Approche d'extraction d'information pour les dialogues transcripts

Narjès Boufaden, Guy Lapalme, Yoshua Bengio
{boufaden, lapalme, bengioy}@iro.umontreal.ca
Département d'Informatique et Recherche Opérationnelle
Université de Montréal, Québec, Canada

Résumé

Nous présentons une approche d'extraction d'information à partir de dialogues basée sur une utilisation intensive des connaissances a priori. Le but est de fournir une approche qui tient compte des particularités des dialogues à savoir les échanges entre locuteurs et la présence d'extra-grammaticalités.

Mots clés : segmentation en thèmes, analyse des conversations, extraction d'information

1 Introduction

L'extraction d'information (EI) a pour but la collecte d'information pertinente dans un domaine d'application particulier. Les premiers travaux en EI étaient essentiellement centrés sur l'analyse de documents électroniques tels que les dépêches journalistiques, qui sont des textes écrits structurés. Dans ce contexte, les conférences MUC (Message Understanding Conferences) ont joué un rôle déterminant dans la définition des sous-tâches de l'EI et l'avancement des recherches en EI à partir de textes écrits [Appelt93, Hobbs96, Cardie97]. Actuellement, les recherches en EI s'orientent de plus en plus vers les textes oraux comme les bulletins du journal télévisé qui sont des monologues. Les conférences HUB [HUB98] témoignent de l'importance accordée à cette nouvelle branche de l'EI.

Le travail que nous présentons s'inscrit dans le cadre de l'extraction d'information à partir des dialogues. Notre projet consiste à définir une approche d'EI pour concevoir un système d'EI qui prend en entrée des transcriptions manuelles de conversations téléphoniques et génère en sortie les réponses aux champs de formulaires. Ces dialogues sont des comptes rendus de missions de recherche et sauvetage maritimes complétées. Les formulaires sont prédéfinis et contiennent des informations spécifiques sur différents aspects d'une mission de recherche tels que *Quel est l'objet recherché ? Quelles sont les moyens entrepris pour la recherche ?* et *Quelles sont les conditions météorologiques pendant la mission de recherche ?*

Une étape cruciale de l'EI est la localisation des énoncés contenant de l'information pertinente. Cette étape, maîtrisée avec les textes écrits structurés, ne l'est pas

pour les textes oraux. Les conversations présentent plusieurs particularités compliquant l'EI, notamment l'aspect collaboratif des conversations et la présence d'extra-grammaticalités qui sont des reformulations telles que les répétitions, omissions, etc. Ces deux caractéristiques font que (1) les éléments d'une réponse ne se trouvent pas nécessairement dans le même énoncé et (2) il faut pouvoir reconstituer la réponse correcte à partir d'un segment dont la structure grammaticale est altérée par les extra-grammaticalités. Pour palier ces problèmes, nous proposons une approche d'extraction composée de trois grandes étapes. La première étape, la segmentation thématique, définit une unité minimale d'extraction qui garantit d'avoir toutes les parties d'une réponse à un champ de formulaire dans la même unité d'extraction. Cette étape est propre aux textes oraux. La deuxième étape est le processus d'extraction. Elle prend en entrée un segment thématique, en extrait les informations pertinentes et les attribuent aux champs du formulaire adéquat. L'approche utilisée pour l'extraction peut être appliquée aux textes oraux comme aux textes écrits structurés, l'unité d'extraction étant alors la phrase. Notre approche se base sur une utilisation intensive des connaissances a priori qui permet, d'une part, de palier la taille insuffisante du corpus qui rend les approches purement probabilistes inefficaces. D'autre part, cela permet de contourner l'utilisation des méthodes d'extraction basées sur les patrons d'extraction, qui dans le cas des dialogues, sont inappropriées. Enfin, la troisième étape, qui est propre aux traitement des dialogues, permet de résoudre la surgénération des réponses pour un champ de formulaire donné. La surgénération est une conséquence directe de la présence d'extra-grammaticalités telles que les reprises et répétitions.

2 Description du projet

Le travail que nous présentons fait partie d'un projet de Recherche et Sauvetage initialement élaboré par le Centre de Recherche de la Défense Canadienne. Le but est de concevoir un outil d'aide à la décision qui utilise les informations extraites de comptes rendus de missions de recherche et sauvetage maritimes pour générer des plans de sauvetage pour des missions futures. Ces conversations sont des dialogues informatifs dans lesquels il s'agit de communiquer l'état d'avancement d'une mission de recherche et sauvetage. La Figure 1 est un exemple de compte rendu d'une mission que nous utilisons dans notre étude.

