*TALN 2004, Fès, 19–21 avril 2004*

# Approche statistique pour le repérage de mots informatifs dans les textes oraux

Narjès Boufaden (1), Yoshua Bengio (2), Guy Lapalme (1)
(1) Laboratoire RALI - Université de Montréal
Québec, Canada
boufaden@iro.umontreal.ca, lapalme@iro.umontreal.ca
(2) Laboratoire LISA - Université de Montréal
Québec, Canada
bengioy@iro.umontreal.ca

## Résumé - Abstract

Nous présentons les résultats de l'approche statistique que nous avons développée pour le repérage de mots informatifs à partir de textes oraux. Ce travail fait partie d'un projet lancé par le département de la défense canadienne pour le développement d'un système d'extraction d'information dans le domaine de la Recherche et Sauvetage maritime (SAR). Il s'agit de trouver et annoter les mots pertinents avec des étiquettes sémantiques qui sont les concepts d'une ontologie du domaine (SAR). Notre méthode combine deux types d'information : les vecteurs de similarité générés grâce à l'ontologie du domaine et le dictionnaire-thésaurus Wordsmyth ; le contexte d'énonciation représenté par le thème. L'évaluation est effectuée en comparant la sortie du système avec les réponses de formulaires d'extraction d'information prédéfinis. Les résultats obtenus sur les textes oraux sont comparables à ceux obtenus dans le cadre de MUC7 pour des textes écrits .

We present results of a statistical method we developped for the detection of informative words from manually transcribed conversations. This work is part of an ongoing project for 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 similarity score vectors and topical information. Similarity vectors are generated using a SAR ontology and the Wordsmyth dictionary-thesaurus. Evaluation is carried out by comparing the output of the system with key answers of predefined extraction templates. Results on speech transcriptions are comparable to those on written texts in MUC7.

## Mots-clefs – Keywords

Étiquetage sémantique, extraction d'information
Semantic tagging, information extraction

# 1 Introduction

Le repérage de mots informatifs consiste à détecter des mots qui apportent de l'information pertinente[1] relativement à un domaine particulier. Cette tâche est une étape charnière pour beaucoup d'applications du Traitement Automatique de la Langue (TAL) telles que l'extraction d'information et la génération automatique de résumé. Dans cet article nous étudions le repérage de mots informatifs pour l'extraction d'information (EI) à partir de textes oraux.

L'extraction d'information (EI) a pour but la collecte d'informations pertinentes dans un domaine d'application particulier. Les approches d'EI pour les textes écrits se basent en général sur le contexte immédiat (partie de phrase) des mots informatifs pour les détecter (Appelt *et al.*, 1993; Aberdeen *et al.*, 1996). Les approches symboliques utilisant des patrons d'extraction ainsi que les approches d'apprentissage basées sur les HMM (Leek, 1997; McCallum *et al.*, 2000) ou les règles d'induction (Riloff, 1998; Soderland *et al.*, 1995) sont des exemples classiques utilisant le contexte immédiat. Toutes ces approches reposent sur l'hypothèse de la grammaticalité des textes et de ce fait sont inadéquates pour les textes oraux.

L'approche que nous présentons diffère des approches classiques d'EI conçues pour les textes écrits notamment par sa robustesse aux extra-grammaticalités présentent dans les textes oraux. Nous utilisons le contenu du mot et le contexte d'énonciation représenté par le thème de l'énoncé pour repérer les mots informatifs. les mots potentiellement pertinents sont identifiés grâce à leur contenu (par opposition au contexte syntaxique définit par une partie de phrase). Cela contourne les problèmes des irrégularités grammaticales causées par les répétitions ou omissions, par exemple, très présentes dans les textes oraux. De plus, le thème que l'on associe à un énoncé définit un contexte de nature sémantique moins vulnérable aux extra-grammaticalités et permet de sélectionner les mots informatifs parmi ceux qui sont potentiellement pertinents. De fait, le thème joue un rôle de désambiguisation.

