Repérage de mots informatifs dans les textes conversationnels

Narjès Boufaden* — **Guy Lapalme*** — **Yoshua Bengio****

* *Laboratoire RALI*
** *Laboratoire LISA*
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, Québec, H3C 3J7 Canada
{boufaden,lapalme,bengioy}@iro.umontreal.ca

RÉSUMÉ. *Nous présentons les résultats d'une approche d'étiquetage sémantique développée pour le repérage de mots informatifs à partir de textes conversationnels. Ce travail entre dans le cadre du développement d'un système d'extraction d'information dans le domaine de la recherche et sauvetage maritime. Il s'agit de détecter et d'annoter les mots pertinents avec des étiquettes sémantiques correspondant aux concepts d'une ontologie du domaine. Notre méthode combine une approche symbolique basée sur un automate à états finis et une approche statistique exploitant deux types d'information : les vecteurs de scores de similarité et le contexte discursif représenté par le thème. Le F-score obtenu sur des transcriptions manuelles de conversations téléphoniques dans le domaine de la recherche et sauvetage maritime est de 82,2 %.*

ABSTRACT. *We present the results of a semantic tagger we developped for the detection of informative words from manually transcribed conversations. This work is part of a project for developing an information extraction system in the field of maritime Search And Rescue (SAR). Our purpose is to automatically detect relevant words and annotate them with concepts from a SAR ontology. Our approach combines a finite state automata and a statistical approach based on similarity score vectors and topical information. We tested our approach on manually transcribed telephone conversations in the domain of maritime search and rescue, and succeeded in semantically annotating relevant informations with an F-score of 82.2 %.*

MOTS-CLÉS : *Étiquetage sémantique, extraction d'information, analyse conversationelle.*

KEYWORDS: *Semantic tagging, information extraction, conversational analysis.*

1. Introduction

Le repérage de mots informatifs consiste à détecter et étiqueter sémantiquement des expressions pertinentes pour un domaine particulier. Cette problématique s'introduit dans le cadre de notre approche d'extraction d'information (EI) proposée pour des textes conversationnels spécialisés. Cette étape précède celle de l'apprentissage de patrons d'extraction[1] nécessaires pour extraire les informations remplissant les champs de formulaires[2] prédéfinis.

Une limite des approches d'EI développées pour les textes structurés[3] tels que les dépêches journalistiques est la difficulté à définir des patrons pour l'ensemble des événements que le système d'EI doit détecter. Les variations langagières compliquent la couverture exhaustive des mots caractérisant ces événements et limitent l'apprentissage des patrons à ceux fréquemment observés dans un corpus. L'approche de repérage de mots informatifs que nous proposons s'attaque aux problèmes des variations langagières pour améliorer l'étape d'apprentissage des patrons d'extraction dans notre approche.

Notre travail est motivé par une analyse comparative de notre corpus et d'un corpus de dépêches journalistiques extraites du corpus Lob[4] montrant que nos textes présentent un taux de variations langagières plus important que ceux des dépêches journalistiques. Cette caractéristique, combinée avec la taille modeste de notre corpus, ont nécessité un étiquetage sémantique des expressions pertinentes pour contourner les problèmes engendrés par les variations langagières.

Nous proposons d'étiqueter sémantiquement deux types d'expressions : les termes du domaine et les expressions sémantiquement similaires à ces termes. Le premier type représente l'ensemble de mots spécialisés définis dans le lexique du domaine que nous avons modélisé dans une ontologie. Le second type représente les expressions complexes (variations langagières) non définies dans l'ontologie du domaine.

La prochaine section décrit les caractéristiques des textes conversationnels dans le domaine de la recherche et sauvetage maritime. La section 3 présente notre approche d'EI pour les textes conversationnels spécialisés ainsi que l'architecture du système d'étiquetage sémantique. La section 4 présente les sources de connaissance utilisées par l'étiqueteur sémantique. En particulier, nous présentons l'approche utilisée pour la conception de l'ontologie du domaine et le dictionnaire-thésaurus Wordsmith. La

1. 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 des informations pertinents au domaine d'application.
2. Un formulaire est une représentation structurée d'information se rapportant à un événement donné pour un domaine particulier.
3. Nous considérons les textes structrés comme l'ensemble de textes composés de phrases grammaticales où les relations de type **sujet-verbe-objet** sont la règle générale.
4. `http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html`

section 5 décrit le module de détection de thème et la section 6 décrit l'étiquetage des termes du domaine et les résultats obtenus. La section 7 aborde la problématique de l'étiquetage des expressions sémantiquement similaires aux termes du domaine. Enfin, la section 8 analyse et critique notre approche.

2. Étude du corpus

Notre corpus est une collection de 95 conversations téléphoniques du domaine de la recherche et sauvetage totalisant 30 000 mots. Ces conversations ont été transcrites et segmentées par des personnes non familières avec le domaine de la recherche et sauvetage. 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 inconsistances dans les noms de personnes utilisés pour l'anonymisation. Un extrait de conversation est donné dans le tableau 1 où sont soulignées les informations étiquetées sémantiquement et extraites par notre système. Les lignes pointillées sont les frontières des unités thématiques détéctées de manière automatique par un module développé dans le cadre de ce projet (Boufaden *et al.*, 2001).