Un des modules de ce projet est un système d'EI qui collecte les informations pertinentes à partir des transcriptions manuelles de conversations téléphoniques. Les informations recherchées sont les circonstances des missions de sauvetage, tels que, les endroits où se sont déroulées les recherches, la description des objets ou personnes recherchées, le nombre de personnes recherchées, l'énumération des ressources utilisées (recherche par radar, avion ou garde côtière) et le temps alloué aux différentes étapes de la mission.

À titre d'exemple, les informations ciblées dans l'extrait de la Figure 1 sont :

- le lieu où a été déclaré l'incident : *south east coast Town2, just in the area quite between Town1 and Town3, the south east coast of Town2*.
- le temps alloué à la recherche : *24 hours*
- les ressources allouées : *an Aircraft1* et *radar search*

1	C	*Maritime operation centre, (INAUDIBLE) hello.*
2	O	*Hi, Mr. Green, it's captain Mr. Red*
3	C	*Yes.*
		. . . . .
4	O	*Ha, I don'nt know if I was handled over to you at all, but we've got an overdue boat on the south coast of Town2, just in the area quite between Town1 and Town3.*
5	O	*It's on the south east coast of Town2.*
		. . . . .
6	O	*This is been going on for, for 24 hours that the case has, or almost anyway, and we had an Airplane1 up flying this morning*
		. . . . .
7	O	*They did a radar search for us in that area.*
8	C	*Yes.*
		. . . . .
9	O	*And their search turned up nothing.*
10	C	*yeah.*
		. . . . .
11	C	*Thanks.*
12	O	*All right.*
13	O	*Bye*

FIG. 1 – Extrait d'un compte rendu entre deux locuteurs : `Caller` (`C`) et `Operator` (`O`). Pour des raisons de confidentialité nous avons remplaçé certaines entités nommées par des noms génériques.

– le résultat de l'étape de la recherche par radar : *their search turned up nothing*

Notre étude a été effectuée sur 95 conversations (totalisant environ 39 000 mots).

3 Particularités des dialogues et conséquences sur les stratégies d'EI

Lorsqu'il s'agit de textes écrits, l'unité d'extraction est la phrase. Il n'y a pas d'ambiguïté dans la délimitation de l'unité à considérer dans le processus d'extraction. Cette caractéristique permet de supposer que chaque phrase jugée pertinente véhicule une réponse complète pour un champ de formulaire donné. Pour les textes oraux, cette hypothèse n'est pas nécessairement garantie. D'une part, un énoncé, qui est l'unité de base d'un texte oral n'est pas aussi bien défini que la phrase [Traum97]. De plus, pour les dialogues, l'aspect collaboratif fait que l'information à localiser ne se trouve pas nécessairement dans le même énoncé, mais plutôt sur des énoncés adjacents. Cette particularité des dialogues soulève la question suivante : comment savoir qu'un énoncé contient toutes les parties de la réponse à un champ de formulaire ?

Nous pensons que le découpage thématique permet de répondre à cette question. Ce processus appliqué permet la définition d'une unité d'extraction qui garantit que

chaque segment jugé pertinent véhicule une réponse complète. La section 5.1 décrit nos travaux à ce sujet.

D'autre part, la présence d'extra-grammaticalité telles que les répétitions, les reprises et omissions [Shriberg94] altère la structure grammaticale des énoncés. Ceci rend difficile l'apprentissage des patrons d'extraction laquelle est l'approche standard utilisée pour les textes écrits. Nous pensons qu'une approche sans patrons permet de palier à ce problème et est plus appropriée pour les dialogues.

Enfin, les extra-grammaticalités telles que les répétitions et les reprises engendrent souvent plus d'une réponse à un même champ de formulaire, ce qui soulève la question suivante : comment gérer la surgénération de réponses due aux extra-grammaticalités et comment choisir la bonne réponse ?

Ce problème a été en partie étudié en analyse syntaxique et en général dans plusieurs travaux dont [Boufaden98, Hindle83, Heeman96, Bear92, Levelt83]. L'exemple de la Figure 2 illustre les problématiques soulevées.

1	C	*How long approach (INAUDIBLE).*
2	O	*It'll be, it'll be less than 2 hours.*
3	O	*It probably would be an hour and a half, an hour 45.*
4	C	*(INAUDIBLE) around ha...*
5	O	*5*
6	C	*An hour and a half.*
7	C	*It would be 9, 9 :32.*
8	O	*Ha, 7, 8, about 9 :30, between 9 :30 and 10 :00 Zulu.*
9	C	*OK.*
10	C	*Good enough.*
11	C	*That would be around 10 o'clock.*
12	C	*Good enough.*
13	O	*All right.*

FIG. 2 – Aspect collaboratif d'un dialogue : cas d'un segment thématique qui répond à un champ du formulaire générique DATE.

