des informations présentes dans la question et la réponse pour, ultimement fournir une unité linguistique dans un format affirmatif plutôt qu'interrogatif.

Également, nous avons ajouté une étape de résolution des anaphores pronominales en amont de l'étape d'apprentissage de patrons. Notre approche a permis un taux de résolution des anaphores de 79,5 % améliorant ainsi le F-score moyen pour l'apprentissage de patrons de 68,6 %. Quelques travaux [Surdeanu et Harabagiu, 2002] ont utilisé une approche similaire pour améliorer l'extraction des faits consistant à résoudre les coréférences aux entités nommées. Dans ces travaux, Surdeanu et Harabagiu [2002] rapporte un F-score de 83 % pour l'extraction des informations des formulaires *Scenario* de MUC-6 (section 3.2.3) comparativement au meilleur score obtenu qui était de 57 %.

Finalement, une amélioration possible sur le plan général de notre approche d'EI consiste à étudier l'organisation des buts informatifs dans les conversations pour dégager une structure du discours pour le domaine de la recherche et sauvetage. L'analyse du discours est une étape importante de l'extraction d'information qui intervient, notamment, au niveau de la

résolution des coréférences. Dans notre cas, cette structure pourrait aussi être utilisée pour la reconstitution des informations manquantes à partir des conversations complémentaires portant sur un même incident.

## 8.2 Portabilité de notre approche

La portabilité d'un système d'EI vers d'autres domaines d'application est un enjeu important de l'EI. Le goulot d'étranglement pour ces applications se situe essentiellement au niveau de l'encodage des connaissances du domaine et au niveau du développement des patrons d'extraction.

Nous évaluons la portabilité de notre approche d'apprentissage de patrons d'extraction en considérant l'effort à développer pour adapter notre approche du domaine de la recherche et sauvetage maritime vers un autre domaine et en évaluant les changements nécessaires pour l'appliquer à des textes structurés. Les modules développés dans les différentes étapes du processus de conception des patrons d'extraction reposent essentiellement sur une approche d'apprentissage supervisé et par conséquent héritent de ses inconvénients, en particulier la nécessité d'annoter manuellement un corpus d'entraînement. Bien que cela soit moins laborieux que le codage manuel de patrons d'extraction, un effort substantiel doit être consacré à l'annotation sémantique du corpus pour l'entraînement de l'étiqueteur sémantique robuste. L'annotation des rôles pour l'apprentissage des relations **prédicat-arguments** nécessite un effort moindre car l'étape d'étiquetage sémantique robuste effectue une première passe en identifiant les expressions potentiellement pertinentes et en suggérant un concept comme étiquette.

Les approches supervisées développées pour la segmentation linguistique et thématique nécessitent, certes l'annotation d'un corpus d'entraînement, toutefois, elles ne dépendent pas du domaine et peuvent donc être réutilisées telles quelles pour des conversations portant sur d'autres sujets.

Une amélioration possible peut être apportée au niveau de l'étiquetage sémantique des expressions sémantiquement similaires aux termes de l'ontologie. Cette étape nécessite l'étiquetage de ces expressions avec les concepts de l'ontologie.

Pour éviter cette étape d'annotation, il serait intéressant d'évaluer la performance de l'étiqueteur entraîné sur un corpus contenant uniquement les annotations sémantiques des termes couverts par l'ontologie. Cette solution est envisageable compte tenu du F-score de ce module, 89,8 %. Nous pourrions ensuite ajuster le seuil de confiance sur un corpus de validation composé uniquement d'expressions non couvertes par l'ontologie. Ces modifications changeraient l'approche actuelle en une approche d'apprentissage non supervisée facilement portable à d'autres domaines d'application. Finalement, en supposant que les performances de l'étiqueteur sémantique robuste soient acceptables avec une approche non supervisée, le seul effort à fournir pour appliquer notre approche d'apprentissage à d'autres domaines se situerait au niveau de l'étiquetage des rôles sémantiques.

Par ailleurs, en tenant compte de l'effort d'annotation nécessaire pour l'apprentissage des rôles, l'application de notre approche à des textes structurés nécessite très peu de modifications. Ces dernières se situeraient essentiellement au niveau de la segmentation des textes. Précisément, la segmentation linguistique n'a plus lieu d'être puisque les phrases sont des unités syntaxico-sémantiques maximales. Par contre, la segmentation thématique reste une étape importante qui permet de gérer le problème des informations distantes, comme souligné par Manning [1998] et Crowe [1995]. Toutefois, dans la mesure où nous avons développé une approche de segmentation propre aux textes conversationnels, cette dernière devrait être remplacée par son homologue pour les textes écrits, par exemple l'approche développée par Hearst [Hearst, 1994] pour le système TextTiling. Enfin, notre approche devrait être plus performante pour ces textes puisque ces derniers ne présentent pas d'irrégularités langagières de l'oral.

## 8.3 Travaux futurs

L'EI est essentielle à plusieurs applications de TAL nécessitant la compréhension des textes. Elle peut être utilisée dans un système de recherche d'information pour sélectionner les documents pertinents étant donné une requête particulière, dans des systèmes de question-réponse pour trouver dans un texte la réponse à une question ou dans les systèmes de résumé automatique pour condenser des textes. Plusieurs scénarios d'exploitation ou de continuation de ce travail nous semblent intéressantes.

### 8.3.1 Systèmes de question-réponse

Les systèmes de question-réponse pour des domaines spécialisés tels que la biomédecine constituent un axe de recherche en pleine expansion où notre approche d'EI pourrait avantageusement être exploitée. Nous nous intéressons au domaine de la biomédecine, en particulier à cause de la disponibilité de ressources telles que l'ontologie GeneOntology[3], nécessaires pour notre approche.

Les approches actuelles limitent les questions à celles dont les réponses sont factuelles (qui, quand, quoi et où). Elles se basent sur un codage manuel d'expressions régulières pour ces types de question. Ceci contraint l'utilisateur à poser des questions simples et courtes dont les réponses figurent telles quelles dans le texte.

Nous proposons d'exploiter nos travaux en intégrant un système d'EI basé sur notre approche et adapté au domaine de la biomédecine. Ce projet peut être réparti en trois étapes comprenant la recherche de documents pertinents étant donné la question d'un utilisateur, l'extraction d'information à partir des documents trouvés et la recherche d'une réponse à la question à partir des informations extraites par le système d'EI.

La première étape peut être réalisée en utilisant un moteur de recherche existant tel qu'OKAPI [Robertson et Walker, 1999] pour collecter les documents pertinents.

La deuxième étape est la plus importante et se décompose en quatre tâches touchant essentiellement l'adaptation de notre approche au domaine de la biomédecine. Tout d'abord,

[3] `http://www.geneontology.org/`

nous devons définir des formulaires d'extraction qui couvrent les classes d'information pertinentes au domaine. Cela peut se faire en collaboration avec un expert du domaine. Ensuite, nous devons concevoir une interface qui convertit les informations présentes dans l'ontologie GeneOntology en un format utilisable par notre étiqueteur sémantique.

Finalement, l'étape d'apprentissage de patrons nécessite l'étiquetage d'un corpus de textes du domaine avec les rôles. Ces textes peuvent être des articles scientifiques extraits de la bibliothèque électronique Medline[4]. À partir des textes annotés, nous pouvons entraîner nos modèles de Markov pour générer les patrons d'extraction qui vont couvrir les informations pertinentes qui répondent aux champs des formulaires définis. La quatrième tâche est le développement d'un module pour la résolution des coréférences afin de compléter le développement du système d'EI.

La troisième partie du projet concerne la recherche de la réponse à une question donnée à partir des formulaires remplis par notre système.

Dans ce projet, nous voyons plusieurs défis dont le plus important est l'expérimentation de la portabilité de notre approche vers un autre domaine que celui de la recherche et sauvetage maritime. Le second réside dans l'évaluation de l'effet de l'intégration de la technologie d'EI dans un système de question-réponse. Ce projet est intéressant puisqu'à notre connaissance, aucun système de question-réponse n'a intégré un processus d'EI complet non limité à l'intégration de l'extraction d'entités nommées [Harabagiu *et al.*, 2000]. Le système qui se rapproche le plus du projet proposé est celui de Grishman *et al.* [2002] qui a introduit un système d'EI dans un moteur de recherche de rapports biomédicaux. Les résultats rapportés montrent une nette amélioration de la recherche des documents.

[4] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi`

### 8.3.2 Génération automatique de rapports à partir de transcriptions de conversations téléphoniques

Une continuation naturelle de notre thèse est la génération de rapports à partir des formulaires d'extraction remplis par un système d'EI développé selon notre approche. La génération de rapports consiste à convertir une représentation structurée des données en une représentation textuelle en langage naturel. Ce projet présente un intérêt particulier dans le domaine de la recherche et sauvetage dans la mesure où une des tâches fastidieuse effectuées par les contrôleurs des centres de coordination de sauvetage est la rédaction d'un rapport qui résume l'ensemble des informations collectées durant la résolution d'un cas de recherche et sauvetage. L'automatisation de la génération des rapports permettrait aux contrôleurs de traiter plus de cas et de mieux exploiter leurs compétences.

Nous voyons trois parties dans ce projet qui touchent trois secteurs du TAL. D'abord la reconnaissance de la parole conversationnelle pour la génération de transcriptions des conversations téléphoniques. Bien que la reconnaissance de ce type de parole soit plus compliquée que celle de monologue , une nette amélioration de la performance a été observée avec un taux d'erreurs d'environ 30 %, comparativement à 90 % aux débuts des années 90 [Pallett, 2003]. Toutefois, dans la mesure où la recherche et sauvetage est un domaine spécialisé avec des textes écrits dans un sous-langage, de meilleurs résultats peuvent être obtenus à l'aide d'un modèle de langue modélisant les termes spécialisés du domaine. La deuxième étape met en application l'approche d'EI que nous avons proposée dans cette thèse. Bien qu'une importante partie de cette tâche ait été développée dans cette thèse, nous distinguons trois parties à étudier.

La première consiste à rassembler les différents modules développés dans cette thèse et de compléter l'étape de résolution des coréférences des événements (section 3.6) pour avoir un système d'EI complet.

La deuxième partie concerne l'élaboration d'une stratégie pour le traitement de la surgénération de réponses due aux irrégularités de l'oral. Cela implique une analyse du contenu des réponses afin de distinguer les répétitions. Une heuristique simple ayant été utilisée en

analyse syntaxique des transcriptions de dialogues consiste à retenir uniquement la dernière réponse extraite à partir d'une unité linguistique [Boufaden, 1998].

La troisième tâche serait d'évaluer l'effet des erreurs présentes dans les transcriptions automatiques de conversations sur les performances de notre système d'EI.

La troisième partie du projet porte sur la génération de rapports à partir des formulaires d'extraction remplis par notre système d'EI. Cette problématique a déjà été expérimentée sur des textes structurés tirés de MUC-6 [Cancedda, 1999]. Bien que les textes sources soient différents, la méthodologie utilisée pour la génération est identique dès lors que les formulaires sont débarrassés des répétitions. Le principal défi de cette étape réside plutôt dans l'utilisation du sous-langage propre au domaine de la recherche et sauvetage.

Les exemples de projets que nous venons de décrire témoignent du potentiel de l'EI en général et plus particulièrement pour les textes conversationnels spécialisés.

Ce travail entre dans le cadre de l'EI, mais cette technologie repose sur un pipeline d'autres applications telles que l'analyse syntaxique. Nous avons proposé des solutions pour différents niveaux du traitement automatique des textes conversationnels spécialisés, en particulier pour la segmentation et l'étiquetage sémantique. Ce travail est une contribution à l'application du TAL aux textes conversationnels considérés parmi les textes les plus complexes. À grande échelle, l'industrie de la téléphonie pourrait être une des premières bénéficiaires de cet avancement.