2.1. *Textes spécialisés*

Ces textes se caractérisent, d'une part, par un vocabulaire spécialisé composé de mots et expressions telles que le type d'incident (`missing` et `overdue`), sa cause (`dead battery`, `fire` et `engine failure`), le moyen utilisé pour signaler l'incident (`witness report`, `flares` et ELT-Emergency Locator Transmitter), le moyen utilisé pour une recherche (`radar`, SARSAT-Search And Rescue Satellite-Aided Tracking, `goggles` et `divers`), les noms d'avions et de bateaux alloués pour la recherche (`Aurora`, `King Air` et `Challenger`), leur disponibilité (`ready` et `available`), la description du bateau en détresse (`Zodiacs`, `20-footer`, `yacht`, `red` ou `trims`) son état (`sinking` ou `drifting`) ainsi que les conditions météorologiques (`foggy`, `haze`) et l'état d'avancement et le résultat d'une mission de recherche (`completed` ou `nothing`). Certains termes sont des mots techniques (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). Enfin, certaines constructions syntaxiques sont typiques du domaine telles que celles décrivant une position en termes de latitude et longitude, telles 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. Ce sont aussi des adjectifs représentant par exemple les conditions météorologiques comme le montre le tableau 2 ou des verbes décrivant des incidents.

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.

Tableau 1. *Exemple de conversation du domaine de la recherche et sauvetage maritime où les termes du domaine sont annotés sémantiquement. Les étiquettes sémantiques sous les barres de soulignement désignent des concepts de l'ontologie du domaine. Les lignes pointillées indiquent les frontières des unités thématiques*

No	Loc	Énoncé
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.

Tableau 2. *Extraits de quatre conversations où les conditions météorologiques sont exprimées par des adjectifs*

2.2. *Transcriptions de l'oral*

Nos textes sont des transcriptions de l'oral segmentés en énoncés correspondant à des unités prosodiques identifiées par une 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.

Nous avons relevé 2,3 % de taux d'erreurs de transcriptions, certaines sont des erreurs d'orthographe comme dans we have a south east flowing à la place de we have a south east blowing tandis que d'autres sont dues au manque de connaissances du domaine comme dans Doray et Dory[5].

Nous avons effectué une analyse comparative des variations lexicales entre notre corpus et un échantillon de 44 dépêches journalistiques analysées par (Biber, 1988) pour mieux cerner la problématique de l'EI et l'étiquetage sémantique des textes conversationnels spécialisés. Nous avons calculé le pourcentage des noms et adjectifs prédicatifs par rapport au nombre total de mots de notre coprus ainsi que le ratio type/occurence. Des résultats obtenus nous tirons deux constatations :

– La diversité du vocabulaire dans notre corpus est plus importante que celle de l'échantillon de dépêches journalistiques, avec respectivement un ratio type/occurence de $5,8$ contre celui de $5,5$ pour les dépêches journalistiques.

– L'information est véhiculée par des noms avec un taux de $17,5\%$ pour notre corpus contre 22% pour les dépêches, mais aussi par des adjectifs prédicatifs en nombre plus important dans notre corpus avec un taux de $1,3\%$ contre $0,3\%$ pour les dépêches journalistiques.

5. *Dory* est un type de bateau de pêche.

Ainsi, nous avons observé que les textes conversationnels présentent une diversité du vocabulaire aussi importante que celle des dépêches journalistiques. L'information est véhiculée par les noms en moindre proportion que celle pour les dépêches journalistiques et que les adjectifs prédicatifs sont des vecteurs d'information en proportion plus importante que celle pour les dépêche journalistiques. Enfin, les classes d'information ne se limitent pas à des noms de personnes ou d'organisation, elles incluent des classes sémantiques propres au domaine telles que les classes AVION, BATEAU et INCIDENT.

Les approches d'EI incluent une étape d'extraction des entités nommées pour contrer les variations langagières des concepts PERSONNE, LIEUX et ORGANISATION. C'est une étape simplifée de l'étiquetage sémantique se limitant au noms propres qui sont des instances de ces concepts. Nous proposons d'étendre cette tâche d'étiquetage sémantique à d'autres classes sémantiques dépendantes du domaine et à des catégories morphosyntaxiques autres que les noms. Le résultat escompté est la généralisation de l'utilisation de classes de mots dans la tâche d'EI pour améliorer le processus d'apprentissage des patrons d'extraction.

3. Approche d'EI pour les textes conversationnels spécialisés

Notre approche d'EI repose sur trois étapes en cascade :

La segmentation des textes conversationnels inclut la détection des paires d'adjacence qui définissent les unités minimales de la conversation que nous utilisons pour l'apprentissage des patrons d'extraction, la segmentation en unités thématiques délimitant les unités discursives et la détection des thèmes utilisés à l'étape de l'étiquetage sémantique.

L´étiquetage sémantique partiel repère les mots informatifs et leur attribue une classe sémantique parmi celles définies dans l'ontologie du domaine.

L'apprentissage des patrons d'extraction prend en entrée les séquences de concepts étiquetant les expressions pertinentes contenues dans une paire d'adjacence. Ces concepts sont eux-même étiquetés par des rôles sémantiques dépendants du formulaire d'extraction considéré dans la tâche d'apprentissage.

Dans cet article nous discutons l'étape d'étiquetage sémantique ; une description complète des première et troisième étapes est donnée respectivement dans (Boufaden *et al.*, 2002) et (Boufaden *et al.*, 2005) .

Nous avons divisé l'étiquetage sémantique en deux étapes complémentaires :

1) Étiquetage des termes du domaine en utilisant l'ontologie.