L'extrait de la Figure 2 est un segment thématique qui illustre l'aspect collaboratif des dialogues. Nous distinguons une discussion pendant laquelle les locuteurs tentent de se mettre d'accord sur l'heure d'arrivée de l'unité de recherche qui est la ressource allouée pour la mission. De cet exemple, il apparaît clairement que sans un découpage thématique, il serait difficile d'identifier l'heure sur laquelle se sont mis d'accord les locuteurs (`That would be around 10 o'clock`). De plus, sans un traitement qui permette de sélectionner la bonne réponse parmi toutes celles identifiées et représentées dans le tableau 1, il est difficile de choisir la bonne réponse pour le champ `temps` du formulaire générique DATE. Le découpage thématique ainsi restreint l'espace de recherche des réponses possibles à un champ tout en s'assurant que toutes les possibilités et parties de la réponse y sont présentes.

Dans cet article, nous mettons l'emphase sur l'approche d'EI.

Expressions temporelles extraites de la Figure 2
less than 2 hours
an hour and a half
an hour 45
5
an hour and a half
9
9 :32
7
8
about 9 :30
between 9 :30 and 10 :00 Zulu
around 10 o'clock

TAB. 1 – Heures possibles en réponse au champ `temps` du formulaire générique DATE

4 EI à partir de conversations téléphoniques transcrites

L'EI est un processus complexe qui met en jeu plusieurs ressources. De manière générale, trois composants sont indispensables :

1. Les connaissances du domaine et les connaissances du monde qui ensemble définissent les connaissances *a priori* (section 4.1).
2. Les formulaires qui définissent le type de l'information recherchée (section 4.2).
3. Les algorithmes d'extraction qui vont permettre la localisation et l'extraction des informations (section 5).

4.1 Connaissances *a priori*

Ce sont des informations spécialisées et propres à un domaine particulier, plus les connaissances du monde modélisées par WordNet [Fellbaum98]. Cela peut être des noms d'organismes gouvernementaux, de pays et de villes. Dans le contexte de notre recherche, les connaissances spécialisées sont le nom des avions et des bateaux utilisés lors des missions de recherche, les organismes auxquels ils appartiennent, les noms de zones de recherche, l'alphabet utilisé pour les codes et les relations qui lient ces différentes entités dans le domaine de la recherche et sauvetage. Cette composante est utilisée pour l'extraction des entités nommées, mais comme nous l'expliquons dans la section 5, elle peut être utilisée tout au long du processus d'extraction afin d'éviter les approches d'EI basées sur les patrons, qui dans le cas des dialogues, s'avèrent inappropriées.

4.2 Les formulaires

Les formulaires regroupent les informations qui définissent les entités, relations et événements sur lesquels nous voudrions mettre l'emphase lors du processus d'EI.

Ils précisent quelles sont les informations recherchées pour chaque entité, relation et événement. Lors des conférences MUC, les différents formulaires étaient fournis. Dans le cas de notre application, ceux-ci n'ont pas été fournis. De ce fait, nous avons construit nos propres formulaires nécessaires à chaque étape de l'extraction à partir des connaissances du domaine. Cette étape de conception consiste à établir un modèle de représentation des données qui prend en compte les entités, les événements et les liens qui les relient, puis faire la projection du modèle en tenant compte à chaque fois du niveau du formulaire considéré tel que défini dans les MUC. Les formulaires que nous avons générés pour le domaine de la recherche et sauvetage sont les suivants :

niveau 1 : Formulaires génériques Ce sont les formulaires (1) AIRCRAFT qui contient un champ pour le type d'avion (hélicoptère, avion), la catégorie de cet avion, sa description ; (2) VESSEL qui contient un champ pour le type de bateau, sa description (Figure 4.2) ; (3) WEATHER-CONDITION qui contient un champ pour les conditions météorologiques telles que la pluie, la neige, la direction du vent, sa vitesse (Figure 4.2) ; (4) DATE qui contient un champ pour la période de la journée, l'heure, la date ; (5) PERSON qui contient le nom et le prénom de la personne, son grade (optionnel), sa description (optionnel) ; (6) LOCATION qui contient un champ pour le nom de la ville, de la région, les coordonnées (latitude, longitude).

niveau 2 : Formulaires des relations Ce sont les formulaires (1) UNIT-SEARCH-OF qui contient un pointeur vers la mission SEARCH-MISSION qui a alloué l'unité de recherche, un pointeur vers la ressource allouée (VESSEL ou AIRCRAFT) et d'autres informations relatives à l'unité de recherche (Figure 4.2) ; (2) MISSED-OBJECT-OF qui contient un pointeur vers le type d'objet recherché (VESSEL ou AIRCRAFT) et l'état de l'objet (`broken`, `under water`, etc).