Dans ce qui suit nous décrivons d'abord le corpus utilisé pour ce projet (section 2), puis, les différentes parties du système d'EI (section 3). Ensuite, nous explicitons le modèle utilisé pour le repérage de mots informatifs (section 4) et présentons les résultats de nos expériences (section 5). Enfin, nous comparons nos résultats à ceux de travaux existants (section 6).

# 2 Cadre de projet et description du corpus

Ce travail fait partie d'un projet qui a pour but d'implémenter un système d'EI pour repérer des informations ayant un lien avec les missions de recherche et sauvetage maritimes (domaine SAR) tels que *la nature de l'incident*, *l'endroit de l'incident*, *les ressources allouées pour la recherche* et *les conditions météorologiques pendant la mission de recherche*. Le projet a été mené par le Centre de Recherche de la Défense Valcartier (CRDV) afin de développer un outil d'aide à la génération de plan de SAR à partir de conversations téléphoniques manuellement transcrites.

Le corpus est une collection de 95 conversations téléphoniques transcrites manuellement (39.000 mots). Dans la plupart des cas ce sont des conversations impliquant deux locuteurs (l'appelant Caller et un opérateur Operator) qui discutent des conditions et circonstances entourant un in-

[1] Dans la suite de l'article, nous utilisons par abus de langage le mot 'pertinent' pour signifier 'pertinent par rapport au domaine de la Recherche et Sauvetage'.

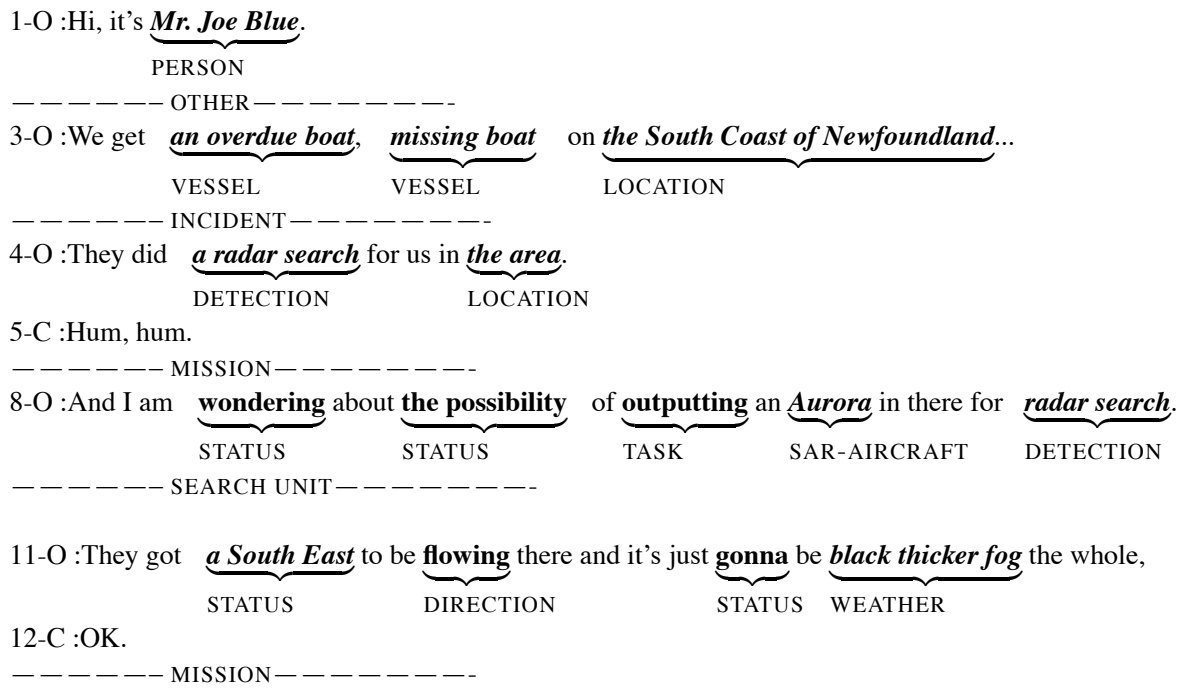


FIG. 1 – Exemple d'une conversation indiquant un incident :***an overdue boat***, une requête pour allouer des avions SAR pour la recherche : ***Aurora***. Les mots en gras italiques sont reconnus par l'étiqueteur sémantique (section 3.3). Tandis que les mots en gras uniquement sont des candidats pour le repérages de mots informatifs (section 4) . Les étiquettes sous les groupes de mots en gras sont des concepts de l'ontologie. Les lignes horizontales sont les frontières des thèmes (MISSION, INCIDENT, SEARCH UNIT, OTHER) ajoutés manuellement.