2) Sélection et filtrage des expressions sémantiquement similaires aux termes du domaine en utilisant une approche statistique qui combine des scores de similarités entre l'expression et les concepts de l'ontologie et le contexte discursif. Les scores

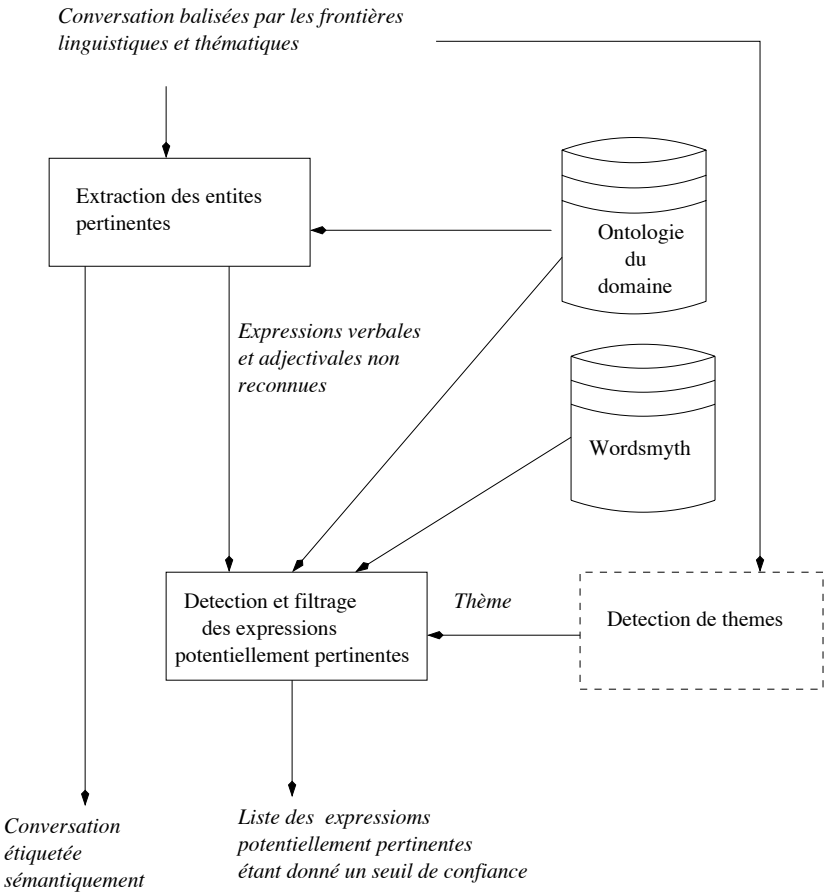

Figure 1. *Architecture de l'étiqueteur sémantique*

de similarité sont obtenus en exploitant l'ontologie et le dictionnaire-thésaurus Wordsmyth, tandis que le contexte discursif est fourni par le système de détection de thème (section 5).

La figure 1 présente l'organisation de ces étapes. Dans la section suivante nous décrivons les connaissances utilisées, tandis que les différents modules de l'architecture sont décrits dans les sections 5, 6 et 7.

4. Connaissances utilisées

Deux types de ressources ont été utilisées dans notre application. La première modélise le lexique du domaine dans une ontologie. La seconde est le dictionnaire-

thésaurus Wordsmith représentant les connaissances du monde que nous utilisons pour le traitement des variations langagières.

4.1. *Ontologie du domaine*

Afin d'effectuer l'étiquetage sémantique des informations pertinentes au domaine, nous avons formalisé le vocabulaire spécialisé du domaine dans une ontologie regroupant en classes les mots utiles pour l'apprentissage des patrons d'extraction. Cette ontologie se base sur les termes spécialisés décrits dans les manuels du domaine de la recherche et sauvetage maritime (et Océans Canada : Garde côtière, 2000).

Le choix d'une ontologie pour représenter les connaissances du domaine est motivé par trois considérations. Premièrement, une ontologie est une spécification formelle de la conceptualisation d'une réalité. Elle comprend un ensemble d'instance de mots clé regroupés en concepts que nous voulons détecter dans notre corpus. Deuxièmement, elle 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). Troisièmement, une ontologie présente des relations hiérarchiques telles que *is-a* et *part-of* que nous exploitons pour réduire le nombre de classes de mots utilisées pour l'apprentissage de patrons d'extraction.

4.1.1. *Étapes de la conception*

Notre approche pour la conception de l'ontologie combine les approches descendante et ascendante (Noy *et al.,* 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 tels que les ressources disponibles (avion, bateau, hélicoptère), les lieux, les organisations et les types d'incident.

Déterminer la portée de l'ontologie L'ontologie contient un ensemble représentatif des informations ciblées par le processus d'EI. Elle doit notamment permettre 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 conditions météorologiques ?

- `High sea`, `Fog`, `Haze`.

– Quels sont les incidents ?

- `Overdue`, `Missing`, `Fire`.

– Quels sont les descriptions de 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 instances d'un concept, par exemple les termes `fog`, `rain`, `wind` et `visibility` sont des instances du concept WEATHER-CONDITIONS, tandis que les termes `broken`, `disabled` et `fire` sont des instances du concept INCIDENT.

La relation *part-of* décrit le lien de composition 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 (et Océans Canada : Garde côtière, 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`).