niveau 3 : Formulaires des scénarios C'est le formulaire SEARCH-MISSION qui contient des pointeurs vers les unités de recherches allouées (UNIT-SEARCH-OF) à cette mission, un pointeur vers l'objet recherché (MISSED-OBJECT-OF), un pointeur vers les conditions météorologiques (WEATHER-CONDITIONS) et d'autres informations relatives à la mission.

niveau 4 : Formulaire principal C'est le formulaire INCIDENT qui contient des pointeurs sur les missions allouées pour la couverture de l'incident, un pointeur sur la date de l'incident (DATE), un pointeur sur la personne (PERSON) chargé de l'incident, ainsi que d'autres informations telles que la cause de l'incident et le type d'alerte.

La figure 4.2 montre les différents formulaires génériques ainsi que les pointeurs vers les champs des formulaires relationnels et événementiels.

5 Extraction des formulaires génériques

L'approche d'EI que nous proposons tient compte de trois contraintes, à savoir :
- l'utilisation de segments thématiques comme unité d'extraction,
- la difficulté à définir des patrons d'extraction pour les raisons décrites dans la section 3,

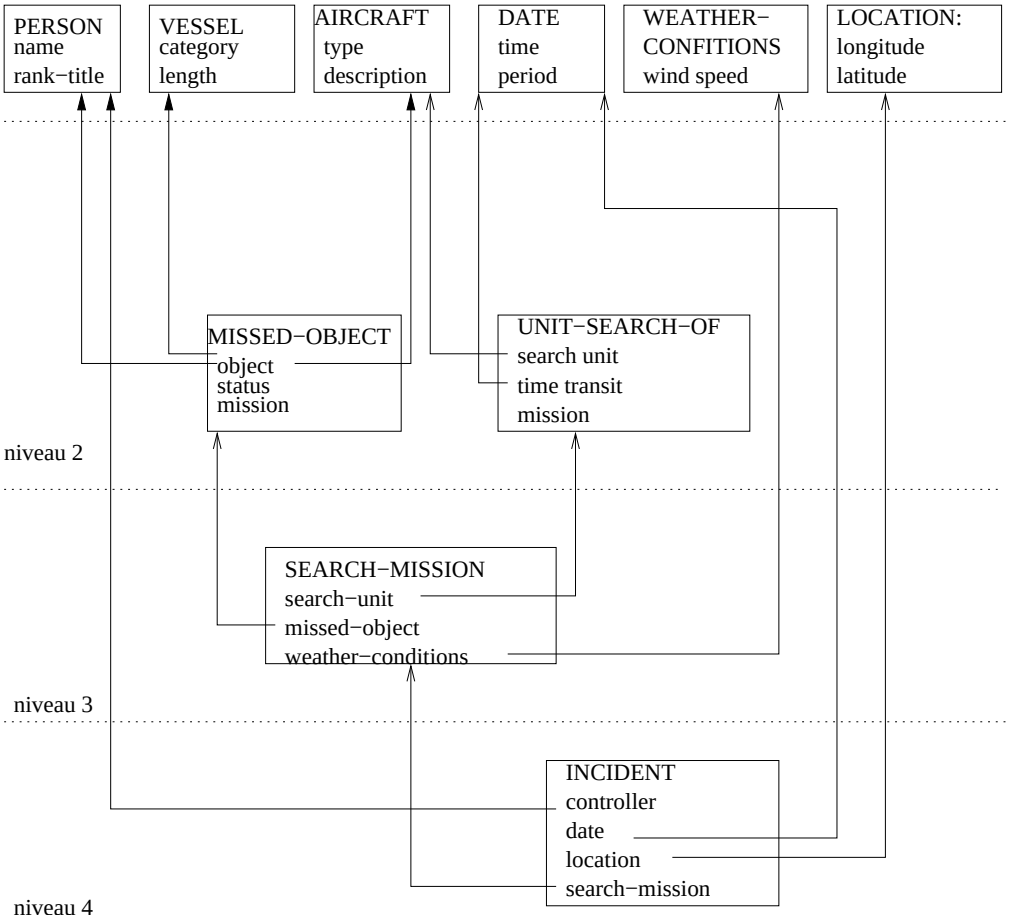


FIG. 3 – Les quatre niveaux de formulaires du projet de Recherche et Sauvetage.

– la taille du corpus d'étude qui contient 39 000 mots et qui est insuffisante pour envisager une approche purement probabiliste.