# Annexe A

# Liste des termes de l'ontologie du domaine de la recherche et sauvetage

## A.1 Description des termes de l'ontologie du domaine de la recherche et sauvetage

Les instances sont représentées selon la nomenclature suivante : NomDeClasse(ListeDeToken,CatégorieMorphoSyntaxique,[ÉtiquetteSémantique,ListeAttributs]).

```
%***********************************
% Instances du champ direction de la classe
%location
%***********************************
position(['position'],'head',['POSITION']).
position(['latitude'],'head',['POSITION-LAT']).
position(['longitude'],'head',['POSITION-LONG']).
position(['lat'],'PROPER-NOUN',['POSITION-LAT']).
```

```
position(['long'],'PROPER-NOUN',['POSITION-LONG']).
direction(['direction'],'head',['DIRECTION']).
direction(['east'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['west'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['north'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['south'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['southeast'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['southwest'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['northeast'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['northwest'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['eastern'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['southern'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['western'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['northern'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['wetecker'],'PROPER-NOUN',['DIRECTION-TYPE']).
direction(['Wetecker'],'PROPER-NOUN',['DIRECTION-TYPE']).

%***********************************
%Instance des champs de la classe province
%
%***********************************

province(['nova','scotia'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['new','brunswick'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['quebec'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['ontario'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['manitoba'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['saskatchewan'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['alberta'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['british','colombia'],'PROPER-NOUN',['LOCATION-TYPE']).
```

```
province(['yukon'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['nunavut'],'PROPER-NOUN',['LOCATION-TYPE']).
province(['newfoundland'],'PROPER-NOUN',['LOCATION-TYPE']).
%***********************************
%Instance des champs de la classe Town
%
%***********************************

town(['trepassey'],'PROPER-NOUN',['LOCATION-TYPE','new foundland']).
town(['fortune','bay'],'PROPER-NOUN',['LOCATION-TYPE','new foundland']).
town(['halifax'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['st-johns'],'PROPER-NOUN',['LOCATION-TYPE','new foundland']).
town(['st-albans'],'PROPER-NOUN',['LOCATION-TYPE','new foundland']).
town(['hungry','grove','pond'],'PROPER-NOUN',['LOCATION-TYPE','new foundland']).
town(['quebec'],'PROPER-NOUN',['LOCATION-TYPE','quebec']).
town(['st-antoine'],'PROPER-NOUN',['LOCATION-TYPE','new brunswick']).
town(['sophie','hill'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['gander'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['cape','st-melanie'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['cape','ste-melanie'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['st-bride'],'PROPER-NOUN',['LOCATION-TYPE','new foundland']).
town(['dartmouth'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['mainguy','cove'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['boston'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['bangor'],'PROPER-NOUN',['LOCATION-TYPE','saskatchewan']).
town(['pascal','cove'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['baie','ste-anne'],'PROPER-NOUN',['LOCATION-TYPE','new brunswick']).
town(['baie','st-anne'],'PROPER-NOUN',['LOCATION-TYPE','new brunswick']).
town(['burnt','church'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['yarmouth'],'PROPER-NOUN',['LOCATION-TYPE','novia scotia']).
```

```
town(['greenwood'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['windsor'],'PROPER-NOUN',['LOCATION-TYPE','ontario']).
town(['sophist','bright'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['richibouctou'],'PROPER-NOUN',['LOCATION-TYPE','new brunswick']).
town(['shelboure'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['province','town'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['maryland'],'PROPER-NOUN',['LOCATION-TYPE','quebec']).
town(['st-peters'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['labrie'],'PROPER-NOUN',['LOCATION-TYPE']).
town(['pennant'],'PROPER-NOUN',['LOCATION-TYPE','saskatshewan']).
town(['penant'],'PROPER-NOUN',['LOCATION-TYPE','saskatshewan']).
town(['point','sapin'],'PROPER-NOUN',['LOCATION-TYPE','new brunswick']).
town(['bight'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['miramichi'],'PROPER-NOUN',['LOCATION-TYPE','new brunswick']).
town(['escouminac'],'PROPER-NOUN',['LOCATION-TYPE','quebec']).
town(['fairland'],'PROPER-NOUN',['LOCATION-TYPE','nova scotia']).
town(['birch','point'],'PROPER-NOUN',['LOCATION-TYPE','manitoba']).
town(['birch','pond'],'PROPER-NOUN',['LOCATION-TYPE','manitoba']).
town(['longueil'],'PROPER-NOUN',['LOCATION-TYPE','quebec']).
town(['ste-margaret'],'PROPER-NOUN',['LOCATION-TYPE','quebec']).
town(['st-margaret'],'PROPER-NOUN',['LOCATION-TYPE','quebec']).

%************************************
%Signature du champ
%description de la classe region
%************************************

region(['fox','island'],'PROPER-NOUN',['LOCATION-TYPE']).
region(['portage','island'],'PROPER-NOUN',['LOCATION-TYPE']).
region(['birch','pond'],'PROPER-NOUN',['LOCATION-TYPE']).
```

```
region(['mcnab','island'],'PROPER-NOUN',['LOCATION-TYPE']).
region(['portage','lake'],'PROPER-NOUN',['LOCATION-TYPE']).
region(['osbourne','bay'],'PROPER-NOUN',['LOCATION-TYPE','ontario']).
region(['osbourne','head'],'PROPER-NOUN',['LOCATION-TYPE','ontario']).
region(['osbourne','point'],'PROPER-NOUN',['LOCATION-TYPE','ontario']).
region(['water'],'head',['LOCATION-TYPE']).
region(['shore'],'head',['LOCATION-TYPE']).
region(['shore'|_],'PROPER-NOUN',['LOCATION-TYPE']).
region(['channel'],'head',['LOCATION-TYPE']).
region(['coast'],'head',['LOCATION-TYPE']).
region(['gulf'],'head',['LOCATION-TYPE']).
region(['bay'],'head',['LOCATION-TYPE']).
region(['peninsula'],'head',['LOCATION-TYPE']).
region(['cape'],'head',['LOCATION-TYPE']).
region(['cap'],'head',['LOCATION-TYPE']).
region(['lake'],'head',['LOCATION-TYPE']).
region(['mountain'],'head',['LOCATION-TYPE']).
region(['scene'],'head',['LOCATION']).

%************************************
%signature de area
%description de la classe area
%************************************

area(['delgada','area'],'PROPER-NOUN',['LOCATION-TYPE']).
area(['delgada'],'PROPER-NOUN',['LOCATION-TYPE']).
area(['madrid','area'],'PROPER-NOUN',['LOCATION-TYPE']).
area(['madrid'],'PROPER-NOUN',['LOCATION-TYPE']).
area(['area'],'head',['LOCATION']).
```

```
 %***********************************
 %Description de la classe Seas de position
 %*************************
 seas(['atlantic'],'PROPER-NOUN',['LOCATION-TYPE']).
 seas(['pacific'],'PROPER-NOUN',['LOCATION-TYPE']).

%***********************************
%
%description de la classe aircraft
%***********************************
aircraft(['speedair'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['aurora'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['king','air'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['kingers'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','DFO']).
aircraft(['challenger'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['labrador'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['lab'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['Hurk'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['hercule'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['jet','ranger'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['heavy','jacks'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['tower','lab'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['herc'],'PROPER-NOUN',['AIRCRAFT-SAR-TYPE','SAR']).
aircraft(['plane'],'head',['AIRCRAFT-TYPE','NULL']).
aircraft(['airplane'],'head',['AIRCRAFT','NULL']).
aircraft(['aircraft'],'head',['AIRCRAFT','NULL']).
aircraft(['aircrafts'],'head',['AIRCRAFT','NULL']). %ca c est parce que le dictio ne reconnai
aircraft(['helicopter'],'head',['AIRCRAFT','NULL']).
aircraft(['a-310'],'PROPER-NOUN',['AIRCRAFT-TYPE','particular']).
```

```
%***********************************
%
%description de la classe vessel
%***********************************

vessel(['sea','king'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['talbot'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['auxiliary','vessel'],'PROPER-NOUN',['VESSEL-CGA-TYPE','SAR']).
vessel(['auxiliary'],'PROPER-NOUN',['VESSEL-CGA-TYPE','SAR']).
vessel(['southwest','harbor'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['coast','harbor'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['zodiac'],'head',['VESSEL-SAR-TYPE']).
vessel(['harbor'],'head',['VESSEL-SAR-TYPE','SAR']).
vessel(['canada','sea'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['shipaggan'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['cga'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['tracadie'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).

vessel(['fishing','boat'],'PROPER-NOUN',['VESSEL-TYPE']).
vessel(['fishing','vessel'],'PROPER-NOUN',['VESSEL-TYPE']).
vessel(['fish','boat'],'PROPER-NOUN',['VESSEL-TYPE']).
vessel(['fish','vessel'],'PROPER-NOUN',['VESSEL-TYPE']).
vessel(['boat'],'head',['VESSEL']).
vessel(['ship'],'head',['VESSEL']).
vessel(['yacht'],'head',['VESSEL']).
vessel(['lobster'],'head',['VESSEL-TYPE','commercial']).
vessel(['sail','vessel'],'PROPER-NOUN',['VESSEL-TYPE']).
vessel(['sailing','vessel'],'PROPER-NOUN',['VESSEL-TYPE']).
vessel(['tanker'],'head',['VESSEL-TYPE']).
vessel(['doray'],'PROPER-NOUN',['VESSEL-TYPE']).
```

```
vessel(['fisher'],'head',['VESSEL-TYPE','commercial']).
vessel(['vessel'],'head',['VESSEL']).