Nous avons enrichi chaque instance de l'ontologie avec une liste de trois synonymes ou mots similaires extraits du dictionnaire thésaurus Wordsmyth dans le but de calculer des scores de similarité pour détecter les expressions sémantiquement similaires aux termes. Dans la section 7.1, nous montrons que cette procédure permet de réduire les chances d'obtenir des scores nuls entre deux expressions sémantiquement similaires. L'étape d'enrichissement a été effectuée de manière semi-automatique[6]. Ensuite, une étape manuelle de désambiguisation a permis de garder uniquement les définitions propres au domaine de la recherche et sauvetage.

4.1.2. *Implémentation*

Nous avons identifié 13 classes principales, groupées sous les super-classes **Physical Entity** et **Conceptual Entity**. Les termes sont organisés en 51 classes selon une hiérarchie *is-a* (43 classes) et une autre *part-of* (8 classes). L'ontologie est un arbre de profondeur 4 contenant 783 feuilles. La figure 2 illustre la hiérarchie *is-a* de cette ontologie. Chaque feuille de l'ontologie contient un terme, la définition textuelle du

6. 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.

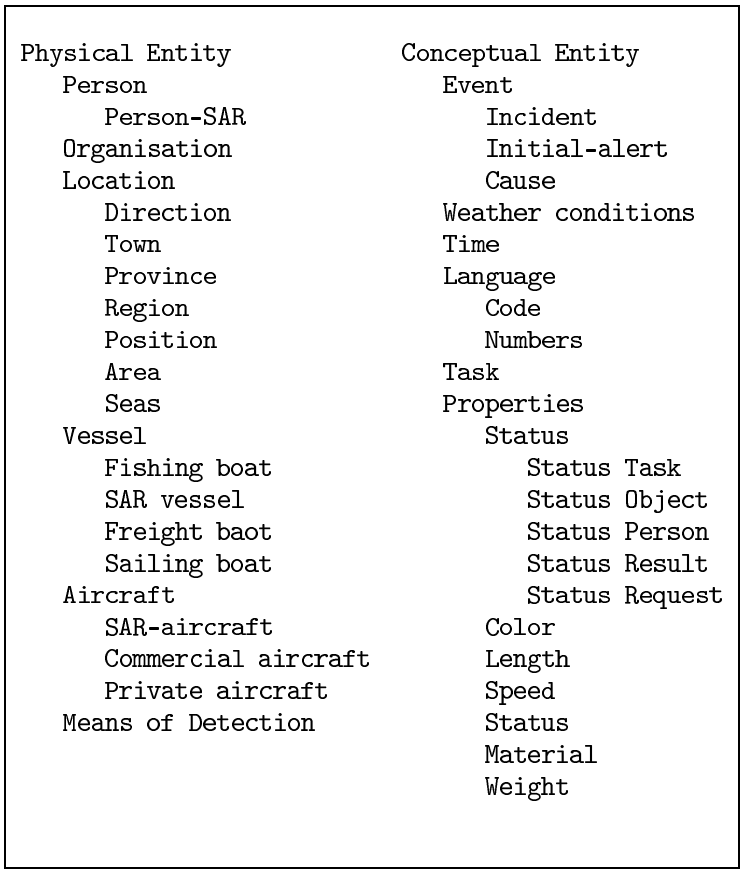

Figure 2. *Hiérarchie is-a de l'ontologie du domaine de la recherche et sauvetage maritime. La première colonne représente les classes des différents niveaux sous la classe Physical Entity, tandis que la deuxième colonne contient les classes sous la classe Conceptual Entity*

sens propre au domaine et une liste de 3 synonymes et/ou mots similaires, tous extraits de Wordsmyth.

4.2. *Dictionnaire-thésaurus Wordsmyth*

Pour détecter et étiqueter les expressions sémantiquement similaires aux termes du domaine nous avons besoin de trois types d'information :

– La définition textuelle d'un mot.

– La liste des synonymes par sens du mot.

– La liste des mots similaires qui sont des hyponymes, des hypéronymes ou mots connexes, pour chaque sens du mot.

La définition textuelle est utilisée dans le calcul du score de similarité selon l'approche Lesk (Lesk, 1986), tandis que les mots similaires et les synonymes améliorent les chances de détecter les expressions nominales, adjectivales ou verbales sémantiquement similaires aux termes du domaine. Ainsi, le choix de cette ressource est conditionné par la couverture et la complétude des informations pour ces trois types d'expression.

Bien que peu de ressources soient disponibles publiquement comme WordNet (Miller *et al.,* 1990), nous avons renoncé à son utilisation en raison de sa couverture limitée des sens, des synonymes et aussi des mots similaires pour les adjectifs. Nous avons choisi le dictionnaire-thésaurus Wordsmyth[7],[8] pour les raisons suivantes :

1) À la différence de WordNet, il couvre à peu près tous les sens d'un adjectif de manière exhaustive sans se limiter aux sens 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 7).

À titre d'exemple, la figure 3 montre une entrée de Wordnet et une de Wordsmyth pour l'adjectif `overdue` qui est un type d'incident défini dans le lexique du domaine. En comparant ces entrées, nous remarquons que le sens propre au domaine correspond à la deuxième définition donnée dans Wordsmith, tandis que Wordnet fournit un seul sens qui n'est pas le bon.