cident ou une mission de recherche et sauvetage. Les conversations sont : (1) des rapports d'incidents survenus tels qu'une personne portée disparue ou un bateau en retard, (2) l'élaboration d'un plan de sauvetage tels que l'allocation d'avions et de bateaux pour les besoins de la recherche, (3) le compte rendu d'une mission de sauvetage et les résultats de cette mission ou une combinaison des ces trois cas. La Figure 1 donne un extrait de ces conversations.

Le corpus est particulièrement bruité et certaines parties d'énoncés sont remplacées par le mot "INAUDIBLE" pour indiquer que l'enregistrement est incompréhensible. Plus de la moitié des énoncés contiennent au moins une extra-grammaticalité (Shriberg, 1994) telles que les répétitions (Ha, do, is there, is there ...) , les omissions et interruptions (we've been, _ actually had a ...). Enfin, nous avons comptabilisé 3% d'erreurs de transcriptions qui apparaissent en majorité dans les mots informatifs comme c'est le cas dans l'énoncé 11-O où le mot `flowing` devrait être `blowing` (Figure 1).

# 3 Architecture du système d'extraction d'information

L'extraction d'information s'élabore en quatre étapes. L'étape I est l'analyse syntaxique et la détection des groupes de mots candidats à l'extraction. Ce sont essentiellement les groupes nominaux, verbaux, adverbiaux et adjectivaux. L'étape II, l'étiquetage sémantique, annote les

groupes de mots avec les concepts qu'il reconnaît grâce à l'ontologie du domaine que nous avons construie. L'étape III, permet le repérage et l'étiquetage sémantique de groupes de mots informatifs qui ne font pas partie de l'ontologie du domaine et par conséquent qui n'ont pu être étiquetés à l'étape précédente. Enfin, les groupes de mots extraits sont utilisés dans le processus de résolution de coréférence pour, ensuite, remplir les formulaires d'extraction. La Figure 2 illustre l'architecture du système d'extraction d'information.

Dans la prochaine section, nous décrivons de manière concise les trois premières étapes, la conception de l'ontologie du domaine SAR et la segmentation en thèmes. (Boufaden, 2003; Boufaden *et al.*, 2002; Boufaden *et al.*, 2001) présentent une description détaillée de ces modules. La résolution de coréférence et le remplissage de formulaires d'extraction sont laissés pour des travaux futurs. Le repérage des groupes de mots informatifs qui fait l'objet de cet article est détaillé dans la section 4.