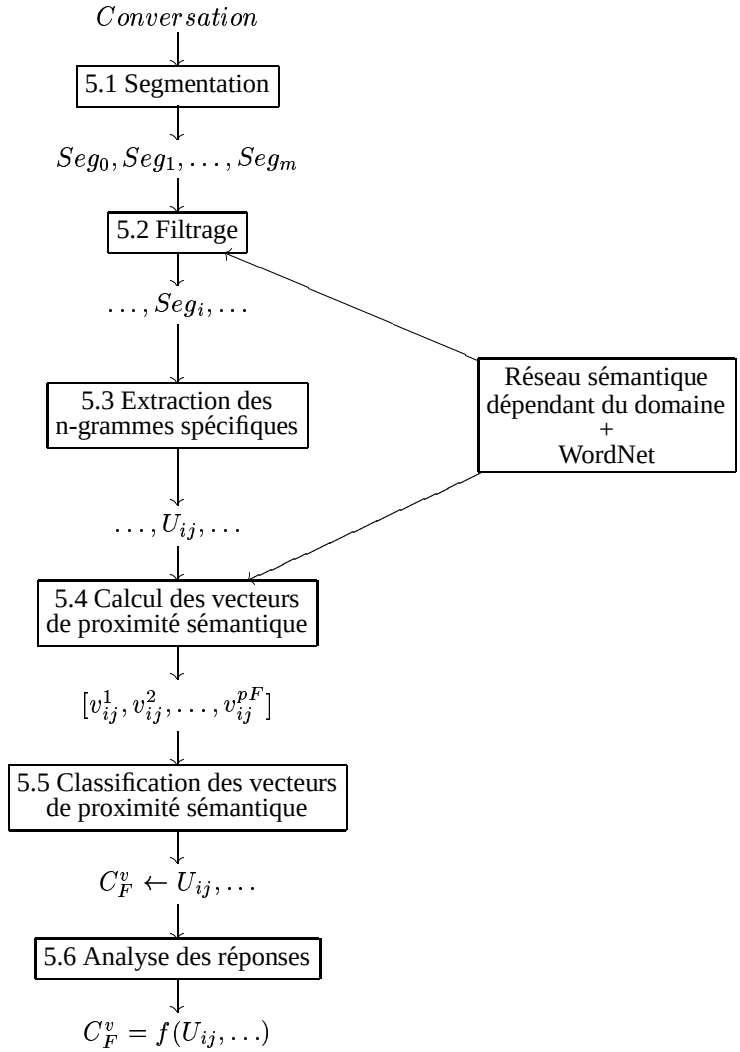
Notre approche est basée sur une utilisation intensive de connaissances du domaine de l'application (ontologie du domaine) ainsi que de WordNet [Fellbaum98]. Nous associons à chaque champ de formulaire une signature qui est constituée d'une liste d'information caractéristiques du champ. Un exemple de signature pour le champ `description` du formulaire générique VESSEL et AIRCRAFT sont les termes **color** et **length**, qui sont des termes récurrents dans la description des avions et des bateaux. Les termes `forecast` ou `weather report` sont aussi caractéristiques du champ `type` du formulaire générique WEATHER-CONDITION (Figure 4.2).

La signature peut aussi contenir le format de la réponse à un champ comme c'est le cas pour le champ `time` du formulaire générique DATE. Ainsi, chaque champ sera identifié par une signature qui sera utilisée lors de la classification des informations contenues dans un segment thématique donné.

Notre méthodologie consiste à localiser des groupements de mots (groupes nominaux, parties de groupes nominaux, chiffres, adverbes temporels, etc.) contenus dans un segment thématique et à ne retenir que ceux qui sont jugés pertinents sur la base de la distance sémantique qui les séparent de notre ontologie. Ensuite, ces groupes de mots pertinents sont classés pour déterminer à quel champ de formulaire ils appartiennent. La classification utilise un vecteur de proximité sémantique qui reflète la distance des mots composant les groupes retenus et les signatures d'un sous ensemble de champs donnés.

L'approche est composée de 6 étapes (figure 4) :

1. Découper la conversation en segments thématiques $S_1, \ldots, S_m$.
2. Filtrer les segments de manière à ne garder que les S_i qui sont pertinents pour l'extraction et choisir le formulaire d'élément générique F le plus proche au thème du segment. En fixant le formulaire générique F concerné par le segment de thème traité, il sera possible de connaître le sous-ensemble de signatures à utiliser lors de l'étape de classification.
3. Extraire les groupes de mots spécifiques U_{ij} de chaque segment S_i qui sont des groupes nominaux, chiffres, formats d'heure ou de date et autres informations pouvant constituer une réponse potentielle à un champ du formulaire F choisi à l'étape de filtrage.
4. Pondérer chaque groupe de mots extrait en lui associant un vecteur de proximité sémantique contenant p_F scores calculés sur la base de la distance sémantique entre les éléments de chaque signature des p_F champs du formulaire F. Pour calculer la distance sémantique, nous utilisons les connaissances a priori (section 4.1) .
5. Déterminer le champ C_F^v pour lequel U_{ij} est une réponse.
6. Si plus d'une unité pertinente U_{ij} est allouée au champ C_F^v, un processus de gestion de la surgénération de réponses sera lancé afin de sélectionner ou de composer la bonne réponse.