vessel(['seal'],'head',['VESSEL-TYPE','commercial']).
vessel(['shrimp'],'head',['VESSEL-TYPE','commercial']).
vessel(['crab'],'head',['VESSEL-TYPE','commercial']).
vessel(['groundfish'],'head',['VESSEL-TYPE','commercial']).
vessel(['scallop'],'head',['VESSEL-TYPE','commercial']).
vessel(['jakman'],'PROPER-NOUN',['VESSEL-SAR-TYPE','SAR']).
vessel(['footer'],'head',['VESSEL']).
vessel(['footer'],'modif',['VESSEL']).


%************************************
%
%description de la classe person
%************************************

person(['duty','officer'],'PROPER-NOUN',['PERSON-SAR-TYPE','SAR']).
person(['captain'],'head',['GRADE','SAR']).
person(['officer'],'head',['PERSON-SAR-TYPE','RCMP']).
person(['major'],'head',['GRADE','SAR']).
person(['duty','watch','officer'],'PROPER-NOUN',['PERSON-SAR-TYPE','SAR']).
person(['marine','controller'],'PROPER-NOUN',['PERSON-SAR-TYPE','SAR']).
person(['commander'],'head',['GRADE','SAR']).
person(['caporal'],'head',['GRADE','SAR']).
person(['flight','controller'],'PROPER-NOUN',['PERSON-TYPE','SAR']).
person(['air','controller'],'PROPER-NOUN',['PERSON-TYPE','SAR']).
person(['commissioner'],'head',['PERSON-TYPE','SAR']).
person(['second','lieutenant'],'PROPER-NOUN',['GRADE','SAR']).
person(['assistant','air','controller'],'PROPER-NOUN',['PERSON-TYPE','SAR']).
```

```
person(['rcp','officer'],'PROPER-NOUN',['PERSON-SAR-TYPE','RCP']).
person(['sergeant'],'head',['GRADE','SAR']).
person(['coast','guard','officer'],'PROPER-NOUN',['PERSON-SAR-TYPE','SAR']).
person(['officer'],'head',['GRADE','SAR']).
person(['watch','director'],'PROPER-NOUN',['PERSON-SAR-TYPE','SAR']).
person(['fishery','officer'],'PROPER-NOUN',['PERSON-SAR-TYPE','DFO']).
person(['fishery','guardian'],'PROPER-NOUN',['PERSON-SAR-TYPE','DFO']).
person(['fisherman'],'head',['PERSON-SAR-TYPE','DFO']).
person(['osc'],'PROPER-NOUN',['PERSON-SAR-TYPE','on-scene commander']).
person(['pilot'],'head',['PERSON-TYPE']).
person(['citizen'],'head',['PERSON']).
person(['civilian'],'head',['PERSON']).
person(['people'],'head',['PERSON']).
person(['person'],'head',['PERSON']).
person(['soul'],'head',['PERSON']).
person(['man'],'head',['PERSON']).
person(['woman'],'head',['PERSON']).
person(['child'],'head',['PERSON']).

%*********************************
%description de la classe organisation
%*********************************

organisation(['search','and','rescue'],'PROPER-NOUN',['ORGANISATION','SAR']).
organisation(['coast','guard'],'PROPER-NOUN',['ORGANISATION','SAR']).
organisation(['coast','surveillance'],'PROPER-NOUN',['ORGANISATION','SAR']).
organisation(['coast','guard','auxiliary'],'PROPER-NOUN',['ORGANISATION','SAR']).
organisation(['rcc'],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['mrcc'],'PROPER-NOUN',['ORGANISATION','MRCC']).
organisation(['mrc'],'PROPER-NOUN',['ORGANISATION','MRCC']).
```

```
organisation(['moc'],'PROPER-NOUN',['ORGANISATION','maritime operation centre']).
organisation(['rescue','coordination','centre'|_],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['rescue','coordination'|_],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['rescue','centre'|_],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['rescue','centre'],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['rescue','center'],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['subcentre'],'PROPER-NOUN',['ORGANISATION']).
organisation(['coast','guard'],'PROPER-NOUN',['ORGANISATION','CG']).
organisation(['maritime','operation','centre'],'PROPER-NOUN',['ORGANISATION','MOC']).
organisation(['rescue'],'PROPER-NOUN',['ORGANISATION','SAR']).
organisation(['rescue','centre'],'PROPER-NOUN',['ORGANISATION','RCC']).
organisation(['maritime','rescue','centre'],'PROPER-NOUN',['ORGANISATION','MRC']).
organisation(['maritime','rescue','sub-centre'],'PROPER-NOUN',['ORGANISATION','MRSC']).
organisation(['operation','centre'],'PROPER-NOUN',['ORGANISATION','OP']).
organisation(['rcmp'],'PROPER-NOUN',['ORGANISATION','RCMP']).
organisation(['occ'],'PROPER-NOUN',['ORGANISATION','OCC']).
organisation(['canadian','mission','control','centre'],'PROPER-NOUN',['ORGANISATION','CMCC'])
organisation(['dfo'],'PROPER-NOUN',['ORGANISATION']).
organisation(['cmre'],'PROPER-NOUN',['ORGANISATION','CMRE']).
organisation(['cbc'],'PROPER-NOUN',['ORGANISATION','CBC']).
organisation(['mso'],'PROPER-NOUN',['ORGANISATION','MSO']).
organisation(['fundy','coast','guard','radio'],'PROPER-NOUN',['ORGANISATION','SAR organisatio
organisation(['fundy','radio'],'PROPER-NOUN',['SAR organistaion']).
organisation(['newfoundland','marine','survey'],'PROPER-NOUN',['ORGANISATION']).
organisation(['newfoundland','marine','surveys'],'PROPER-NOUN',['ORGANISATION']).
organisation(['fishery','patrol'],'PROPER-NOUN',['ORGANISATION']).
organisation(['cor'],'PROPER-NOUN',['ORGANISATION','COR']).
organisation(['dngcc'],'PROPER-NOUN',['ORGANISATION','DNGCC']).
organisation(['marine','communication'|L],'PROPER-NOUN',['ORGANISATION',['MCTS'|L]]).
organisation(['mcts'],'PROPER-NOUN',['ORGANISATION','MCTS']).
```

```
organisation(['owl'],'PROPER-NOUN',['ORGANISATION','OWL']).
organisation(['airport'],'head',['ORGANISATION']).
organisation(['base'],'head',['ORGANISATION']).

%*********************************
%description de la classe Means_of_detection
%*********************************

detection(['surface','search'],'head',['DETECTION-MEANS']).
detection(['radar','search'],'head',['DETECTION-MEANS']).
detection(['visual','search'],'head',['DETECTION-MEANS']).
detection(['diver'],'head',['DETECTION-MEANS']).
detection(['sar','sat'],'PROPER-NOUN',['DETECTION-MEANS']).
detection(['satellite'],'head',['DETECTION-MEANS']).
detection(['radar'],'head',['DETECTION-MEANS']).
detection(['dive','team'],'head',['DETECTION-MEANS']).
detection(['dive'],'head',['DETECTION-MEANS']).
detection(['dive'],'verb',['DETECTION-MEANS']).
detection(['search','light'],'head',['DETECTION-MEANS']).
detection(['goggle'],'head',['DETECTION-MEANS']).
detection(['slbmb'],'PROPER-NOUN',['DETECTION-MEANS']). %qui est en réalité SLDMB
detection(['buoy'],'head',['DETECTION-MEANS']).

%*********************************
%description de la classe Event et des
%ses sous-classes Incident et Initial-alert
%*********************************

incident(['drift'],'verb',['INCIDENT-TYPE']).
incident(['broken'],'modif',['INCIDENT-TYPE']).
```

```
incident(['breakdown'],'verb',['INCIDENT-TYPE']).
incident(['disable'],'verb',['INCIDENT-TYPE']).
incident(['disorient'],'verb',['INCIDENT-TYPE']).
incident(['fire'],'head',['INCIDENT-TYPE']).
incident(['smoke'],'head',['INCIDENT-TYPE']).
incident(['dead','battery'],'PROPER-NOUN',['INCIDENT-TYPE']).
incident(['overdue'],'modif',['INCIDENT-TYPE']).
incident(['overdue'],'head',['INCIDENT-TYPE']).
incident(['miss'],'verb',['INCIDENT-TYPE']).
incident(['crash'],'verb',['INCIDENT-TYPE']).
incident(['lost'],'verb',['INCIDENT-TYPE']).
incident(['lost'],'modif',['INCIDENT-TYPE']).
incident(['out','of','gas'],'PROPER-NOUN',['INCIDENT-TYPE']).
incident(['swamp'],'verb',['INCIDENT-TYPE']).
incident(['incident'],'head',['INCIDENT']).

initial_alert(['epirb'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['e-perp'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['e-perp','beacon'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['plb'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['ple'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['sldmp'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['emergency'],'head',['ALERT']).
initial_alert(['alert'],'verb',['ALERT']).
initial_alert(['alert'],'head',['ALERT']).
initial_alert(['elt'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['beacon'],'head',['ALERT-TYPE']).
initial_alert(['flare'],'head',['ALERT-TYPE']).
initial_alert(['witness','report'],'head',['ALERT-TYPE']).
initial_alert(['mayday','relay'],'head',['ALERT-TYPE']).
```

```
initial_alert(['slbmb'],'PROPER-NOUN',['ALERT-TYPE']).
initial_alert(['report'],'head',['ALERT-TYPE']).
initial_alert(['report'],'verb',['ALERT-TYPE']).
initial_alert(['problem'],'head',['ALERT']).
initial_alert(['relay'],'head',['ALERT']).

%*********************************
%description de la classe Weather-conditions et de
%ses sous-classes fog, rain, snow, wind et sea
%*********************************

weather(['fog'],'head',['WEATHER-TYPE']).
weather(['haze'],'head',['WEATHER-TYPE']).
weather(['rain'],'head',['WEATHER-TYPE']).
weather(['wind'],'head',['WEATHER-TYPE']).
weather(['visibility'],'head',['WEATHER-TYPE']).
weather(['sea'],'head',['WEATHER-TYPE']).
weather(['swirl'],'head',['WEATHER-TYPE']).
weather(['weather'],'head',['WEATHER']).
weather(['forecast'],'head',['WEATHER']).

%*********************************
%description de la classe Date
%*********************************

time(['secure','time'],'PROPER-NOUN',['TIMEX']).
time(['local','time'],'PROPER-NOUN',['TIMEX']).
time(['scene','time'],'PROPER-NOUN',['TIMEX']).
time(['bingo','time'],'PROPER-NOUN',['TIMEX']).
time(['task','time'],'PROPER-NOUN',['TIMEX']).
```

```
time(['depart','time'],'PROPER-NOUN',['TIMEX']).
time(['ground','time'],'PROPER-NOUN',['TIMEX']).
time(['transit','time'],'PROPER-NOUN',['TIMEX']).
time(['time','on','scene'],'PROPER-NOUN',['TIMEX']).
time(['zulu'],'PROPER-NOUN',['TIMEX']).
time(['hour'],'PROPER-NOUN',['TIMEX']).
time(['minute'],'PROPER-NOUN',['TIMEX']).
time(['second'],'PROPER-NOUN',['TIMEX']).
time(['midnight'],'head',['TIMEX']).
time(['afternoon'],'head',['TIMEX']).
time(['evening'],'head',['TIMEX']).
time(['overnight'],'head',['TIMEX']).
time(['utc'],'PROPER-NOUN',['TIMEX']).
time(['eta'],'PROPER-NOUN',['TIMEX']). %estimated time of arrival
time(['oclock'],'head',['TIMEX']).
time(['pm'],'PROPER-NOUN',['TIMEX']).
time(['am'],'PROPER-NOUN',['TIMEX']).
time(['time'],'head',['TIMEX']).
time(['tomorrow',Period],'head',['TIMEX']):-period(Period).
time(['yesterday',Period],'head',['TIMEX']):-period(Period).
time(['today',Period],'head',['TIMEX']):-period(Period).
time(['day'],'head',['TIMEX']).
time(['monday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
time(['tuesday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
time(['wednsday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
time(['thursday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