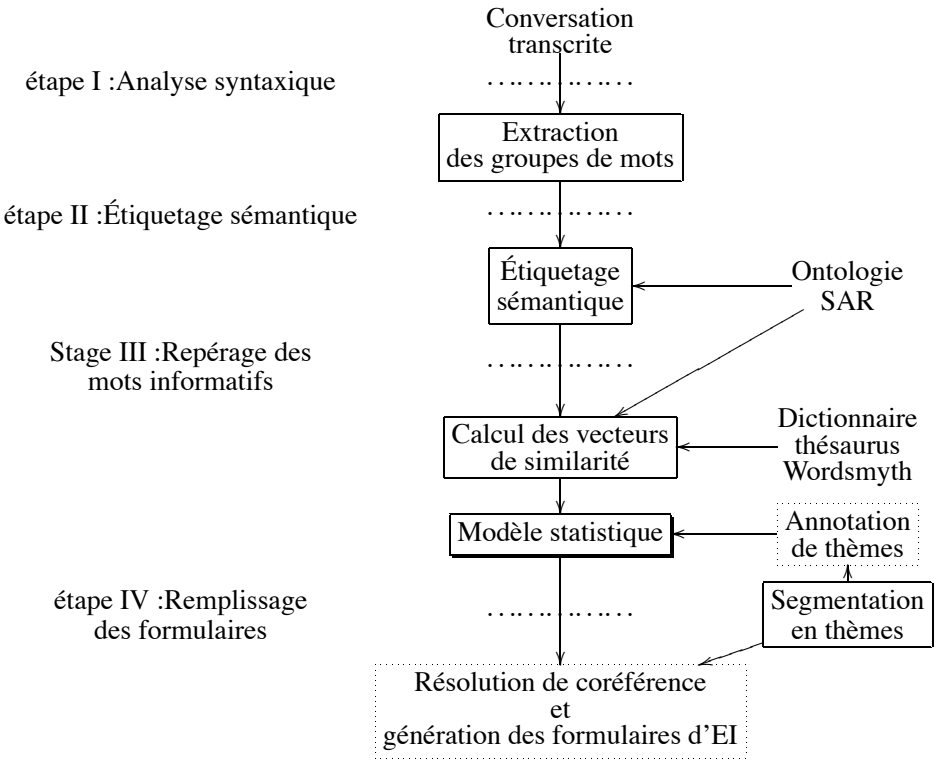


FIG. 2 – Cette figure présente les principales composantes du système d'EI. Les rectangles simples représentent les modules qui ont déjà été développés et sont décrits brièvement dans les sections 3.1, 3.2 et 3.3. Le rectangle en gras est le module qui fait l'objet de cet article. Les rectangles en pointillés sont des modules laissés pour des travaux futurs.

## 3.1 Segmentation en thèmes

Le mot *thème* utilisé dans plusieurs travaux (Carbonell *et al.*, 1999; Hearst, 1994) ne jouit pas d'une définition formelle. Selon l'application cible, le *thème* peut varier du sujet d'un texte au propos d'une partie d'un texte. (Hearst, 1994; Brown & Beorge, 1983) s'accordent pour dire que la notion de *thème* dans un contexte de segmentation de textes implique que les phrases

sont regroupées naturellement selon leur 'propos'[2]. Dans le cadre de notre application, nous avons développé un module de segmentation en thèmes qui permet de regrouper les énoncés adjacents qui portent sur un aspect de la mission, tels que l'annonce d'un incident (énoncé 3-O Figure 1) ou le résultat d'une mission de recherche et sauvetage. Dans (Boufaden *et al.*, 2001; Boufaden *et al.*, 2002), nous montrons qu'en utilisant des connaissances pragmatiques, sémantiques, syntaxiques et lexicales, il est possible moyennant un modèle de Markov de générer les changements de thèmes[3] avec un rappel de 61,4% et une précision de 67,3%.

## 3.2 Ontologie du domaine

L'ontologie du domaine est une composante fondamentale dans notre approche de repérage des mots informatifs. Elle est utilisée lors de l'étiquetage sémantique (section 3.3) et pour la génération des vecteurs de similarité (Boufaden, 2003). L'ontologie du domaine est utilisée pour quantifier la pertinence d'un mot par rapport au domaine. Dans la section 4.2, nous montrons que la probabilité qu'un mot soit informatif est une fonction du degré de similarité de ce mot par rapport aux concepts du domaine SAR.

L'ontologie a été construite à partir de manuels fournis par le Secrétariat de la Recherche et Sauvetage Nationale et d'un échantillon de 10 conversations choisies au hasard. Elle est constituée de mots ou groupes de mots informatifs tels que `radar search`, `diving` pour les moyens de détections, `drifting`, `overdue` pour les incidents et `wind`, `rain`, `fog` pour les conditions météorologiques. Ces mots sont des exemples de réponses pour les champs des formulaires d'EI. Ils sont regroupés en 24 classes et organisées en une hiérarchie IS-A et une autre PART-OF. Les classes de l'ontologie forment les concepts pertinents du domaine SAR. Ils sont utilisés pour étiqueter les mots informatifs comme nous l'expliquons dans la section 4. Enfin, chaque entrée de l'ontologie contient un mot informatif, une liste exhaustive de synonymes extraite de Wordsmyth[4] et leur définitions textuelles aussi extraites du dictionnaire-thésaurus. La Figure 3 est un exemple des entrées de Wordsmyth que nous avons utilisé pour la construction de l'ontologie.