- $i \in \{0, \ldots, m\}$ avec m le nombre de segments
de la conversation
- $j \in \{0, \ldots, n_i\}$ avec n_i le nombre
de n-grammes extraits à partir du segment Seg_i
- p_F est le nombre total des champs du formulaire F
- C_F^v est le vi ème champ du formulaire F
- $f(\ldots)$ implique qu'il y a eu traitement des
n-grammes extraits pour
remplir le champ C_F^v du formulaire F

FIG. 4 – Étapes de l'extraction des réponses des formulaires génériques.

5.1 Segmentation thématique

Nous avons effectué une segmentation en sous thèmes [Hearst, 94] qui a permit de regrouper les énoncés adjacents qui portent sur un aspect de la mission. Le but de cette étape est de garantir la présence de toutes les parties d'une réponse à un champ de formulaire (Figure 4.2) dans un même segment thématique. Dans la Figure 1, les segments thématiques sont composés des énoncés situés entre les lignes en pointillé. Les énoncés 4 et 5 fournissent l'endroit présumé du bateau en retard. Ces deux énoncés appartiennent à un même thème, c'est-à-dire qui est le lieu du bateau en retard. Concrètement, un thème est composé d'énoncés adjacents qui indiquent par exemple le type d'incident, comme dans le cas des énoncés `4 et 5` de l'extrait de la Figure 1, l'heure du début de la mission de recherche et sauvetage (énoncé `6`) ou les ressources allouées pour cette mission (énoncés `7 et 8`) et les résultats de la recherche (énoncés `9 et 10`). En procédant à ce découpage, le système d'EI peut localiser les éléments d'information qui forment la réponse à un champ de formulaire.

Une segmentation manuelle de 65 conversations montre que 60 % des thèmes sont composés de 2 à 6 énoncés et plus de 35% des thèmes sont déclenchés par une question dont la réponse s'étend sur plus de deux énoncés (Figure 2). D'autre part, seulement 6% des thèmes sont constitués d'un seul énoncé. La variation dans la distribution des longueurs des thèmes a nécessité le développement d'un modèle de langage pour déterminer la distribution des frontières de thème.

Dans nos travaux [Boufaden2001, Boufaden2002], nous décrivons les expériences et résultats de l'automatisation de la segmentation thématique. À cette fin, nous avons identifié les informations discriminantes pour la localisation des frontières et avons utilisé une approche probabiliste basée sur un modèle de Markov caché d'ordre 1. La détermination des marques discriminantes pour la segmentation repose sur deux aspects : (1) la cohésion entre les énoncés d'un même thème et (2) l'aspect collaboratif des conversations. Notre approche basée sur les traits linguistiques, les interruptions et les catégories d'entités nommées nous a permis d'obtenir un taux de détection de changement de thème de l'ordre de 61,4 % avec une précision de 67,3 % et un rappel de 61,4 %.

5.2 Filtrage et classification des thèmes

Le but de cette étape est d'éliminer les segments de conversation qui n'apportent pas de réponses aux champs des formulaires et de prédire quel formulaire sera rempli par les éléments extraits du segment en cours de traitement. Ce procédé permet de retenir les segments où l'on cite des informations relatives à un élément générique. Le processus de filtrage utilise essentiellement sur les connaissances a priori. La méthode que nous proposons se base sur le calcul d'une distance sémantique entre les noms ou verbes non modaux contenus dans le segment et l'ensemble des concepts définis dans le réseau sémantique dépendant du domaine. Cette distance sera calculée en utilisant les liens de WordNet et sera ensuite traduite en un score qui sera comparé à un seuil de confiance.

5.3 Extraction des unités spécifiques

Cette étape permet d'extraire tous les groupes de mots qui pourraient constituer une réponse potentielle à un champ C_F^v du formulaire F. Ces unités sont, par exemple, des groupes nominaux, des chiffres, des codes ou une information indiquant une date, une condition météorologique. Le tableau 1 montre les unités pertinentes retenues pour l'extrait de la figure 2.