5. Identification des thèmes

Cette étape détermine le sujet (le thème) de chaque unité thématique d'une conversation relatant un événement pertinent dans le domaine de la recherche et sauvetage, c'est-à-dire les thèmes *Incident*, *Search-unit*, *Search-mission* et *Missing-object* ou le thème *Other* regroupant tout autre thème. Les unités thématiques sont obtenues de manière automatique grâce à un segmenteur que nous avons développé dans le cadre de ce projet (Boufaden *et al.,* 2001, Boufaden *et al.,* 2002) qui détecte les changements

7. Wordsmyth nous a été prêté gracieusement pour la durée de ce travail.
8. `http://www.wordsmyth.net`

Wordnet

```
1. delinquent, overdue - (past due; not paid at the scheduled
time); "an overdue installment"; "a delinquent account"
```

Wordsmyth

```
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
```

Figure 3. *Comparaison des définitions de l'adjectif* `overdue` *extraites à partir de Wordnet et du dictionnaire-thésaurus Wordsmyth. Les acronymes* `ENT`, `SYL`, `PRO`, `POS`, `INF`, `DEF`, `EXA`, `SYN`, `SIM` *sont respectivement l'entrée, la séparation en syllabes, 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*

de thèmes[9] avec un rappel de 61,4% et une précision de 67,3%. Les lignes horizontales dans le tableau 1 montrent les changements de thèmes. Par exemple, les énoncés 4-5 décrivent un incident (thème *Incident*), tandis que les énoncés 6-10 une mission de recherche (*Search-mission*).

Pour identifier le thème d'un segment situé entre deux frontières, nous avons utilisé la cooccurrence des concepts étiquetant les termes présents dans ce segment. Les concepts considérés pour cette tâche sont uniquement ceux étiquetant les termes pré-

9. Les changements de thèmes sont représentés par une étiquette. Quatre autres étiquettes sont utilisées pour distinguer les énoncés qui font partie d'un segment de thème de celles qui terminent un segment, qui initient une conversation ou qui indiquent la fin d'une conversation

Concept	Thème
INCIDENT	*Incident*
AIRCRAFT-SAR	*Search-unit*
VESSEL	*Missing-object*, *Search-unit*, *Incident*
TIME	*Search-unit*, *Incident*
MEANSOFDETECTION	*Search-unit*
STATUS	*Search-unit*
LOCATION	*Incident*, *Search-unit*
WEATHER	*Incident*, *Search-unit*

Tableau 3. *Exemples de concepts utilisés pour l'identification des thèmes et les thèmes pouvant être associés*

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 %

Tableau 4. *Pourcentage de la précision et du rappel par classe de thèmes avec l'algorithme ID3*

sents dans l'ontologie. Le tableau 3 donne des exemples de concepts utilisés pour l'identification des thèmes et les thèmes associés.

Nous avons expérimenté deux approches : un classifieur bayésien naïf et un arbre de décision généré avec l'algorithme ID3 (Quinlan, 1986) qui s'est avéré être le plus performant avec un F-score[10] moyen de 81,7 %. Les résultats de la classification par thème sont présentés au tableau 4.

L'algorithme ID3 montre une meilleure performance que le classifieur bayésien naïf car certains concepts sont plus discriminants que d'autres pour la classification des thèmes. Par exemple, 78 % des occurrences de l'entité MEANSOFDETECTION apparaissent dans le thème *Search-unit*.

10. Le F-score utilisé est défini par la formule $F = \frac{(\beta+1)P.R}{\beta^2.P+R}$ et $\beta = 0.5$ où P représente la précision et R le rappel.

Nous remarquons que le taux d'erreurs le plus important 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 classées *Other*. L'analyse de erreurs montre deux causes possibles. 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). Les chances d'observer des coocurrences de concepts sur des petites unités thématiques est inférieure à celles observées sur des longues unités. D'autre part, moins de concepts discriminants apparaissent avec le thème *Incident*, car les incidents sont souvent décrits par des adjectifs ou des verbes (`overdue`, `missing`) moins présents que les noms dans l'ontologie.

6. Étiquetage des termes du domaine

Cette étape de l'étiquetage sémantique étend la tâche d'extraction d'entités nommées telles que décrites dans (Sundheim, 1995). Elle reconnaît des entités nommées telles que les lieux, les personnes, les organisations et d'autres entités propres au domaine telles que les noms d'avions, de bateaux et de matériel de détection. Les termes considérés sont des noms, des verbes (`drifting`) et des adjectifs prédicatifs (`foggy`) non ambigus définis dans l'ontologie.

6.1. *Approche*

Notre approche se base sur un automate à états finis qui prend en entrée une analyse syntaxique partielle d'un énoncé. L'étiquetage morphosyntaxique utilise le système de Brill (Brill, 1992) entraîné sur le corpus Brown, tandis que l'étiquetage morphosyntaxique a été réalisé avec le système d'analyse syntaxique partielle CASS de Abney (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 des mots composés NP → NP NP. Notre approche s'effectue 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 mots composés tels que `Rescue Coordination Centre`. Également, cette étape reconnaît les expressions spécialisées du domaine qui font référence à des positions ou des périodes de temps telles que `54 degree` ou `18 Zulu on 8`.

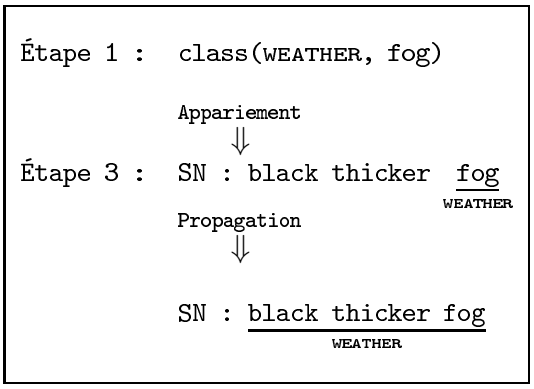

Figure 4. *Le syntagme nominal* SN *: black thicker fog est étiqueté avec le concept* WEATHER. *La première étape de l'analyse sémantique reconnaît la tête fog comme un type de conditions climatiques. La troisième étape propage le concept à tout le syntagme nominal*

3) La propagation de l'étiquette sémantique de la tête du syntagme à tout le syntagme.