## 3.3 Étiquetage sémantique

L'étiqueteur sémantique est similaire à un module d'extraction d'entités nommées (MUC, 1998). Il reconnaît des entités nommées telles que les lieux, les personnes, les organisations, 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 deux étapes illustrées dans la Figure 4. La première étape recherche un appariement entre la tête du syntagme analysé et les instances des concepts de l'ontologie. Lorsque un appariement réussit, le tête est annotée par le concept dont le mot est une instance. La deuxième étape sert à propager l'étiquette sémantique de la tête du syntagme vers tout le syntagme. La sortie de l'étiqueteur sémantique est représentée par les mots en gras italique dans la Figure 1. Dans (Boufaden, 2003), nous montrons que l'étiqueteur sémantique attribue

[2]Ce terme est la traduction de 'aboutness' selon le glossaire français-anglais de terminologie linguistique SIL http ://www.sil.org/linguistics/

[3]Les changements de thèmes sont représentés par une étiquette. Quatre autres sont utilisées pour distinguer les énonces qui font partie d'un segment de thème de celles qui clos le segment de thème, qui initient une conversation ou qui indiquent la fin d'une conversation

[4]URL http ://www.wordsmyth.net/.

```
ENT:     wonder
SYL:     won-der
PRO:     wuhn dEr
POS:     intransitive verb
INF:     wondered, wondering, wonders
DEF:     1. to experience a sensation of admiration or amazement (of-
ten fol. by at):
EXA:     She wondered at his bravery in combat.
SYN:     marvel
SIM:     gape, stare, gawk
DEF:     2. to be curious or skeptical about something:
EXA:     I wonder about his truthfulness.
SYN:     speculate (1)
SIM:     deliberate, ponder, think, reflect, puzzle, conjecture
...
```

FIG. 3 – Description d'une entrée du dictionnaire-thésaurus Wordsmyth pour le verbe **wonder** qui est un verbe généralement utilisé pour formuler une requête pour allouer du matériel de recherche. Ce verbe a pour étiquette le concept STATUS (`8-O` Figure 1). Les acronymes ENT, SYL, PRO, POS, INF, DEF, EXA, SYN, SIM sont respectivement l'entrée, la syllabe, la prononciation, la catégorie syntaxique, les formes fléchies, la définition textuelle, un exemple, les mots synonymes et les mots similaires. Pour construire notre ontologie nous avons utilisé les informations contenues dans les champs ENT, DEF, SYN et SIM.

les concepts avec un rappel de 85,3% et une précision de 94,8%.

Étape 1 : ...SN : black thicker $\underbrace{\textbf{\textit{fog}}}_{\text{WEATHER-TYPE}}$ ...
← `Propagation`

Étape 2 : ...SN : $\underbrace{\textbf{\textit{black thicker fog}}}_{\text{WEATHER-TYPE}}$ ...

FIG. 4 – Le syntagme nominal SN : ***black thicker fog*** est étiqueté avec le concept WEATHER (énonce 11-O). 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.

# 4 Repérage des mots informatifs

Le repérage de mots informatifs est une fonction du contexte d'énonciation représenté par le thème $T$ et de la pertinence du mot par rapport aux concepts $C_k$ de l'ontologie. Pour calculer la pertinence, nous utilisons le contenu du mot $w$ représenté par sa définition textuelle $D_w$, la mesure de similarité OC[5] (Manning & Schutze, 2001) (section 4.2) et l'ontologie. Chaque mot est représenté par un vecteur de similarité qui contient les scores de similarité $sim(w, C_k)$ obtenus pour chaque concept $C_k$ de l'ontologie. Le thème permet d'écarter les faux positifs qui sont les mots $w$ tel que $w$ proche (selon la mesure de similarité ) d'un concept $C_k$ et $C_k$

[5]Overlap Coefficient.

est rarement observé dans le contexte d'énonciation ayant pour thème $T$. Dans ce qui suit, nous présentons le modèle de repérage. La section 4.2 décrit la modélisation de la distribution des concepts sachant les mots. Ce modèle permet de calculer $P(C_k|w_t)$ à partir des scores de similarité $sim(w_t, C_k)$ (Équation 2). La section 4.3 explicite la modélisation de la distribution des concepts $C_k$ étant donné les thèmes $T_t$.