Cette étape sera elle-même divisée en sous-étapes où l'on extraira successivement :

- Les groupes de mots figés qui représentent, par exemple, le nom d'une organisation, le nom d'une ville, le nom d'un avion, les expressions temporelles. Le réseau sémantique dépendant du domaine sera utilisé pour la reconnaissance de ces entités nommées.
- Les unités pertinentes qui sont difficiles à identifier par des patrons, tels que par exemple, `in the area quite between Town1 and Town2`, `overdue 20-foot open boat` ou `10-horse power upboard` et les formes elliptiques telles que `auxiliaries` pour `coast guard auxiliaries`, etc.

5.4 Calcul du vecteur de proximité sémantique

Cette étape permet de remplacer chaque mot d'une unité spécifique U_{ij} extraite par une représentation vectorielle v_{ij} qui nous renseigne sur la pertinence de ce mot. Le vecteur de proximité représentant l'unité spécifique U_{ij} sera une combinaison linéaire des vecteurs de chaque mot qui forme l'unité spécifique. La dimension p_F du vecteur v_{ij} est le nombre de champs du formulaire F prédit à l'étape de filtrage. Chaque élément du vecteur est un score calculé en fonction de la proximité sémantique des éléments qui forment l'unité spécifique par rapport aux signatures des p_F champs du formulaire générique F. Un exemple de signature pour le champ `description` du formulaire VESSEL est composé des termes **length**, **color** et **weight**. Ainsi, le mot `pounds` qui est sémantiquement proche du concept `weight` va permettre l'attribution de l'unité `42000 pounds of GP-4` au champ `description` du formulaire `vessel`. Le même raisonnement s'applique avec l'unité `white, with blue prints` qui contient les mots `white` et `blue` sémantiquement proches du concept clé `color`.

5.5 Classification des vecteurs de proximité sémantique

Le but de cette étape est d'identifier si une unité pertinente U_{ij} constitue une réponse à un champ de formulaire et lequel des champs est le plus approprié. L'approche préconisée prend comme entrée le vecteur de proximité sémantique d'une unité U_{ij} et génère en sortie le champ attribué ou un symbole qui indique que U_{ij} ne constitue pas une réponse à un des champs du formulaire générique F considérés. Dans notre approche, nous comptons utiliser les réseaux de neurones pour classifier les vecteurs de proximité sémantique.

5.6 Analyse des réponses

Cette étape est spécifique à l'extraction d'information à partir de dialogues. Elle va permettre de résoudre la surgénération de réponses (pour un champ) qui est due à la présence d'extra-grammaticalités telles que les reprises (Figure 5), les répétitions et l'aspect collaboratif des conversations (Figure 2) :

- Les répétitions et reprises peuvent créer des versions modifiées d'une même information. Plusieurs unités pertinentes potentielles U_{ij} seront extraites et attribuées à un même champ de formulaire C_F^v. Par exemple, dans l'extrait de la Figure 5, l'opérateur (`O`) formule une première fois l'endroit présumé d'un bateau porté disparu `the south coast of Town2, just in the area quite between Town1 and Town3`, puis reformule une partie de l'énoncé pour apporter des modifications sur l'endroit présumé du bateau, à savoir `the south east coast of Town2`.
- L'aspect collaboratif des conversations engendre un autre type de problème qui lui aussi a pour conséquence la surgénération de réponses potentielles pour un champ, comme le montre le Tableau 1. Tous les temps cités sont des réponses potentielles au champ `time` du formulaire DATE. Parmi toutes ces unités pertinentes, c'est la dernière qui résume le mieux la discussion et qui est approuvée par les deux locuteurs (`around 10 o'clock`) et donc sera considéré comme une réponse pour le champ `time`.

4	O	*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 Town2, just in the area quite between Town1 and Town3.*
5	O	*It's on the south east coast of Town2.*

FIG. 5 – Surgénération de réponses pour le champ `area` du formulaire LOCATION causée par une reprise effectuée par le locuteur (`O`).

Jusqu'à présent, nous avons implémenté et évalué l'étape de segmentation thématique [Boufaden2001, Boufaden2002]. Actuellement, nous procédons à l'annotation des 95 conversations afin d'identifier les réponses aux champs des formulaires génériques. Les réponses des formulaires de 80 conversations seront utilisées pour la partie apprentissage du classifieur et les réponses des formulaires de 15 conversations serviront à l'évaluation de la tâche d'extraction des réponses des formulaires génériques.

6 Conclusion

Nous avons présenté une approche d'EI basée sur une utilisation intensive des connaissances a priori ainsi que sur des réseaux de neurones pour la classification des informations extraites. La plupart des approches élaborées pour l'extraction d'information pour les textes écrits se basent sur l'utilisation de patrons formulés par des règles ou par des modèles stochastiques [Huffman96, Cardie97, Leek97, Riloff96]. 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. Pour les dialogues, ce type d'approche est difficile à appliquer. Les dialogues sont des textes oraux qui présentent deux caractéristiques sources de problèmes pour les traitements standards du langage : la présence d'extra-grammaticalités, qui altèrent la structure syntaxique des énoncés, et la présence d'échanges entre locuteurs, qui constituent l'aspect collaboratif des dialogues. Les extra-grammaticalités rendent l'utilisation de patrons d'extraction difficile sinon impossible. D'autre part, l'aspect collaboratif des dialogues nécessite la prise en compte de la contribution de chacun des locuteurs dans le processus d'extraction pour assurer la complétude des informations extraites.