time(['friday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
time(['sunday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
time(['saturday',Period],'PROPER-NOUN',['TIMEX']):-period(Period).
```

```
time(['tomorrow'],'head',['TIMEX']).
time(['morning'],'head',['TIMEX']).
time(['yesterday'],'head',['TIMEX']).
time(['today'],'head',['TIMEX']).
time(['last','day'],'PROPER-NOUN',['TIMEX']).
time(['monday'],'PROPER-NOUN',['TIMEX']).
time(['tuseday'],'PROPER-NOUN',['TIMEX']).
time(['wednsday'],'PROPER-NOUN',['TIMEX']).
time(['thursday'],'PROPER-NOUN',['TIMEX']).
time(['friday'],'PROPER-NOUN',['TIMEX']).
time(['sunday'],'PROPER-NOUN',['TIMEX']).
time(['saturday'],'PROPER-NOUN',['TIMEX']).
time(['january'],'PROPER-NOUN',['TIMEX']).
time(['february'],'PROPER-NOUN',['TIMEX']).
time(['march'],'PROPER-NOUN',['TIMEX']).
time(['april'],'PROPER-NOUN',['TIMEX']).
time(['may'],'PROPER-NOUN',['TIMEX']).
time(['june'],'PROPER-NOUN',['TIMEX']).
time(['july'],'PROPER-NOUN',['TIMEX']).
time(['august'],'PROPER-NOUN',['TIMEX']).
time(['september'],'PROPER-NOUN',['TIMEX']).
time(['october'],'PROPER-NOUN',['TIMEX']).
time(['november'],'PROPER-NOUN',['TIMEX']).
time(['december'],'PROPER-NOUN',['TIMEX']).
time(['couple','of','days'],'PROPER-NOUN',['TIMEX']).

period('morning').
period('evening').
period('afternoon').
period('night').
```

```
%************************************
%description de la classe code de la classe langage
%************************************

code(['alpha'],'PROPER-NOUN',['CODE-TYPE']).
code(['bravo'],'PROPER-NOUN',['CODE-TYPE']).
code(['charlie'],'PROPER-NOUN',['CODE-TYPE']).
code(['carl'],'PROPER-NOUN',['CODE-TYPE']).
code(['delta'],'PROPER-NOUN',['CODE-TYPE']).
code(['echo'],'PROPER-NOUN',['CODE-TYPE']).
code(['endurance'],'PROPER-NOUN',['CODE-TYPE']).
code(['foxtrot'],'PROPER-NOUN',['CODE-TYPE']).
code(['fox'],'PROPER-NOUN',['CODE-TYPE']).
code(['golf'],'PROPER-NOUN',['CODE-TYPE']).
code(['hotel'],'PROPER-NOUN',['CODE-TYPE']).
code(['india'],'PROPER-NOUN',['CODE-TYPE']).
code(['juliet'],'PROPER-NOUN',['CODE-TYPE']).
code(['kilo'],'PROPER-NOUN',['CODE-TYPE']).
code(['lima'],'PROPER-NOUN',['CODE-TYPE']).
code(['lemon'],'PROPER-NOUN',['CODE-TYPE']).
code(['mike'],'PROPER-NOUN',['CODE-TYPE']).
code(['niner'],'PROPER-NOUN',['CODE-TYPE']).
code(['oscar'],'PROPER-NOUN',['CODE-TYPE']).
code(['papa'],'PROPER-NOUN',['CODE-TYPE']).
code(['quebec'],'PROPER-NOUN',['CODE-TYPE']).
code(['romeo'],'PROPER-NOUN',['CODE-TYPE']).
code(['sierra'],'PROPER-NOUN',['CODE-TYPE']).
code(['tango'],'PROPER-NOUN',['CODE-TYPE']).
code(['uniform'],'PROPER-NOUN',['CODE-TYPE']).
```

```
code(['victor'],'PROPER-NOUN',['CODE-TYPE']).
code(['whiskey'],'PROPER-NOUN',['CODE-TYPE']).
code(['x-ray'],'PROPER-NOUN',['CODE-TYPE']).
code(['yankee'],'PROPER-NOUN',['CODE-TYPE']).
code(['zulu'],'PROPER-NOUN',['CODE-TYPE']).

code(['code'],'head',['CODE']).
code(['number'],'head',['CODE']).
%************************************
%description de la classe acknowledgment de la classe langage
%************************************

ackword(['ok'],'modif',['ACK']).
ackword(['okay'],'modif',['ACK']).
ackword(['all-right'],'PROPER-NOUN',['ACK']).
ackword(['alright'],'modif',['ACK']).
ackword(['hum'],'modif',['ACK']).
ackword(['right'],'modif',['ACK']).
ackword(['ronald'],'PROPER-NOUN',['ACK']).
ackword(['ronald'],'PROPER-NOUN',['ACK']).
ackword(['yes'],'modif',['ACK']).
ackword(['good-morning'],'PROPER-NOUN',['ACK']).
ackword(['excuse-me'],'PROPER-NOUN',['ACK']).
ackword(['good-afetrnoon'],'PROPER-NOUN',['ACK']).
ackword(['good-evening'],'PROPER-NOUN',['ACK']).
ackword(['just-a-moment'],'PROPER-NOUN',['ACK']).
ackword(['just-a-second'],'PROPER-NOUN',['ACK']).
ackword(['a-second'],'PROPER-NOUN',['ACK']).
ackword(['good-night'],'PROPER-NOUN',['ACK']).
ackword(['good-luck'],'PROPER-NOUN',['ACK']).
```

```
ackword(['no','problem'],'PROPER-NOUN',['ACK']).
ackword(_,_,_):- fail,!.

%************************************
%description de la sous-classe
%Color
%************************************

color(['red'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['white'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['blue'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['blank'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['magenta'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['yellow'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['green'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['orange'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['black'],'modif',['PROPERTIES-COLOR-TYPE']).
color(['color'],'head',['PROPERTIES-COLOR']).
color(['color'],'modif',['PROPERTIES-COLOR']).

%************************************
%description de la sous-classe
%length
%************************************

mylength(['foot'],'PROPER-NOUN',['PROPERTIES-DISTANCE-TYPE']).
mylength(['meter'],'PROPER-NOUN',['PROPERTIES-DISTANCE-TYPE']).
mylength(['mile'],'PROPER-NOUN',['PROPERTIES-DISTANCE-TYPE']).
mylength(['length'],'head',['PROPERTIES-DISTANCE']).
mylength(['distance'],'head',['PROPERTIES-DISTANCE']).
```

```
%************************************
%description de la sous-classe
%speed
%************************************

speed(['knot'],'PROPER-NOUN',['PROPERTIES-SPEED-TYPE']).
speed(['speed'],'head',['PROPERTIES-SPEED']).
speed(['horse','power'],'PROPER-NOUN',['PROPERTIES-ENGINE-TYPE']).
speed(['power'],'head',['PROPERTIES-ENGINE-TYPE']).

%************************************
%description de la sous-classe
%status de la classe propriétés
%************************************

status(['status'],'head',['STATUS']).
status(['bad'],'modif',['PROPERTIES-TYPE']).
status(['good'],'modif',['PROPERTIES-TYPE']).
status(['dead'],'modif',['PROPERTIES-TYPE']).
status(['available'],'modif',['PROPERTIES-TYPE']).

%************************************
%description de la sous-classe
%status object de la classe status
%************************************

status(['anchor'],'verb',['STATUS-OBJECT']).
status(['secure'],'verb',['STATUS-OBJECT']).
status(['tow'],'verb',['STATUS-OBJECT']).
```

```
status(['escort'],'verb',['STATUS-OBJECT']).
status(['fly'],'verb',['STATUS-OBJECT']).
status(['takeoff'],'verb',['STATUS-OBJECT']).
status(['airborne'],'modif',['STATUS-OBJECT']).
status(['airborne'],'verb',['STATUS-OBJECT']).
status(['airborne'],'head',['STATUS-OBJECT']).
status(['land'],'verb',['STATUS-OBJECT']).
status(['arrive'],'verb',['STATUS-OBJECT']).
status(['go'],'verb',['STATUS-OBJECT']).
status(['leave'],'verb',['STATUS-OBJECT']).

%************************************
%description de la sous-classe
%status person de la classe status
%************************************

status(['alive'],'verb',['STATUS-PERSON']).

%************************************
%description de la sous-classe
%status task de la classe status
%************************************

status(['complete'],'verb',['STATUS-TASK-RESULT']).
status(['nothing'],'head',['STATUS-TASK-RESULT']).
status(['continue'],'verb',['STATUS-TASK']).
status(['cancel'],'verb',['STATUS-TASK']).
status(['start'],'verb',['STATUS-TASK']).
status(['plan'],'verb',['STATUS-TASK-PLANNED']).
status(['alternative'],'head',['STATUS-TASK-PLANNED']).
```

```
status(['request'],'verb',['STATUS-TASK-REQUEST']).
status(['request'],'head',['STATUS-TASK-REQUEST']).
status(['wonder'],'verb',['STATUS-TASK-REQUEST']).
status(['need'],'verb',['STATUS-TASK-REQUEST']).

%************************************
%description de la classe materiel qui
%forme un bateau ou un avion. C'est utilisê
%pour la description d'un objet perdu
%************************************

materiel(['wood'],'head',['PROPERTIES-MATERIAL-TYPE']).
materiel(['aluminum'],'head',['PROPERTIES-MATERIAL-TYPE']).
materiel(['fiberglass'],'head',['PROPERTIES-MATERIAL-TYPE']).

%************************************
%description de la sous-classe
%weight
%************************************

weight(['pound'],'PROPER-NOUN',['PROPERTIES-CAPACITY-TYPE']).
weight(['kilo'],'PROPER-NOUN',['PROPERTIES-CAPACITY-TYPE']).
weight(['seat'],'head',['PROPERTIES-CAPACITY-TYPE']).

%**********************************88
%DESCRIPTION DE LA HIERARCHIE PART-OF
%
%**********************************8

search_mission(['case'],'PROPER-NOUN',['MISSION']).
```

```
search_mission(['mission'],'PROPER-NOUN',['MISSION']).

%*********************************************
%Description de la sous-classe search-unit
%
%*********************************************
search_unit(['crew'],'head',['SEARCH-UNIT-TEAM']).
search_unit(['squadron'],'head',['SEARCH-UNIT-TEAM']).

status(['search'],'head',['SEARCH-UNIT-TASK']).
status(['rescue'],'head',['SEARCH-UNIT-TASK']).
status(['search'],'verb',['SEARCH-UNIT-TASK']).
status(['rescue'],'verb',['SEARCH-UNIT-TASK']).
status(['anchor'],'verb',['SEARCH-UNIT-TASK']).
status(['secure'],'verb',['SEARCH-UNIT-TASK']).
status(['tow'],'verb',['SEARCH-UNIT-TASK']).
status(['escort'],'verb',['SEARCH-UNIT-TASK']).

task(['search'],'head',['SEARCH-UNIT-TASK']).
task(['rescue'],'head',['SEARCH-UNIT-TASK']).
task(['search'],'verb',['SEARCH-UNIT-TASK']).
task(['rescue'],'verb',['SEARCH-UNIT-TASK']).
task(['anchor'],'verb',['SEARCH-UNIT-TASK']).
task(['secure'],'verb',['SEARCH-UNIT-TASK']).
task(['tow'],'verb',['SEARCH-UNIT-TASK']).
task(['escort'],'verb',['SEARCH-UNIT-TASK']).

%*********************************************
%Description de la sous-classe component
%
```

```
%*********************************************
component(['engine'],'head',['PARTOF-PHYSICAL-OBJECT']).
component(['stern'],'head',['PARTOF-VESSEL']).
component(['battery'],'head',['PARTOF-VESSEL']).
component(['anchor'],'head',['PARTOF-VESSEL']).
component(['bow'],'head',['PARTOF-VESSEL']).
component(['compass'],'head',['PARTOF-VESSEL']).
component(['life','jacket'],'head',['PARTOF-VESSEL']).

engine(['pump'],'head',['PARTOF-ENGINE']).
fuel(['fuel'],'head',['FUEL']).
fuel(['gas'],'head',['FUEL-TYPE']).
fuel(['gaz'],'head',['FUEL-TYPE']).

%*********************************************
%Description de la sous-classe inside
%
%*********************************************

inside(['on-board'],'modif',['PARTOF-PHYSICAL-OBJECT']).
inside(['cabine'],'head',['PARTOF-PHYSICAL-OBJECT']).
inside(['hall'],'head',['PARTOF-PHYSICAL-OBJECT']).
```