La figure 4 donne un exemple de propagation du trait sémantique WEATHER.

6.2. *Expérience et résultat*

Nous avons évalué notre système sur 64 conversations totalisant 1021 expressions qui sont des termes de l'ontologie. Nous avons basé l'évaluation sur les scores de rappel et précision où 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 % obtenus lors de MUC 6. Parmi les erreurs fréquentes qui ont influencées le rappel, lequel est relativement bas sont :

– La présence de noms propres transcrits de manières différentes tels que les noms de lieux (`St-Johns` et `Saint Johns`) ou les noms d'avions (`Hurk` et `Hercules`).

– Les 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.

7. Étiquetage des expressions non couvertes par l'ontologie

Cette étape détecte et étiquette les expressions qui sont des instances des concepts de l'ontologie du domaine mais qui ne font pas partie du lexique du domaine. Le but étant d'associer une expression à un concept de l'ontologie même si celui-ci ne figure pas explicitement dans le lexique du domaine. Les expressions visées par ce processus sont les noms communs, les adjectifs et les verbes qui sont pertinents pour le domaine. La table 2 montrent des expressions qui sont des instances du concept WEATHER exprimés sous différentes variations langagières.

Nous voulons étiqueter ces expressions avec les concepts de l'ontologie les plus vraisemblables étant donné le contexte discursif.

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^t, W^t,T^t sont des variables aléatoires associées respectivement aux concepts, mots et thèmes du corpus. $C^{1,\ldots,n}$ est la séquence des concepts qui étiquettent les mots $W^{1,\ldots,n}$ et $T^{1,\ldots,n}$ la séquence des thèmes des contextes discursifs des $W^{1,\ldots,n}$.

L'identification d'un concept k est conditionnée par deux sources d'information complémentaires : le mot w, qui apparaît dans un énoncé, et un thème j de cet énoncé. Dans notre approche, nous exploitons le lien de similarité entre un mot $W^t = w$ et un concept $C^t = k$ et la fréquence des concepts étant donné un thème $T^t = j$.

Une approche statistique qui formalise cette contrainte est celle basée sur le produit de modèles probabilistes (Hinton, 2002). Ce modèle exploite l'expertise de chaque composante dans une problématique plus simple et la combine de manière à ce que seules les données qui maximisent les probabilités générées par chaque composante soient retenues. La formulation du produit de probabilités qui correspond à notre problématique est la suivante :

$$P(C^t = k|W^t, T^t) \quad = \quad \frac{P(C^t = k|W^t)^{\beta_1} P(C^t = k|T^t)^{\beta_2}}{\Sigma_{l=1}^{K} P(C^t = l|W^t)^{\beta_1} P(C^t = l|T^t)^{\beta_2}} \qquad [1]$$

k est un des concepts de l'ontologie, $P(C^t = k|W^t)$ représente la probabilité d'observer le concept k étant donné un mot W^t, que nous estimons à la section 7.1, et $P(C^t = k|T^t)$ est la probabilité d'observer le concept k étant donné un thème T^t que nous estimons à la section 7.2. K est le nombre de concepts dans l'ontologie. Les coefficients β_1 et β_2 sont des poids indépendants des concepts qui reflètent la contribution de chaque modèle 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 [1] 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 δ sont étiquetés. Ainsi, une seconde condition s'ajoute à notre système d'étiquetage :

$$P(C = k^*|W^t, T^t) > \delta, \text{ avec} \qquad [2]$$
$$k^* = \underset{k}{\operatorname{argmax}}\, P(C^t = k|W^t, T^t)$$

Les paramètres β_1 et β_2 ont été estimés avec l'algorithme de Newton-Raphson (Press *et al.,* 1988) sur un corpus d'entraînement contenant des observations $o^t = (W^t, C^t, T^t)$ où W^t est un mot que nous voulons étiqueter, C^t est le concept correct attribué à W^t et T^t est le thème du contexte discursif de W^t.

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

7.1. *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) correspondant au nombre de mots [11] (lemmatisés) en commun contenus dans la définition textuelle de deux mots w et de u pour un sens donné, tel que décrit par l'équation [3].

$$sim(w, u) = \frac{\mid D_w \mid \cap \mid D_u \mid}{min(\mid D_w \mid, \mid D_u \mid)} \qquad [3]$$

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 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 formulé par l'équation [4].

$$\begin{aligned} P(C|w) &= P(C|s(w) = s_1, s_i, \ldots, s_l) \\ &= \sum_{s_l \in S(w)} P(C|w)P(s_l|w) \end{aligned} \qquad [4]$$

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

$s_l \in S(w)$ 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 le deux étapes suivantes :

Soit :

– $C = \{c_1, \ldots, c_t, \ldots, c_K\}$ l'ensemble des K concepts des classes principales de l'ontologie.

– $I = \{i_1, \ldots, i_l, \ldots, i_N\}$ l'ensemble des N instances d'un concept c_t.

– $M = \{m_1, m_2, \ldots, m_T\}$ l'ensemble des T mots synonymes et des mots similaires d'une instance i_l.