## 4.1 Modèle

Une formulation de notre problématique est : chercher le concept $C_k$ qui maximise la similarité pour un mot $w_t$ ***et*** $C_k$ fréquemment observé étant donné le thème $T_t$. Cela se traduit par un produit d'experts. Un expert, $P(C_t|w_t)$, modélise la distribution des concepts sachant le mot ; l'autre, $P(C_t|T_t)$, modélise la distribution des concepts sachant le thème. Les coefficients $\beta_1$ et $\beta_2$ sont des poids qui reflètent la contribution de chacun des experts dans le modèle de repérage.

Lorsque les deux experts $P(C_t|T_t)$ et $P(C_t|w_t)$ s'entendent sur le concept qui maximise ces deux probabilités, il est facile de conclure que $w_t$ est informatif. Le cas non trivial est lorsque les experts ne s'entendent pas sur le concept. Dans ce cas la décision repose sur un seuil de confiance $\delta$ déterminé de manière empirique (Équation 1). Un mot $w_t$ est considéré informatif lorsque $P(C^*|w_t, T_t)$ est supérieur à $\delta$. L'équation 1 décrit le modèle $P(C^*|w_t, T_t)$.

$$P(C_t = k|w_t, T_t) = \frac{P(C_t = k|w_t)^{\beta_1} P(C_t = k|T_t)^{\beta_2}}{\Sigma_{l=1}^{K} P(C_t = l|w_t^{\beta_1} P(C_t = l|T_t)^{\beta_2}} \quad (1)$$

$$et$$

$$C^* = \underset{C_t}{\operatorname{argmax}}\, P(C_t|w_t, T_t), P(C^*|w_t, T_t) > \delta$$

$k$ est un des $K$ concepts de l'ontologie, $\log P(C_t = k|w_t)$ représente la log probabilité d'observer le concept $k$ étant donné le mot $w_t$ et $\log P(C_t = k|T_t)$ est la log probabilité d'observer le concept $k$ étant donné le thème $T_t$.

## 4.2 Distribution des concepts par rapport à un mot

La pertinence d'un mot est quantifiée en utilisant la mesure de similarité OC. Cette mesure correspond à la proportion de mots en commun, contenus dans la définition textuelle du mot et celle d'un concept (Équation 2). Un mot est jugé proche du domaine lorsqu'un des scores de similarité est élevé pour un concept donné. (Boufaden, 2003) présente l'algorithme qui permet de calculer les scores de similarités et de générer les vecteurs de similarité.

$$sim(w(l), C_k) = \frac{\mid D_{w(l)} \mid \cap \mid D_{C_k} \mid}{min(\mid D_{w(l)} \mid, \mid D_{C_k} \mid)} \quad (2)$$

$w(l)$ représente un sens particulier $l$ du mot $w$ et $C_k$ un concept de l'ontologie du domaine. $D_{w(l)}$ et $D_{C_k}$ sont respectivement les ensembles de mots lemmatisés extraits des définitions textuelles de $w(l)$ et $C_k$. Les définitions textuelles sont extraites à partir de Wordsmyth.

La distribution d'un concept par rapport à un mot $P(C_k|w_t)$ s'exprime en fonction de la probabilité d'observer un concept $C_k$ étant donné un sens particulier $w(l)$ du mot $w$ et la probabilité d'observer un sens particulier $w(l)$ sachant $w$. $P(C_k|w(l))$ est obtenue par une redistribution

des scores du vecteur de similarité afin d'attribuer une probabilité très faible aux scores nuls. Aussi, pour simplifier nos calculs nous supposons que les sens d'un mot sont équiprobables (Équation 3).

$$P(C_k|w) = \sum_{w(l)\in S(w)} P(C_k|w(l))P(w(l)|w), \tag{3}$$

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

Avec $w(l) \in S(w)$ sont les différents sens du mot $w$, $P(C_k|w(l))$ est le score de similarité normalisé entre le concept $C_k$ étant donné un sens $w(l)$ de $w$ et $P(w(l)|w)$ est la probabilité d'observer le sens $w(l)$ étant donné le mot $w$.