# Annexe B

# Liste des marques utilisées pour les segmentations linguistique et thématique

## B.1 Liste des marques lexicales utilisées pour la segmentation linguistique

Cette liste présente l'ensemble des marques lexicales et syntaxiques utilisées pour l'étape de segmentation linguistique.

```
%***********************************
% Liste des marques lexicales
%***********************************

list_of_lex_cue('...','I', L, L ,_).
list_of_lex_cue('absolutely','Ack', L, L ,_).
list_of_lex_cue('alright','Ack', L, L ,_).
```

```
list_of_lex_cue('anyway','anyway', L, L ,'RL').
list_of_lex_cue('bonjour','BC', L, L ,_).
list_of_lex_cue('bye','EC', L, L ,_).
list_of_lex_cue('cheers','EC', L, L ,_).
list_of_lex_cue('cool','Ack', L, L ,_).
list_of_lex_cue('exactly','Ack', L, L ,_).
list_of_lex_cue('excellent','Ack', L, L ,'LR').
list_of_lex_cue('fine','Ack', L, L ,_).
list_of_lex_cue('gees','FB', L, L ,_).
list_of_lex_cue('good',Cue, [Word|L], R,'LR'):-
                                completion([Word|L],Cue,R).
list_of_lex_cue('good','Ack', [],[],_).
list_of_lex_cue('good',Cue, [Word|L], R,'RL'):-
                                completion([Word|L],Cue,R).
list_of_lex_cue('great','Ack', L, L ,_).


list_of_lex_cue('ha', 'ha', L, L ,_).

list_of_lex_cue('hung','EC', [Word|L], L ,_) :- val(Word,lexMin,'up').
list_of_lex_cue('hello','BC', L, L ,_).
list_of_lex_cue('hum','FB', L, L ,_).
list_of_lex_cue('i','CT',  [Word|L], L ,_):- val(Word,lexMin,'mean').%
list_of_lex_cue('mean','CT',  [Word|L], L ,'RL'):- val(Word,lexMin,'I').%
list_of_lex_cue('i','Ack', [Word|L], L ,_):- val(Word,lexMin,'see').%
list_of_lex_cue('i','Ack', [Word|L], L ,_):- val(Word,lexMin,'know').%
list_of_lex_cue('laughing','FB', L, L ,_).%
list_of_lex_cue('no','Ack', [Word|L], L ,_):-val(Word,lexMin,'problem').
list_of_lex_cue('no','Ack', L, L ,_).
list_of_lex_cue('oh','FB', L, L ,'LR').
```

```
list_of_lex_cue('ok','Ack', L, L ,'LR').
list_of_lex_cue('okay','Ack', L, L ,'LR').
list_of_lex_cue('oui','Ack', L, L ,_).
list_of_lex_cue('pardon','CT', L, L ,'LR').%
list_of_lex_cue('perfect','Ack', L, L ,_).
list_of_lex_cue('right','Ack', L, L ,'LR').
list_of_lex_cue('ronald','Ack', L, L ,_).
list_of_lex_cue('sure','Ack', L, L ,'LR').
list_of_lex_cue('thanks','thank', L, L ,_).
list_of_lex_cue('thank','thank', [Word|L], L ,'LR'):- val(Word,lexMin,'you').
list_of_lex_cue('wait','wait', L, L ,_).
list_of_lex_cue('your','Ack', [Word|L], L ,'LR'):- val(Word,lexMin,'welcome').
list_of_lex_cue('yes','Ack', L, L ,'LR').
list_of_lex_cue('yah','Ack', L, L ,'LR').
list_of_lex_cue('you','Emp', [Word|L], L ,'LR'):-val(Word,lexMin,'see').
list_of_lex_cue('how','BC', [W1,W2|L], L ,'LR'):-val(W1,lexMin,'are'),
                                  val(W2,lexMin,'you').
list_of_lex_cue('you','BC', [W1,W2|L], L ,'RL'):-val(W1,lexMin,'are'),
                                  val(W2,lexMin,'how').
list_of_lex_cue('yourself','BC', L, L ,'LR').
list_of_lex_cue('talk','EC', [W1,W2,W3|L], L ,_):-
                                          val(W1,lexMin,'to'),
                                    val(W2,lexMin,'you'),
                                    val(W3,lexMin,'later').
list_of_lex_cue('later','EC', [W1,W2,W3|L], L ,'RL'):-
                                          val(W1,lexMin,'you'),
                                    val(W2,lexMin,'to'),
                                    val(W3,lexMin,'talk').

list_of_lex_cue('you','Emp', [Word|L], L ,_):- val(Word,lexMin,'know').
```

```
list_of_lex_cue('go','CT', [Word|L], L ,_):- val(Word,lexMin,'ahead').%
list_of_lex_cue('goodbye','EC', L, L ,_).
list_of_lex_cue('excuse','CT', L, L ,_).%
list_of_lex_cue('allo','BC', L, L ,_).
list_of_lex_cue('hi','hi', L, L ,_).
list_of_lex_cue('hey','Emp', L, L ,_).%
list_of_lex_cue('of','Ack', [Word|L], L ,_):- val(Word,lexMin,'course').%
list_of_lex_cue('yeah','Ack', L, L ,'LR').
list_of_lex_cue('not','Ack', [Word|L], L ,_):- val(Word,lexMin,'bad') .
list_of_lex_cue('not','Ack', [_,W2|L], L ,_):- val(W2,lexMin,'bad'). %too bad, so bad, that ba
list_of_lex_cue('all','Ack', [Word|L], L ,'LR'):- val(Word,lexMin,'right').

%***********************************
% List des marques syntaxiques
%***********************************

list_of_synt_cue('?','?', L, L ,_).

%***********************************
% ajout de la composante longueur de la phrase
%***********************************

list_of_synt_cue('but','but', L, L ,'LR').%
list_of_synt_cue('so','so', L, L ,'LR').
list_of_synt_cue('well','well', L, L ,_).%
list_of_synt_cue('now','now', L, L ,_).
list_of_synt_cue('cause','CT', L, L ,'LR').%
list_of_synt_cue('because','CT', L, L ,'LR').%
list_of_synt_cue('then','then', L, L ,'LR').%
list_of_synt_cue('or','CT', L, L ,'LR').%
```

```
completion([Word|L],Cue,L) :-
 val(Word, lexMin,Lex),!,
 expression(Lex,Cue).

completion(L,'Ack',L).

expression('morning','BC').
expression('evening','BC').
expression('afternoon','BC').
expression('night','EC').
expression('very','Ack').
expression('pretty','Ack').
expression(',','Ack').


%****************************************
% expressions typiques
%****************************************

formulaic_speech([],[]).

formulaic_speech([Token|List],['wait']):-
val(Token, lexMin, Lex),
wait_expression(Lex,List).

formulaic_speech([_|List],R):-
            formulaic_speech(List,R).
```

```
wait_expression('hang',[Token|_]):-
val(Token, lexMin,'on').
wait_expression('hold',[Token|_]):-
val(Token, lexMin,'on').
wait_expression('just',[Token1,Token2|_]):-
val(Token1, lexMin,'a'),
        val(Token2, lexMin,'second').
wait_expression('just',[Token1,Token2|_]):-
val(Token1, lexMin,'a'),
        val(Token2, lexMin,'minute').
wait_expression('just',[Token1,Token2|_]):-
val(Token1, lexMin,'a'),
        val(Token2, lexMin,'moment').

wait_expression('wait',L):-
wait_expression('just',L).

wait_expression('standby',_).
```

## B.2 Liste des marques lexicales utilisées pour la segmentation thématique

Cette liste présente l'ensemble des marques lexicales et syntaxiques utilisées pour l'étape de segmentation thématique. Cette liste est une version augmentée de la liste précédente. En particulier, certaines marques lexicales présentes en fin d'un énoncé sont prises en compte pour la segmentation thématique et non pour la segmentation linguistique. des exemple de ces marques sont celles se terminant par une '*'.

```
%**********************************
```

```
% Liste des marques lexicales
%**********************************

list_of_lex_cue('...','I', L, L ,_).
list_of_lex_cue('absolutely','Ack', L, L ,_).
list_of_lex_cue('alright','Ack', L, L ,_).
list_of_lex_cue('anyway','anyway', L, L ,'RL').
list_of_lex_cue('bonjour','BC', L, L ,_).
list_of_lex_cue('bye','EC', L, L ,_).
list_of_lex_cue('cheers','EC', L, L ,_).
list_of_lex_cue('cool','Ack', L, L ,_):- length(L, Long), Long < 3.
list_of_lex_cue('exactly','Ack', L, L ,_).
list_of_lex_cue('excellent','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('fine','Ack', L, L ,_):- length(L, Long), Long < 3.
list_of_lex_cue('gees','FB', L, L ,_).
list_of_lex_cue('good',Cue, [Word|L], R,'LR'):-
                                  completion([Word|L],Cue,R).
list_of_lex_cue('good','Ack', [],[],_).
list_of_lex_cue('good',Cue, [Word|L], R,'RL'):-
                                  completion([Word|L],Cue,R).
list_of_lex_cue('great','Ack', L, L ,_):- length(L, Long), Long < 3.
list_of_lex_cue('hung','EC', [Word|L], L ,_) :- val(Word,lexMin,'up').
list_of_lex_cue('hello','BC', L, L ,_).
list_of_lex_cue('hum','FB', L, L ,_).
list_of_lex_cue('i','CT',  [Word|L], L ,_):- val(Word,lexMin,'mean').
list_of_lex_cue('mean','CT',  [Word|L], L ,'RL'):- val(Word,lexMin,'I').
list_of_lex_cue('i','Ack', [Word|L], L ,_):- val(Word,lexMin,'see').
list_of_lex_cue('i','Ack', [Word|L], L ,_):- val(Word,lexMin,'know').
list_of_lex_cue('laughing','FB', L, L ,_).
list_of_lex_cue('no','Ack', [Word|L], L ,_):-val(Word,lexMin,'problem').
```

```
list_of_lex_cue('no','Ack', L, L ,_).
list_of_lex_cue('oh','FB', L, L ,'LR').
list_of_lex_cue('ok','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('okay','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('ok','Ack*', L, L ,'RL').
list_of_lex_cue('yah','Ack*', L, L ,'RL').
list_of_lex_cue('oui','Ack', L, L ,_):- length(L, Long), Long < 3.
list_of_lex_cue('pardon','CT', L, L ,'LR').
list_of_lex_cue('perfect','Ack', L, L ,_):- length(L, Long), Long < 3.
list_of_lex_cue('right','Ack*', L, L ,'RL').
list_of_lex_cue('right','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('ronald','Ack', L, L ,_).
list_of_lex_cue('sure','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('thanks','thank', L, L ,_).
list_of_lex_cue('thank','thank', [Word|L], L ,'LR'):- val(Word,lexMin,'you').
list_of_lex_cue('your','Ack', [Word|L], L ,'LR'):- val(Word,lexMin,'welcome').
list_of_lex_cue('yes','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('yah','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('you','Emp', [Word|L], L ,'LR'):-val(Word,lexMin,'see').
list_of_lex_cue('how','BC', [W1,W2|L], L ,'LR'):-val(W1,lexMin,'are'),
                                    val(W2,lexMin,'you').
list_of_lex_cue('you','BC', [W1,W2|L], L ,'RL'):-val(W1,lexMin,'are'),
                                    val(W2,lexMin,'how').
list_of_lex_cue('yourself','BC', L, L ,'LR').
list_of_lex_cue('talk','EC', [W1,W2,W3|L], L ,_):-
                                                val(W1,lexMin,'to'),
                                          val(W2,lexMin,'you'),
                                          val(W3,lexMin,'later').
list_of_lex_cue('later','EC', [W1,W2,W3|L], L ,'RL'):-
                                                val(W1,lexMin,'you'),
```

```
                                        val(W2,lexMin,'to'),
                                        val(W3,lexMin,'talk').

list_of_lex_cue('you','Emp', [Word|L], L ,_):- val(Word,lexMin,'know').
list_of_lex_cue('go','CT', [Word|L], L ,_):- val(Word,lexMin,'ahead').
list_of_lex_cue('goodbye','EC', L, L ,_).
list_of_lex_cue('excuse','CT', L, L ,_).
list_of_lex_cue('allo','BC', L, L ,_).
list_of_lex_cue('hi','BC', L, L ,_).
list_of_lex_cue('of','Ack', [Word|L], L ,_):- val(Word,lexMin,'course').
list_of_lex_cue('yeah','Ack', L, L ,'LR'):- length(L, Long), Long < 3.
list_of_lex_cue('yeah','Ack*', L, L ,'RL').
list_of_lex_cue('not','Ack', [Word|L], L ,_):- val(Word,lexMin,'bad') .
list_of_lex_cue('not','Ack', [_,W2|L], L ,_):- val(W2,lexMin,'bad'). %too bad, so bad, that ba
list_of_lex_cue('all','Ack', [Word|L], L ,'LR'):- val(Word,lexMin,'right').
list_of_lex_cue('right','Ack*', [Word|L], L ,'RL'):- val(Word,lexMin,'all').

%***********************************
% Liste des marques syntaxiques
%***********************************

list_of_synt_cue('?','?', L, L ,_).
list_of_synt_cue('actually','actually', L, L ,'LR').
list_of_synt_cue('but','but', L, L ,'LR'):- length(L, Long), Long > 2.
list_of_synt_cue('so','so', L, L ,'LR'):- length(L, Long), Long > 2.
list_of_synt_cue('well','well', L, L ,_):- length(L, Long), Long > 2.
list_of_synt_cue('well','well*', L, L ,'RL').
list_of_synt_cue('well','well*', ['.'], ['.'] ,'RL').
list_of_synt_cue('cause','CT', L, L ,'LR').
list_of_synt_cue('because','CT', L, L ,'LR').
```

```
list_of_synt_cue('so','so*', L, L ,'RL').
list_of_synt_cue('then','then*', L, L ,'RL').
list_of_synt_cue('then','then', L, L ,'LR').
list_of_synt_cue('or','CT', L, L ,'LR').

completion([Word|L],Cue,L) :-
 val(Word, lexMin,Lex),!,
 expression(Lex,Cue).

completion(L,'Ack',L).

expression('morning','BC').
expression('evening','BC').
expression('afternoon','BC').
expression('night','EC').
expression('very','Ack').
expression('pretty','Ack').
expression(',','Ack').


%***************************************
% expressions typiques
%***************************************

formulaic_speech([],[]).

formulaic_speech([Token|List],['wait']):-
val(Token, lexMin, Lex),
wait_expression(Lex,List).
```

```
formulaic_speech([_|List],R):-
            formulaic_speech(List,R).

wait_expression('hang',[Token|_]):-
val(Token, lexMin,'on').
wait_expression('hold',[Token|_]):-
val(Token, lexMin,'on').
wait_expression('just',[Token1,Token2|_]):-
val(Token1, lexMin,'a'),
        val(Token2, lexMin,'second').
wait_expression('just',[Token1,Token2|_]):-
val(Token1, lexMin,'a'),
        val(Token2, lexMin,'minute').
wait_expression('just',[Token1,Token2|_]):-
val(Token1, lexMin,'a'),
        val(Token2, lexMin,'moment').

wait_expression('wait',L):-
wait_expression('just',L).

wait_expression('standby',_).
```