– $S(w) = \{s_1, s_2, \ldots, s_P\}$ l'ensemble des P sens d'un mot w.

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 [3]. Le calcul des scores de similarité $sim(m, s)$ pour chaque $m \in M$ définit un vecteur de scores de similarité $\vec{v_{M,s}}$ 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 $\vec{v_{M,s}}$ qui a donné de meilleurs résultats que la moyenne.

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 $\vec{v_{I,s}}$. La moyenne des scores de ce vecteur détermine le score de similarité $sim(c_t, s)$ entre le concept c_t et le sens 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 toutes** instances $i_l \in I$ **Faire**
2: **Pour toutes** 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: $sim(s, i_l) = médiane(\vec{v_{M,s}})$
7: **Fin Pour**
8: $\vec{v_{I,s}} \stackrel{\text{def}}{=} (sim(i_1, s), \ldots, sim(i_N, s))$
9: $sim(s, c_t) = max(\vec{v_{I,s}})$

Le choix du meilleur score de similarité entre w et le concept c_t repose sur une heuristique qui donne la priorité au score de similarité le plus significatif entre w et une instance de ce concept.

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 (voir (Ide *et al.,* 1998) pour une revue de l'état de l'art).

Ainsi, nous supposons que les sens d'un mot sont équiprobables. Ce qui se traduit par l'équation suivante :

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

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 [5].

$$P(C|w_i^l) = \frac{sim(w_i^l, C) + \epsilon}{\sum_{t=1}^{K} sim(w_i^l, C^t) + K\epsilon} \qquad [5]$$

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

Ces approximations modifient l'équation [4] comme suit :

$$\begin{aligned} P(C|w) &= \sum_{w_i \in S(w)} P(C|w_i) P(w_i|w) \\ &= \frac{1}{\mid S(w) \mid} \sum_{w_i \in S(w)} \frac{sim(w_i, C) + \epsilon}{\sum_{t=1}^{K} sim(w_i, C^t) + K\epsilon} \end{aligned} \qquad [6]$$

7.2. *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 5 donne un exemple de faux positifs où, dans le premier énoncé, le verbe `land` n'est pas pertinent, tandis que dans le deuxième énoncé il représente une instance du concept STATUS.

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 avons remplacé 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 [7].

$$P_\alpha(C^t|T^t) = \alpha P(C^t) + (1-\alpha)P(C^t|T^t) \quad [7]$$

α est le coefficient de pondération estimé par l'algorithme EM (Dempster *et al.*, 1977) et les fréquences relatives sont définies par :

– $P(C = k)$, 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 de l'ontologie (figure 2).

– $P(C = k|T = j)$, 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.

7.3. *Expériences et résultats*

Nous avons entraîné notre modèle d'étiquetage des expressions sémantiquement similaires sur un corpus d'entraînement constitué de 3413 mots extraits de 65 % des 64 conversations annotées manuellement avec les concepts de l'ontologie et les thèmes. Les expressions annotées manuellement sont des instances des concepts de l'ontologie mais qui ne font pas partie de l'ontologie. Ce choix a eu pour conséquence 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 δ 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. Les mots étiquetés manuellement ont comme étiquette un concept de l'ontologie (figure 2) s'ils sont pertinents ou ont l'étiquette OTHER qui rassemble toutes les expressions qui ne sont pas pertinentes au domaine. Cela signifie

Thème	No Loc Énoncé
Incident	7 a : On the way to go, he AIRCRAFT had to land in emergency in the STATUS 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 to land there AIRCRAFT *

Tableau 5. *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 comme décrivant l'état d'un objet*

	$P(C^t\|T^t)$		$P(C^t\|W^t)$		$P(C^t\|T^t, W^t)$	
Modèles	Précision	Rappel	Précision	Rappel	Précision	Rappel
mots informatifs	51,1 %	51,1 %	70,6 %	70,6 %	86,0 %	67,4%

Tableau 6. *Résultats du modèle* $P(C^t|T^t, W^t)$ *pour le seuil de confiance* $\delta = 0,35$

qu'un même mot peut selon le thème 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 informations ciblées par le processus d'EI.

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 et celle de notre modèle d'étiquetage. Le tableau 6 donne le rappel et la précision obtenus pour le seuil δ=0,35.

La précision et le rappel du modèle de classification des concepts étant donné les thèmes $P(C^t|T^t)$, ainsi que celui de la classification des mots par concepts $P(C^t|W^t)$ affichent un même score car les deux modèles formalisent une tâche de classification pour laquelle chaque modèle génère toujours une réponse. Le résultat 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 OTHER et STATUS.

Le modèle exponentiel $P(C^t|T^t, W^t)$ a une meilleure performance, avec un F-score de 75,3 %, que celle de notre *baseline* $P(C^t|T^t)$ et le modèle basé uniquement sur la similarité $P(C^t|W^t)$, 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 modèle composé de 7,1 %.

L'analyse des erreurs du modèle combiné montre que celles-ci sont en partie dues aux erreurs d'étiquetage lexical causées par les irrégularités langagières de l'oral influençant le score de similarité. Aussi, à cause de la taille modeste de notre corpus d'entraînement, nous avons opté pour des paramètres β_1 et β_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 β_1 et β_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.

Afin d'évaluer la performance du système, nous avons comparé nos résultats avec ceux obtenus lors de MUC 6 pour la tâche d'extraction d'entités nommées. Le F-score obtenu pour les mots informatifs est de 74,65 % alors que le meilleur résultat obtenu lors de MUC 6 est de 96,4 %.