Plusieurs travaux ont été effectués en segmentation thématique de monologues dans le but de préparer les données à la tâche d'extraction. Les conférences TDT (Topic Detection and Tracking) [Allan98] qui font partie du programme TIDES (Translingual Information Detection, Extraction, and Summarization) de la DARPA étudient la segmentation en thèmes de monologue, tels que les bulletins du journal télévisé, dans le but de faciliter certains traitements de la langue naturelle dont l'EI. Ceci explique notre intérêt à effectuer une segmentation thématique avant l'étape d'EI.

Actuellement, très peu de travaux ont étudié le problème de l'EI à partir de dialogues. Les conférences HUB [HUB98] organisées en partie par la DARPA étudient la tâche d'extraction à partir de monologues tels que les bulletins du journal télévisé, et plus particulièrement l'extraction des entités nommées. À notre connaissance, peu ou pas de travaux se sont spécialisés sur la tâche d'extraction proprement dite à partir de dialogues. Ainsi, cet article a présenté une analyse exploratoire des difficultés et contraintes liées à l'EI à partir de dialogues et a décrit l'approche proposée pour palier ces difficultés.

Références

[HUB98] *Proceedings of the DARPA Broadcast Transcriptions and Understanding Workshop*, Lansdowne, Virginia, février 1998.

[Allan98] J. Allan, J. Carbonnel, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[Appelt93] E.D. Appelt, J.R. Hobbs, J. Bear, D. Israel, and M. Tyson. Fastus : A Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of IJCAI*, pages 1172–1178, 1993.

[Bear92] John Bear, John Dowding, and Elizabeth Shriberg. Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-computer Dialog. In *Proceedings of ACL*, pages 56–63, 1992.

[Boufaden2001] N. Boufaden, G. Lapalme, and Y. Bengio. Topic Segmentation : A First Stage to Dialog-based Information Extraction. In *Natural Language Processing Rim Symposium, NLPRS'01*, pages 273–280, Tokyo, Japan, 2001.

[Boufaden2002] N. Boufaden, G. Lapalme, and Y. Bengio. Segmentation en topiques de conversations téléphoniques : traitement en amont pour l'extraction d'information. In *TAL 2002*, Nancy, France, Juin 2002.

[Boufaden98] Narjès Boufaden. Analyse syntaxique robuste des textes de dialogues oraux. Master's thesis, Université LAVAL, Janvier 1998.

[Cardie97] C Cardie. Empirical Methods in Information Extraction. In *AI Magazine*, volume 18 of *4*, pages 65–79. American Association for Artificial Intelligence, 1997.

[Fellbaum98] Fellbaum Christiane. *WordNet an Electronic Lexical Database*. MIT Press Cambridge Massachusetts, London England, 1998.

[Hearst, 94] Marti Hearst. Multi-paragraph Segmentation of Expository Text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16, New Mexico State University, Las Cruces, New Mexico, 1994.

[Heeman96] P. Heeman, Ho Kyung, Kim Lokem, and J. Allen. Combining the Detection and Correction of Speech Repair. In *Proceedings of the International Conference on Spoken Language Processing*, pages 358–361, Philadephia, October 1996.

[Hindle83] D. Hindle. Deterministic Parsing of Syntactic Nonfluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, 1983.

[Hobbs96] J. Hobbs, D Appelt, J. Bear, D. Israel, and Kame. *FASTUS : A Cascaded Finite-State Transducer for Extracting Information from Natural Language Text*. MIT Press, 1996.

[Huffman96] 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.

[Leek97] T.R. Leek. Information Extraction using Hidden Markov model. Master's thesis, University of California, San Diego, CA, 1997.

[Levelt83] W. J. M Levelt. Monitoring and self-repair in speech. *Cognition*, 1983.

[Riloff96] E. Riloff. Automatically Generating Extraction Patterns from Untagged Tex. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pages 1044–1049, 1998.

[Shriberg94] E.E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley, 1994.

[Traum97] D. Traum and P. Heeman. Utterance Units in Spoken Dialogue. In E. Maier, M. Mast, and S. LuperFoy, editors, *Dialogue Processing in Spoken Language Systems*, LNAI. Springer-Verlag, 1997.