# Annexe C

# Estimation des paramètres du modèle d'étiquetage sémantique robuste

## C.1 Estimation de $\beta$

La formulation de la problématique de l'étiquetage des expressions pertinentes sémantiquement similaires aux termes de l'ontologie est exprimée par le produit d'experts suivant :

$$\begin{aligned} P(C^t = k|w^t, T^t) &= \frac{P(C^t = k|w^t)^{\beta_1} P(C^t = k|T^t)^{\beta_2}}{\Sigma_{l=1}^{K} P(C^t = l|w^t)^{\beta_1} P(C^t = l|T^t)^{\beta_2}} \\ &= \frac{\exp^{\beta_1 \log P(C^t=k|w^t) + \beta_2 \log P(C^t=k|T^t)}}{\Sigma_{l=1}^{K} \exp^{\beta_1 \log P(C^t=l|w^t) + \beta_2 \log P(C^t=l|T^t)}} \\ &= \frac{\exp^{\beta^T X_k^t}}{\Sigma_{l=1}^{K} \exp^{\beta^T X_l^t}} \end{aligned} \tag{C.1}$$

$k$ est un des concepts du niveau supérieur de l'ontologie (tableau 5.1), $K$ est le nombre de ces concepts, $X_k^t = (\log P(C^t = k|w^t), \log P(C^t = k|T^t))$ est une observation, $P(C^t = k|w^t)$ représente la probabilité de similarité entre le concept $k$ et le mot $w^t$, $P(C^t = k|T^t)$ est la

probabilité d'observer le concept $k$ étant donné le thème $T^t$, et $\beta = (\beta_1, \beta_2)^T$.

La valeur de $\beta$ est obtenue en minimisant la -log vraisemblance ($-l(\beta)$) définie par :

$$l(\beta) = \Sigma_{t=1}^{n} \log P(C^t|w^t, T^t)$$

$n$ est la longueur de la séquence des observations. Pour minimiser $-l(\beta)$ nous estimons les $\beta_i$, $i \in \{1, 2\}$ qui satisfont l'équation suivante :

$$\frac{\partial l(\beta)}{\partial \beta_i} = 0$$

Soit $p_k^t$ la densité de probabilité associée à distribution $P(C|w, T)$ définie par l'équation suivante :

$$p_k^t \stackrel{\text{def}}{=} \frac{\exp^{\beta^T X_k^t}}{\Sigma_{l=1}^{K} \exp^{\beta^T X_l^t}}$$

La dérivée partielle de $l(\beta)$ s'exprime par l'équation suivante :

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta_i} &= \frac{\partial \log p_k^t}{\partial \beta_i} \\ &= \frac{X_{ki}^t \Sigma_{k'=1}^{K} p_{k'}^t X_{k'i}^t - \exp^{\beta^T X_k^t} (\Sigma_{k'=1}^{K} p_{k'}^t X_{k'i}^t)^2}{(\Sigma_{k'=1}^{K} p_{k'}^t X_{k'i}^t)^2} \\ &= X_{ki}^t - \Sigma_{k'=1}^{K} p_{k'}^t X_{k'i}^t \end{aligned} \tag{C.2}$$

Nous remarquons que l'équation C.2 ne peut être calculée de manière analytique puisque la partie droite contient $p_k^t$.

Une des approches connues pour calculer les racines d'une fonction est la méthode Newton-Raphson [Press *et al.*, 1988, p. 270-274] qui évalue de manière itérative une racine $\beta$ en se basant sur une approximation de la fonction au voisinage de la racine utilisant la série de Taylor [Press *et al.*, 1988, p. 270]. L'intérêt de cette méthode est qu'elle converge vers la racine de manière quadratique lorsque les dérivées sont continues au voisinage de

cette racine. La formule de Newton-Raphson est donnée par l'équation C.3.

$$\beta^{t+1} = \beta^t - \frac{f'(\beta)}{f(\beta)} \tag{C.3}$$

Afin d'approximer le $\beta$ qui annule l'équation C.2, nous définissons la fonction $f$ et sa dérivée $f'$ telles que :

$$\begin{aligned} f(\beta) &= \frac{\partial \log p_k^t}{\partial \beta_i} \\ f'(\beta) &= \frac{\partial \partial \log p_k^t}{\partial \beta_i \partial \beta_j} \end{aligned}$$

$f(\beta)$ représente un vecteur et $f'(\beta)$ une matrice. Le développement de la dérivée $f'(\beta)$ est donné par l'équation C.4.

$$\begin{aligned} f'(\beta) &= \frac{\partial \partial \log p_k^t}{\partial \beta_i \partial \beta_j} \\ &= -(\Sigma_{k'=1}^K p_{k'}^t X_{k'j}^t X_{k'i}^t - (\Sigma_{k'=1}^K p_{k'}^t X_{k'j}^t)(\Sigma_{k'=1}^K p_{k'}^t X_{k'i}^t)) \\ &\quad - \Sigma_{k'=1}^K p_{k'}^t (X_{k'j}^t - \Sigma_{k''=1}^K p_{k''}^t X_{k''j}^t) X_{k'i}^t \\ &= -(E[X_{k'i}^t, X_{k'j}^t] - E[X_{k'i}^t] E[X_{k'j}^t]) \\ &= -Cov(X_{k'i}^t, X_{k'j}^t) \end{aligned} \tag{C.4}$$

Du fait que $f'(\beta)$ est semi-définie négative (la matrice de Covariance est positive définie ou positive semi définie et $f(\beta)$ est positive), le pas de convergence $\delta$ défini par l'équation C.5 est positif et donc garantit la convergence de l'algorithme vers un minimum.

$$\begin{aligned} \delta &= \beta^{t+1} - \beta^t \\ &= -\frac{f'(\beta)}{f(\beta)} \end{aligned} \tag{C.5}$$

Soit le vecteur gradient $\vec{g} = (\frac{\partial \log p_k^t}{\partial \beta_i})_{i \in \{1,2\}}$ et la matrice hessienne $H = (\frac{\partial \partial \log p_k^t}{\partial \beta_i \partial \beta_j})_{i,j \in \{1,2\}}$, alors la valeur de $\beta$ est obtenue en itérant l'équation C.6 jusqu'à convergence de $\beta$ vers une valeur $\beta^*$.

$$\beta = \beta - H^{-1}g \tag{C.6}$$

## C.2 Estimation du paramètre $\alpha$

La distribution des variables $C^t$ et $T^t$ est une combinaison linéaire des fréquences relatives des concepts dans le corpus $P(C)$ et de leurs fréquences étant donné les thèmes $P(C|T)$ tel que décrit par l'équation C.7.

$$P_\alpha(C^t = k|T^t = j) = \alpha P(C^t = k) + (1 - \alpha)P(C^t = k|T^t = j) \tag{C.7}$$

$\alpha$ est le paramètre de lissage que nous estimons avec l'algorithme EM [Dempster *et al.*, 1977].

L'estimation de $\alpha$ revient à maximiser une fonction auxiliaire définie par l'équation C.8 :

$$Q(\alpha, \alpha^i) = E[\log P(C^{1,\ldots,n}, X^{1,\ldots,n}|T^{1,\ldots,n}, \alpha)|c^{1,\ldots,n}, t^{1,\ldots,n}, \alpha^i] \tag{C.8}$$

$X^t$ est la variable cachée indiquant quel modèle (dépendant du thème, ou indépendant du thème) fournit la réponse correcte pour la $t^{\text{eme}}$ observation, $n$ est la longueur de la séquence d'observation, $C^{1,\ldots,n}$ et $T^{1,\ldots,n}$ sont les variables observées provenant d'une loi i.i.d, $c^{1,\ldots,n}$ est la séquence des concepts observés, $t^{1,\ldots,n}$ la séquence des thèmes observés, $\alpha$ est le paramètre de notre modèle et $\alpha^i$ est une valeur de ce paramètre.

L'algorithme EM se compose de deux étapes :

- L'étape E-step (Expectation step) dans laquelle la valeur de la fonction auxiliaire $Q(\alpha, \alpha^i)$ est calculée.

- L'étape M-step (Maximization step) choisit la nouvelle valeur $\alpha^{i+1}$ qui est une valeur de $\alpha$ maximisant $Q(\alpha, \alpha^i)$.

Dans ce qui suit, nous développons chacune de ces étapes.

### C.2.1 E-Step

Considérons $P(X^t)$ la probabilité qu'un des modèles dépendant ($X^t = 0$) ou indépendant du thème ($X^t = 1$) fournisse la bonne réponse à la $t^{\text{eme}}$ observation. L'équation C.8 se réécrit comme suit :

$$\begin{aligned}(C.8) &= \sum_{t=1}^{n} E[\log P(C^t = c^t, X^t = j | T^t = t^t, \alpha) | \alpha^i c^t, t^t],\ j \in \{0,1\} \\ &= \sum_{t=1}^{n} [P(X^t = 1 | C^t = c^t, T^t = t^t, \alpha^i) \log P(C^t = c^t, X^t = 1 | T^t = t^t, \alpha) \\ &\quad + P(X^t = 0 | C^t = c^t, T^t = t^t, \alpha^i) \log P(C^t = c^t, X^t = 0 | T^t = t^t, \alpha)]\end{aligned} \tag{C.9}$$

Soit $p^t$ la densité de probabilité définie par l'équation C.10.

$$p^t = P(X^t = 1 | C^t = c^t, T^t = t^t, \alpha^i) \tag{C.10}$$

Nous supposons que $X^t$ est indépendant de $C^t$, ainsi :

$$P(C^t = c^t, X^t = j | T^t = t^t) = P(C^t = c^t | X^t = j, T^t = t^t) P(X^t = j | T^t = t^t),\ j \in \{0,1\} \tag{C.11}$$

Nous réécrivons l'équation C.9 en remplaçant chaque composante par les équations C.10 et C.11. Le résultat est présenté par l'équation C.12.

$$
\begin{aligned}
(C.9) = \sum_{t=1}^{n} [p^t (\log P(C^t = c^t | X^t = 1, T^t = t^t, \alpha) \\
+ \log P(X^t = 1 | T^t = t^t, \alpha)) \\
+ (1 - p^t)(\log P(C^t = c^t | X^t = 0, T^t = t^t, \alpha) \\
+ \log P(X^t = 0 | T^t = t^t, \alpha))]
\end{aligned}
\tag{C.12}
$$

Les probabilités de l'équation C.12 peuvent être réécrites de la manière suivante :

- $P(X^t = 1|T^t) = P(X^t)$, tel que $X^t = 1$ implique que le modèle indépendant du thème est choisi et
- $P(C^t = c^t|X^t = 1, T^t = t^t, \alpha) = P(C^t = c^t)$ pour les mêmes raisons.
- $P(C^t = c^t|X^t = 0, T^t = t^t, \alpha) = P(C^t = c^t|T^t = t^t)$, tel que $X^t = 0$ implique que le modèle dépendant du thème est choisi.