8. Conclusion

Nous avons proposé une approche d'étiquetage sémantique étend l'extraction des entités nommées. Le F-score moyen incluant les résultats des deux modules d'étiquetage sémantique (section 6 et 7) est de 82,2 % avec un rappel de 74,65 % et une précision de 89,9 %. Nous avons développé une ontologie du domaine pour systématiser 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 [12].

Nous avons utilisé l'ontologie pour l'étiquetage sémantique des termes du domaine, mais aussi pour calculer des probabilités de similarité entre des expressions du corpus et les concepts de l'ontologie, tandis que le contexte discursif a permis d'écarter les faux positifs.

Le résultat obtenu est inférieur au meilleur résultat obtenu pour l'extraction des entités nommées lors de MUC 6, cependant notre tâche est plus ambitieuse car elle s'attaque aux expressions non couvertes par l'ontologie, en plus d'étiqueter des verbes et des adjectifs et de couvrir plus de classes sémantiques.

12. http://www.racai.ro/EUROLAN-2003/html/workshop/cfp-Onto-IE.pdf

Le F-score obtenu pour l'étiquetage des expressions sémantiquement similaires aux termes du domaine est 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.

Toutefois, quelques améliorations peuvent être apportées en particulier à notre modèle statistique. La première est l'utilisation de l'approche Lesk (Lesk, 1986) pour la désambiguïsation de sens. Cette modification peut être simplement intégrée dans notre module puisque notre algorithme s'inspire de cette approche 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.

Bien que l'approche présentée soit conçue pour les textes oraux, celle-ci présente des avantages intéressants pour l'EI à partir de textes structurés. En effet, les variations langagières sont présentes aussi bien dans les textes conversationnels que les textes écrits. L'utilisation d'une approche d'apprentissage de patrons basée sur une généralisation de l'utilisation de classes sémantiques en incluant, d'une part les verbes et adjectifs, et d'autre part, en élargissant l'éventail des classes sémantiques considérées pour l'étiquetage sémantique est un moyen de contourner les effets pervers des variations langagières sur le processus d'apprentissage.

Remerciements

Nous remercions Robert Parks pour nous avoir donné accès à la version électronique de Wordsmyth ainsi que le Secrétariat National de la Recherche et Sauvetage pour les manuels de SAR.

9. Bibliographie

Abney S., « Partial Parsing via Finite-State Cascades », *Natural Language Engineering*, vol. 2, n° 4, p. 337-344, 1996.

Biber D., *Variation Across Speech and Writing*, Cambridge University Press, 1988.

Boufaden N., Lapalme G., « Apprentissage de relations prédicat-argument pour l'extraction d'information à partir de textes conversationnels », *Actes de la conférence annuelle sur le traitement automatique des langues naturelles (TALN 2005)*, vol. I, Dourdan, France, juin, 2005.

Boufaden N., Lapalme G., Bengio Y., « Topic Segmentation : A First Stage to Dialog-based Information Extraction », *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, p. 273-280, 2001.

Boufaden N., Lapalme G., Bengio Y., « Découpage thématique des conversations : un outil d'aide à l'extraction », *Actes de la 9^{e} conférence annuelle sur le traitement automatique des langues naturelles (TALN 2002)*, vol. I, Nancy, France, p. 377-382, juin, 2002.

Brill E., « A Simple Rule-based Part-of-Speech Tagger », *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italie, 1992.

Dempster A. P., Laird N. M., Rubin D. B., « Maximum Likelihood from Incomplete Data via the EM », *Journal of the Royal Statistical Society, Series B*, vol. 39, p. 1-38, 1977.

et Océans Canada : Garde côtière P., *SAR Seamanship Reference Manual*, Canadian Goverment Publishing – Public Works and Goverment Services Canada, Ottawa, Canada, 2000.

Graeme H., « Ontology and the lexicon », *in* , S. Staab, , R. Studer (eds), *Handbook on Ontologies*, International Handbooks on Information Systems, Springer, Berlin, chapter 11, p. 209-229, 2004.

Hinton G. E., « Training Products of Experts by Minimizing Contrastive Divergence », *Neural Computation*, vol. 14, n° 8, p. 1771-1800, 2002.

Ide N., Véronis J., « Introduction to the Special Issue on Word Sense Disambiguation : The State of the Art », *Computational Linguistics*, vol. 24, n° 1, p. 1-40, 1998.

Lesk M., « Automatic Sense Disambiguation Using Machine Readable Dictionaries : How to Tell a Pine Cone from an Ice Cream Cone », *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC 1986)*, Toronto, Canada, p. 24-26, 1986.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K., Five Papers on WordNet, Technical Report n° CSL Report 43, Cognitive Science Laboratory, Princeton University, juillet, 1990.

Noy N. F., McGuinness D. L., Ontology Development 101 : A Guide to Creating Your First Ontology, Technical Report n° KSL-01-05 et SMI-2001-0880, Stanford University, mars, 2001.

Press W. H., Teukolsky S. A., Vetterling W. T., *Numerical Recipes in C*, Cambridge Press University, chapter 9, 1988.

Quinlan J. R., « Induction of Decision Trees », *Machine Learning*, vol. 1, p. 81-106, 1986.

Sundheim B., « Design of the MUC-6 Evaluation », *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Morgan Kaufmann Publishers, p. 1-12, 1995.