### 4.3 Distribution des concepts par rapport à un thème

Selon le découpage effectué à l'étape de segmentation (section 3.1), un segment est composé d'énoncés dont le thème peut être classé en cinq catégories :(1)MISSING_OBJECT qui englobe toutes les informations faisant référence à l'objet impliqué dans un incident ; (2)INCIDENT qui décrit l'incident, sa cause et l'endroit où il s'est produit ; (3)SEARCH_UNIT qui rapporte les faits et actes des équipes de recherches ; (4)MISSION qui décrit les conditions météorologiques lors de la mission, l'endroit où sont effectuées les recherches ; (5)OTHER qui contient toutes autres informations qui n'a pas de lien directe avec le type d'information recherchée (section 2). La probabilité $P(C_t|T_t)$ est définie par l'équation :

$$P(C_t|T_t) = \alpha P_0(C_t) + (1-\alpha)P_1(C_t|T_t) \tag{4}$$

$C_T$ est la séquence des concepts observés, $T_T$ la séquence des thèmes observés. $\alpha$ est le paramètre libre de notre modèle. $P_0(C_t)$ est la fréquence relative des concepts dans le corpus d'entraînement et $P_1(C_t|T_t)$ la fréquence relative des concepts sachant le thème.

## 5 Expériences et résultats

Le corpus d'entraînement est constitué de 1850 mots, soit 65% des 64 conversations annotées manuellement avec les concepts de l'ontologie et les thèmes. Les résultats sont obtenus en comparant les concepts générés par le modèle de repérage aux réponses des formulaires préalablement annotées avec les concepts de l'ontologie. Le Tableau 1 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. Afin de comparer le modèle basé sur la similarité et le modèle exponentiel nous avons considéré uniquement les $P(C_t|w_t) > 0.02$. Pour des rappels équivalents ($38,5\%$ pour $P(C_t|w_t)$ et $36,8\%$ pour $P(C_t|w_t, T_t)$) le modèle exponentiel performe mieux que la modèle basé uniquement sur la similarité. Bien que le modèle basé sur les thèmes ait une faible performance, celui-ci a permis d'augmenter la précision du repérage de mots informatifs de 16,2 %. La moyenne performance du modèle basé sur la similarité est probablement due à l'approximation faite pour passer du vecteur de scores de similarité vers $P(C_t|w_t)$. 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. Le résultat modeste du modèle de repérage de mots informatifs est en partie dû aux erreurs d'étiquetage syntaxique causées par les extra-grammaticalités qui engendrent un score de similarité erroné. Par ailleurs, à cause

| | $P(C_t\|T_t)$ | | $P(C_t\|w_t)$ | | $P(C_t\|T_t, w_t)$ | |
|---|---|---|---|---|---|---|
| Mots | Précision | Rappel | Précision | Rappel | Précision | Rappel |
| Tous | 37,4% | 44% | 73,33% | 76,1% | 61,45% | 55% |
| Informatifs | 34,6% | 21,8% | 64,7% | 38,5% | 75,2% | 36,8% |

TAB. 1 – Classification des mots $w_t$ par rapports aux 24 concepts $C_t$ de l'ontologie. Les mots informatifs sont les réponses des champs des formulaires d'extraction. Les résultats du modèle $P(C_t|w_t)$ sont obtenus en ne considérant que les probabilités $P(C_t|w_t) > 0.02$. Le résultat du modèle $P(C_t|T_t, w_t)$ est obtenu pour le seuil de confiance optimale $\delta = 0.35$

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 (le concept STATUS à lui seul représente 29,5% du corpus d'entraînement) dans le 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.