Ainsi l'équation C.12 se réécrit par l'équation C.13 :

$$
\begin{aligned}
(C.12) = \sum_{t=1}^{n} \quad & [p^t (\log P(C^t = c^t) + \log \alpha) \\
& + (1 - p^t)(\log P(C^t = c^t) + \log(1 - \alpha))]
\end{aligned}
\tag{C.13}
$$

Ainsi la fonction auxiliaire (équation C.8) se réécrit par l'équation C.14.

$$
\begin{aligned}
Q(\alpha, \alpha^i) = \sum_{t=1}^{n} \quad & [p^t (\log P(C^t = c^t) + \log \alpha) \\
& + (1 - p^t)(\log P(C^t = c^t) + \log(1 - \alpha))]
\end{aligned}
\tag{C.14}
$$

### C.2.2 M-Step

Choisir une valeur qui maximise l'équation C.14 revient à chercher la valeur $\alpha^{i+1}$ de $\alpha$ qui annule la dérivée partielle de $Q(\alpha, \alpha^i)$. Nous développons la dérivée partielle dans l'équation

C.15 :

$$\begin{aligned} \frac{\partial Q(\alpha, \alpha^i)}{\partial \alpha} &= 0 \\ \sum_{t=1}^{n} [\frac{p^t}{\alpha} + \frac{1-p^t}{1-\alpha}] &= 0 \\ \sum_{t=1}^{n} \frac{p^t - \alpha}{\alpha(1-\alpha)} &= 0 \end{aligned} \tag{C.15}$$

Ainsi la valeur $\alpha^{i+1}$ est obtenue par l'équation C.16 :

$$\begin{aligned} \sum_{t=1}^{n} \alpha &= \sum_{t=1}^{n} p^t \\ \alpha^{i+1} &= \frac{\sum_{t=1}^{n} p^t}{n} \end{aligned} \tag{C.16}$$

la valeur de $\alpha$ est obtenue en itérant l'équation C.16 jusqu'à convergence de $\alpha$ vers une valeur $\alpha^*$.

# BIBLIOGRAPHIE


[Aberdeen *et al.*, 1996] J. Aberdeen, J. Burger, D. Day, L. Hirschman, D. Palmer, P. Robinson et M. Vilain. MITRE : Description of the Alembic System as Used in MET. In *Proceedings of the TIPSTER 24-Month Workshop*, mai 1996.

[Abney, 1996] S. Abney. Partial Parsing via Finite-State Cascades. *Natural Language Engineering*, 2(4) :337–344, 1996.

[Aitken, 2002] J. S. Aitken. Learning Information Extraction Rules : An Inductive Logic Programming Approach. In *Proceedings of the 15th European Conference on Artificial Intelligence*, F. Van Harmelen, éditeur, pages 355–359, Lyon, France, 2002.

[Appelt *et al.*, 1995] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers et M. Tyson. SRI International FASTUS System : MUC-6 Test and Analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 237–248. Morgan Kaufmann Publishers, 1995.

[Ballim et Russell, 1994] A. Ballim et G. Russell. LHIP : Extended DCG's for Configurable Robust Parsing. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, volume 1, pages 501–507, Kyoto, Japon, 1994.

[Bear *et al.*, 1992] J. Bear, J. Dowding et E. Shriberg. Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL 1992)*, pages 56–63, 1992.

[Biber, 1988] D. Biber. *Variation Across Speech and Writing*. Cambridge University Press, 1988.

[Bigi *et al.*, 2001] B. Bigi, A. Brun, J.-P. Haton, K. Smaili et I. Zitouni. A Comparative Study of Topic Identification on Newspaper and E-mail. In *String Processing and Information Retrieval (SPIRE 2001)*, pages 238–241, 2001.

[Bikel *et al.*, 1997] D. Bikel, S. Miller, R. Shwartz et R. Weischedel. NYMBLE : A High Performance Learning Name-Finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington D.C., 1997.

[Boufaden *et al.*, 1998] N. Boufaden, S. Delisle et B. Moulin. Analyse syntaxique robuste de dialogues retranscrits : peut-on vraiment traiter l'oral à partir de l'écrit ? In *Actes de la 5$^e$ conférence annuelle sur le traitement automatique des langues naturelles (TALN 1998)*, Paris, France, juin 1998.

[Boufaden *et al.*, 2001] N. Boufaden, G. Lapalme et Y. Bengio. Topic Segmentation : A First Stage to Dialog-based Information Extraction. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 273–280, 2001.

[Boufaden *et al.*, 2002] N. Boufaden, G. Lapalme et Y. Bengio. Découpage thématique des conversations : un outil d'aide à l'extraction. In *Actes de la 9$^e$ conférence annuelle sur le traitement automatique des langues naturelles (TALN 2002)*, volume I, pages 377–382, Nancy, France, juin 2002.

[Boufaden *et al.*, 2004a] N. Boufaden, Y. Bengio et G. Lapalme. Approche statistique pour le repérage de mots informatifs à partir de textes oraux. In *Actes de la 11$^e$ conférence sur le traitement automatique des langues naturelles (TALN 2004)*, pages 71–80, Fès, Maroc, avril 2004.

[Boufaden *et al.*, 2004b] N. Boufaden, Y. Bengio et G. Lapalme. Extended Semantic Tagger for Entity Extraction. In *Proceedings of the Workshop "Beyond Named Entity Recognition Semantic Labeling for NLP Tasks" held jointly with LREC 2004 conference*, pages 49–54, Lisbonne, Portugal, mai 2004.

[Boufaden, 1998] N. Boufaden. Analyse syntaxique robuste des textes de dialogues oraux. Mémoire de maîtrise, Université Laval, Québec, Canada, janvier 1998.

[Boufaden, 2003] N. Boufaden. An Ontology-based Semantic Tagger for IE System. In *41st Annual Meeting of the Association for Computational Linguistics (ACL 2003) : Student Workshop*, pages 7–14, Sapporo, Japon, juillet 2003.

[Brill, 1992] E. Brill. A Simple Rule-based Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italie, 1992.

[Byron et Stent, 1998] D. Byron et A. Stent. A Preliminary Model of Centering in Dialog. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998) and 36th Annual Meeting of the Association for Computational Linguistics (ACL 1998)*, pages 1475–1477, Montréal, Canada, 1998.

[Byron, 2002] D. K. Byron. Resolving Pronominal Reference to Abstract Entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 80–87, Philadelphie, Pennsylvanie, juillet 2002.

[Cancedda, 1999] Nicola Cancedda. Text Generation from MUC Templates. In *Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99)*, Toulouse, France, 1999.

[Childs *et al.*, 1995] L. Childs, D. Brady, L. Guthrie, J. Franco, D. Valdes-Dapena, B. Reid, J. Kielty, G. Dierkes et I. Sider. Lockheed Martin : LOUELLA PARSING and NLToolset System for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 97–112. Morgan Kaufmann Publishers, 1995.

[Chinchor et Dungca, 1995] N. Chinchor et G. Dungca. Four Scores and Seven Years Ago : The Scoring Method for MUC-6. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pages 33–38. Morgan Kaufmann Publishers, 1995.

[Cohen, 1960] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psycological Measurement*, 20(37-46), 1960.

[Crowe, 1995] J. Crowe. Constraint-based Event Recognition for Information Extraction. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, pages 296–298, Cambridge, Massachusetts, 1995.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM. *Journal of the Royal Statistical Society, Series B*, 39 :1–38, 1977.

[Eckert et Strube, 1999] M. Eckert et M. Strube. Dialogue Acts, Synchronising Units and Anaphora Resolution. In *Amstelogue'99 : Workshop on Dialogue*, pages 1–5, Amsterdam, Pays-Bas, 1999.

[Ford et Thompson, 1991] C. Ford et S. Thompson. On Projectability in Conversation : Grammar, Intonation, and Semantics. In *The Second International Cognitive Linguistics Association Conference*, août 1991.

[Freitag et McCallum, 1999] D. Freitag et A. K. McCallum. Information Extraction with HMMs and Shrinkage. *The AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[Fries, 1952] C. C. Fries. *The Structure of English : An Introduction to the Construction of English Sentences.* Harcourt, New York, 1952.

[Gildea et Palmer, 2002] D. Gildea et M. Palmer. The Necessity of Syntactic Parsing for Predicate Argument Recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL 2002)*, pages 239–246, Philadelphie, Pennsylvanie, 2002. `http://www.cs.rochester.edu/~gildea/gildea-acl02.pdf`.

[Graeme, 2004] H. Graeme. Ontology and the lexicon. In *Handbook on Ontologies*, S. Staab et R. Studer, éditeurs, International Handbooks on Information Systems, chapitre 11, pages 209–229. Springer, Berlin, 2004.

[Grice, 1975] H. P. Grice. Logic and Conversation. In *Syntax and Semantics*, P. Cole, éditeur, volume 3 de *Speech Acts*. Academic Press, New York, 1975.

[Grishman *et al.*, 2002] R. Grishman, S. Huttunen et R. Yangarber. Information Extraction from Enhanced Access to Disease Outbreak Reports. *Journal of Biomedical Informatics*, 35(4) :236–246, août 2002.

[Grishman, 1995] R. Grishman. The NYU System for MUC-6 or Where's the Syntax ? In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 167–176. Morgan Kaufmann Publishers, 1995.

[Grishman, 1998] R. Grishman. Information Extraction and Speech Recognition. In *Proceedings of the DARPA Broadcast Transcription and Understanding Workshop*, Lansdowne, Virginie, février 1998. Morgan Kaufmann Publishers.

[Gross *et al.*, 1993] D. Gross, J. Allen et D. Traum. The Trains 91 Dialogs. Trains technical note 92-1, Department of Computer Science, University of Rochester, juin 1993.

[Grosz *et al.*, 1995] B. Grosz, A. Joshi et S. Weinstein. Centering : A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2) :203–225, 1995.

[Grosz et Sidner, 1986] B. Grosz et C. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3) :175–204, 1986.

[Halliday et Hassan, 1976] M. A. K. Halliday et R. Hassan. *Cohesion in English*. Longman Group Ltd, 1976.

[Halliday, 1989] M. A. Halliday. *Spoken and Written Language*. Oxford University Press, 1989.

[Harabagiu *et al.*, 2000] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus et P. Morarescu. FALCON : Boosting Knowledge for Answer Engines. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pages 479–486, Washington, D.C., 2000. National Institute of Standards and Technology.

[Hearst, 1994] M. Hearst. Multi-paragraph Segmentation of Expository Text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, pages 9–16, Las Cruces, Nouveau-Mexique, 1994. `citeseer.nj.nec.com/151333.html`.

[Heeman *et al.*, 1996] P. Heeman, H. Kyung, L. Kim et J. Allen. Combining the Detection and Correction of Speech Repair. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996)*, pages 358–361, Philadephie, Pennsylvanie, octobre 1996.

[Heeman, 1999] P. Heeman. Modeling Speech Repairs and Intonational Phrasing to Improve Speech Recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 265–268, Keystone, Colorado, décembre 1999.

[Hindle, 1983] D. Hindle. Deterministic Parsing of Syntactic Nonfluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL 1983)*, pages 123–128, 1983.

[Hinton, 2002] G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8) :1771–1800, 2002.

[Hirschberg et Nakatani, 1996] J. Hirschberg et C. Nakatani. A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In *Proceedings of the 11th Annual Meeting of the Association for Computational Linguistics (ACL 1996)*, pages 286–293, new york, juin 1996.

[Hirschman, 1991] L. Hirschman. Comparing MUCK-II and MUC-3 Test Results and Analysis. In *Proceedings of the Third Message Understanding Conference (MUC-3)*, pages 25–30. Morgan Kaufmann Publishers, 1991.

[Hobbs *et al.*, 1997] J. R. Hobbs, D. Appelt, J. Bear, D. Israel et M. Kameyama. FASTUS : A Cascaded Finite-State Transducer for Extracting Information from Natural-Language

Text. In *Finite State Devices for Natural Language Processing*, pages 383–406. MIT Press, 1997.

[Hobbs, 2002] J. Hobbs. Information Extraction from Biomedical Text. *Journal of Biomedical Informatics*, 35(4) :260–264, août 2002.

[Huffman, 1996] S. B. Huffman. Learning Information Extraction Patterns from Examples. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 246–260. Springer, 1996.

[Humphreys et Lindberg, 1993] B. L. Humphreys et D. A. B. Lindberg. The UMLS Project : Making the Conceptual Connection Between Users and the Information They Need. *Bulletin of the Medical Library Association*, 81(2) :170, 1993.

[Ide et Véronis, 1998] N. Ide et J. Véronis. Introduction to the Special Issue on Word Sense Disambiguation : The State of the Art. *Computational Linguistics*, 24(1) :1–40, 1998.

[Kameyama, 1997] M. Kameyama. Recognizing Referential Links : An Information Extraction Perspective. In *Proceedings of the ACL/EACL 1997 Workshop on Anaphora Resolution, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, Madrid, Espagne, juillet 1997.

[Karp *et al.*, 2002] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides et S. Gama-Castro. The EcoCyc Database. *Nucleic Acids Research*, 30(1) :56–58, 2002.

[Kilgarriff et Palmer, 2000] A. Kilgarriff et M. Palmer. Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, 34(1-2) :1–13, 2000.

[Kim et Moldovan, 1995] J. T. Kim et D. Moldovan. Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. In *IEEE Transactions on Knowledge and Data Engineering*, volume 7, pages 713–724, 1995.

[Kingsbury et Palmer, 2002] P. Kingsbury et M. Palmer. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 2002.

[Krupka, 1995] G. Krupka. Description of the SRA System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 221–236. Morgan Kaufmann Publishers, 1995.

[Kubula *et al.*, 1998] F. Kubula, R. Schwartz, R. Stone et R. Weischedel. Named Entity Extraction From Speech. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginie, février 1998. Morgan Kaufmann Publishers.

[Langer, 1990] H. Langer. Syntactic Normalization of Spontaneous Speech. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 1990)*, volume 3, pages 180–183, Helsinki, Finlande, 1990.

[Lavie *et al.*, 1997] A. Lavie, D. Gates, N. Coccaro et L. Levin. Input Segmentation of Spontaneous Speech in Janus : A Speech to Speech Translation System. In *Dialogue*

*Processing in Spoken Language Systems*, E. Maier, M. Mast et S. Luperfoy, éditeurs, volume 1236 de *Lecture Notes in Artificial Intelligence*, pages 86–99. Heidelberg, Springer-Verlag édition, 1997.

[Leek, 1997] T. R. Leek. Information Extraction Using Hidden Markov Model. Mémoire de maîtrise, University of California, San Diego, Californie, 1997.

[Lesk, 1986] M. Lesk. Automatic Sense Disambiguation Using Machine Readable Dictionaries : How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC 1986)*, pages 24–26, Toronto, Canada, 1986.

[Levelt, 1989] W. J. M. Levelt. *Speaking : From Intention to Articulation*. ACL-MIT Press Series in Natural Language Processing. MIT Press, 1989.

[Levinson, 1983] S. Levinson. *Pragmatics*. Cambridge University Press, Cambridge, 1983.

[MacDonald et Zucchini, 1997] I. L. MacDonald et W. Zucchini. *Hidden Markov and Other Models for Discrete-valued Times Series*. Chapman and Hall, 1997.

[Manning, 1998] C. D. Manning. Rethinking Text Segmentation Models : An Information Extraction Case Study. Rapport technique, University of Sydney, 1998.

[Maynard et Zimmerman, 1984] W. D. Maynard et D. H. Zimmerman. Topical Talk, Ritual and the Social Organiztion of Relationships. *Social Psychology Quarterly*, 47(4) :301–306, 1984.

[McCallum *et al.*, 1999] A. McCallum, K. Nigam, J. Rennie et K. Seymore. A Machine Learning Approach to Building Domain-specific Searchs Engines. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

[Meeter et Iyer, 1996] M. Meeter et R. Iyer. Modeling Conversational Speech for Speech Recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphie, Pennsylvanie, mai 1996.

[Miller *et al.*, 1990] G. Miller, R. Beckwith, C. Fellbaum, D. Gross et K. Miller. Five Papers on WordNet. Rapport Technique CSL Report 43, Cognitive Science Laboratory, Princeton University, juillet 1990.

[Miller, 1990] G. Miller. WordNet : An On-line Lexical Database. *International Journal of Lexicography*, 3(4) :235–312, 1990.

[Morato *et al.*, 2004] J. Morato, M. A. Marzal, J. Llorens et J. Moreiro. WordNet Applications. In *Proceedings of the Second International WordNet Conference*, P. Sojka, K. Pala, P. Smrz, C. Fellbaum et P. Vodden, éditeurs, pages 270–278, Brno, République Tchèque, 2004.

[Morris et Hirst, 1991] J. Morris et G. Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Texts. *Computational Linguistics*, 17 :21–48, 1991.

[Nakajima et Allen, 1993] S. Nakajima et J. Allen. A Study on Prosody and Discourse Structure in Cooperative Dialogues. *Phonetica*, 50(3) :197–210, 1993.

[Ney *et al.*, 1995] H. Ney, U. Essen et R. Kneser. On the Estimation of 'Small' Probabilities by Leaving-One-Out. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12) :1202–1212, 1995.

[Noy et Hafner, 1997] F. N. Noy et C. D. Hafner. The State of the Art in Ontology Design : A Survey and Comparative Review. *AI Magazine*, 18(3) :53–74, 1997.

[Noy et McGuinness, 2001] N. F. Noy et D. L. McGuinness. Ontology Development 101 : A Guide to Creating Your First Ontology. Rapport Technique KSL-01-05 et SMI-2001-0880, Stanford University, mars 2001.

[Pallett, 2003] David S. Pallett. A Look at NIST's Benchmark ASR Tests : Past, Present, and Future. Rapport technique, National Institute of Standards and Technology, 2003.

[Passonneau et Litman, 1993] R. J. Passonneau et D. J. Litman. Intention-Based Segmentation : Human Reliability and Correlation with Linguistic. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 1993)*, pages 148–155, Columbus, Ohio, 1993.

[Press *et al.*, 1988] W. H. Press, S. A. Teukolsky et W. T. Vetterling. *Numerical Recipes in C*, chapitre 9. Cambridge Press University, 1988.

[Pêches et Océans Canada, 2000] Pêches et Océans Canada : Garde côtière. *SAR Seamanship Reference Manual.* Canadian Goverment Publishing – Public Works and Goverment Services Canada, Ottawa, Canada, 2000.

[Quinlan, 1986] J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1 :81–106, 1986.

[Rabiner, 1989] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[Reeker *et al.*, 1983] L. H. Reeker, M. Elena, P. E. Chmora et P. E. Blower. Specialized Information Extraction : Automatic Chemical Reaction Coding From English Descriptions. In *Proceedings of Applied Natural Language Processing (ANLP)*, pages 109–116, 1983.

[Renkema, 1993] J. Renkema. *Discourse Studies.* John Benjamins Publishing Co., 1993.

[Reynar, 1999] J. C. Reynar. Statistical Models for Topic Segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 357–364, College Park, Maryland, 1999.

[Riloff, 1993] E. Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pages 811–816, Washington, D.C., 1993.

[Riloff, 1996] E. Riloff. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pages 1044–1049, Portland, Oregon, 1996.

[Robertson et Walker, 1999] S. E. Robertson et S. Walker. Okapi/Keenbow at TREC-8. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. National Institue of Standards and Technology, 1999.

[Robinson *et al.*, 1999] P. Robinson, E. Brown, J. Burger, N. Chinchor, A. Douthat, L. Ferro et L. Hirschman. Overview : Information extraction from broadcast news. In *Proceedings of DARPA Broadcast News Worksho*, pages 27–30, Herndon, VA, 1999.

[Sacks *et al.*, 1974] H. Sacks, E. A. Schegloff et G. Jefferson. A Simplest Systematics for the Organization of Turn-taking in Conversation. *Language*, 50 :696–735, 1974.

[Sacks et Schegloff, 1973] H. Sacks et E. A. Schegloff. Opening up Closing. *Semiotica*, 7 :289–327, 1973.

[Sager *et al.*, 1987] N. Sager, C. Friedman et M. S. Lyman. *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley, Menlo Park, Californie, 1987.

[Seymore *et al.*, 1999] K. Seymore, A. McCallum et R. Rosenfeld. Learning Hidden Markov Structure for Information Extraction. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 37–42, Orlando, Floride, juillet 1999.

[Soderland *et al.*, 1995] S. Soderland, D. Fisher, J. Aseltine et W. Lenhert. CRYSTAL : Inducing a Conceptual Dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1314–1319, Menlo Park, Californie, 1995.

[Soderland, 1998] S. Soderland. Learning Information Extraction Rules for Semi-structured and Free Text. *Machine Learning*, 44(1-3) :233–272, 1998.

[Stolcke et Shriberg, 1996] A. Stolcke et E. Shriberg. Statistical Language Modeling for Speech Disfluencies. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 405–408, Atlanta, 1996.

[Stolcke, 1997] A. Stolcke. Modeling Linguistic Segment and Turn Boundaries for N-best Rescoring of Spontaneous Speech. In *Proceedings of EUROSPEECH 1997*, volume 5, pages 2779–2782, Rhodes, Grèce, 1997.

[Sundheim, 1995] B. Sundheim. Design of the MUC-6 Evaluation. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pages 1–12. Morgan Kaufmann Publishers, 1995.

[Surdeanu *et al.*, 2003] M. Surdeanu, S. Harabagiu, J. Williams et P. Aarseth. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, E. Hinrichs et D. Roth, éditeurs, pages 8–15, 2003. `http://www.aclweb.org/anthology/P03-1002.pdf`.

[Surdeanu et Harabagiu, 2002] M. Surdeanu et S. M. Harabagiu. Infrastructure for Open-Domain Information Extraction. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, M. Mitchell, éditeur, pages 325–330, San Diego, Californie, 2002.

[Takagi et Itahashi, 1996] K. Takagi et S. Itahashi. Segmentation of Spoken Dialogue by Interjection, Disfluent Utterances and Pauses. In *Proceedings of the 4th International*

*Conference on Spoken Language Processing (ICSLP 1996)*, pages 693–697, Philadelphie, Pennsylvanie, octobre 1996.

[Traum et Heeman, 1997] D. Traum et P. Heeman. Utterance Units in Spoken Dialogue. In *Dialogue Processing in Spoken Language Systems*, E. Maier, M. Mast et S. Luperfoy, éditeurs, volume 1236 de *Lecture Notes in Artificial Intelligence*, pages 125–140. Springer-Verlag, Heidelberg, 1997.

[Weischedel, 1995] R. Weischedel. BBN : Description of the PLUM System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 55–77. Morgan Kaufmann Publishers, 1995.

[Witten et Frank, 2000] I. H. Witten et E. Frank. *Data Mining : Practical Machine Learning Tools and Techniques with JAVA Implementations*, chapitre 8, pages 265–319. Morgan Kaufmann Publishers, 2000.

[Youmans, 1990] G. Youmans. Measuring Lexical Style and Competence : Token Vocabulary Curve. *Style*, 24 :584–599, 1990.