# 6 Conclusion

Le repérage de mots informatifs est une tâche fondamentale pour plusieurs applications de TAL tels que l'extraction d'information. La plupart des approches élaborées pour les textes écrits se basent sur l'utilisation de patrons formulés par des règles ou par des modèles stochastiques (Leek, 1997; Riloff, 1998). Dans les deux cas, la structure des phrases pertinentes constitue une partie importante des connaissances a priori prises en compte dans la définition de la démarche d'extraction. Ces approches performantes pour les textes écrits sont inadéquates pour les textes oraux qui ne présentent pas de structures phrastiques régulières. Pour pallier ce problème, nous avons développé une approche basée sur le contenu des mots et sur le thème associé au contexte d'énonciation.

Afin d'évaluer la performance du système, nous avons comparé nos résultats avec ceux obtenus lors de MUC7 pour la tâche d'extraction des objets[6] 'Template Object' à partir de textes écrits (MUC, 1998). Le F-score[7] obtenu pour les mots informatifs est de 74.65% alors que le meilleur résultat obtenu lors de MUC7 est de 80%.

Enfin, bien que l'approche présentée soit conçue pour les textes oraux, celle-ci présente des avantages intéressant pour les textes écrits. En particulier, une approche basée sur le contenu des mots permet un raisonnement sur le sens des mots plutôt que sur le mot en tant qu'unité lexicale. Une telle approche est un atout pour remédier aux variations langagières très présentes dans les textes écrits. De plus, l'absence d'extra-grammaticalités dans les textes écrits permet d'envisager un meilleur résultat pour les textes écrits.

[6]Cela correspond à l'extraction des objets qui participent aux évènements. Dans notre cas les évènements sont les incidents

[7]Le F-score utilisé pour MUC7 est $F = \frac{(\beta+1)P.R}{\beta^2.P+R}$ et $\beta = 0.5$

## 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.

## Références

MUC(1998). *Proceedings of the seventh Message Understanding Conference (MUC-7)*. Morgan Kauffman.

ABERDEEN J., BURGER J., DAY D., HIRSCHMAN L., PALMER D., ROBINSON P. & VILAIN M. (1996). MITRE :Description of the Alembic System as Used in MET. In *Proceedings of the TIPSTER 24-Months Workshop*.

APPELT E., HOBBS J., BEAR J., ISRAEL D. & TYSON M. (1993). FASTUS : A Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of IJCAI*, p. 1172–1178.

BOUFADEN N. (2003). An Ontology-based Semantic Tagger for IE System. In *41st. Annual Meeting of the Association for Computational Linguistics(ACL) : Student Workshop*, p. 7–14, Sapporo, Japon.

BOUFADEN N., LAPALME G. & BENGIO Y. (2001). Topic Segmentation : A First Stage to Dialog-based Information Extraction. In *Natural Language Processing Rim Symposium, NLPRS'01*, p. 273–280.

BOUFADEN N., LAPALME G. & BENGIO Y. (2002). Découpage thématique des conversations : un outil d'aide à l'extraction. In *Actes de Traitement Automatique de la Langue*, volume I, p. 377–382, Nancy, France.

BROWN G. & BEORGE Y. (1983). *Discourse Analysis*. Cambridge Textbooks in Linguistics Series. Canbridge University Press.

CARBONELL J., YIMMING Y., LAFERTY J., R.D B., PIERCE T. & LIU X. (1999). CMU Report on TDT-2 : Segmentation, Detection and Tracking. In *DARPA Broadcast News Workshop*.

HEARST M. (1994). Multi-paragraph Segmentation of Expository Text. In *32nd. Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 9–16, New Mexico State University, Las Cruces, New Mexico.

LEEK T. (1997). Information Extraction Using Hidden Markov Model. Master's thesis, University of California, San Diego, CA.

MANNING C. D. & SCHUTZE H. (2001). *Foundations of Statistical Natural Language Processing*, chapter Word Sense Disambiguiation, p. 294–303. The MIT Press Cambridge, Massachussets London England.

MCCALLUM A., FREITAG D. & PEREIRA F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Prroceedings of ICML-2000*.

RILOFF E. (1998). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, p. 1044–1049.

SHRIBERG E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley.

SODERLAND S., FISHER D., ASELTINE J. & W. L. (1995). Crystal : Inducing a Conceptual Dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, p. 1314–1319, Menlo Park.