EXIBUM : Un système expérimental d'extraction d'information bilingue*

Leila Kosseim et Guy Lapalme
RALI, Département d'informatique et de recherche opérationnelle
Université de Montréal, CP 6128, Succ. Centre Ville
Montréal (Québec) Canada, H3C 3J7
e-mail : {kosseim, lapalme}@iro.umontreal.ca

Abstract

This paper presents the EXIBUM system : a bilingual information extraction system currently being developed at the University of Montréal. EXIBUM is an experimental system for analyzing French and English terrorist events from newswires and producing a template of the most pertinent information. EXIBUM consists of the following modules : a language identifier, a part-of-speech tagger, a transcriber, a filter, a syntactic analyzer, a semantic analyzer, a discourse analyser, and a formatter. An interesting characteristic of EXIBUM is that many of its components are off-the-shelf systems not initially intended for information extraction. The rapid results obtained through this experiment demonstrate the great advantage of system re-use in this domain, and leave us optimistic for the future development of multilingual information extraction systems.

1 Introduction

L'extraction d'information consiste à extraire de l'information précise du contenu d'un document et à la représenter sous forme structurée. Cette forme structurée peut ensuite être stockée dans une base de données ou être utilisée comme base à la génération automatique de résumés. Par exemple, à partir d'un rapport sur un attentat terroriste, un système d'extraction

[0]blabla

*in Proceedings of *Rencontre Internationale sur l'extraction, le filtrage et le résumé automatiques (RIFRA-98)*. pp. 129-140. November. Sfax, Tunisia.

d'information sera capable d'identifier la date de l'attentat, le type d'attentat, l'auteur du crime ainsi que les victimes.

L'extraction d'information soulève de plus en plus d'intérêt en ingénierie linguistique, car ce domaine répond à un besoin concret et immédiat de l'industrie : le développement de systèmes robustes capables de manipuler rapidement de grandes quantités de documents. En effet, d'énormes quantités d'information "brute" sont maintenant disponibles (sur le Web, sur CD-ROM, ...) et la grande majorité est sous forme de textes en langue naturelle ; des outils permettant d'identifier, de filtrer et de résumer ces textes deviennent donc de plus en plus importants.

Dans cet article, nous présentons les travaux de recherche en extraction d'information effectués à l'Université de Montréal. Le but de notre recherche est de concevoir un système d'extraction d'information bilingue (anglais et français), c'est-à-dire capable d'analyser des documents rédigés aussi bien dans une langue que dans l'autre et d'en extraire de l'information dans la langue originale.

2 L'extraction d'information

Bien que l'extraction d'information soit en pleine effervescence, les ancêtres de ce domaine remontent au tout début de l'intelligence artificielle. En effet, dès 1950-1960 on tentait de créer des systèmes capables de structurer toutes les données d'un texte en langue naturelle.

Après cette période optimiste, les travaux se sont limités à extraire et à structurer de l'information bien précise plutôt que toute l'information retrouvée dans le texte. Dans cette première génération de travaux en extraction d'information, on retrouve, par exemple, le système d'extraction de noms propres et de titres à partir d'extraits de journaux de [Borkowski, 1967], le système FRUMP de [DeJong, 1979] qui extrait de l'information sur des désastres naturels à partir d'extraits de journaux et le système de [Sager, 1981] sur les rapports de radiologie.

C'est à partir de la fin des années 1980, que la recherche en extraction d'information a joui d'une grande popularité. Aux États-Unis, l'armée s'est particulièrement intéressée à l'extraction d'information (d'où l'organisation des *Message Understanding Conferences*) et de nombreux projets de recherche ont vu le jour, par exemple [Hobbs et al., 1996, Grishman, 1997]. De son côté, l'Europe s'est aussi intéressée au domaine et de nombreux systèmes ont été développés pour répondre aux besoins de l'industrie [Gaizauskas et Humphreys, 1997, Ecran, 1998].

En ce qui concerne l'extraction multilingue, un certain nombre de travaux ont été effectués ; cependant, il s'agit généralement d'une nouvelle version ou d'une adaptation d'un système monolingue existant. En effet, la majorité des projets multilingues ont initialement été développés pour l'anglais, puis une nouvelle version des ces systèmes a été développée pour traiter une seconde langue. C'est le cas, par exemple, du système JV-FASTUS, la version japonaise de FASTUS [Appelt et al., 1993]. À toutes fins pratiques, il s'agit de deux systèmes monolingues parallèles. Plus récemment, des systèmes initialement développés pour l'anglais, ont été adaptés pour traiter une seconde langue. C'est le cas, par exemple, de SOLOMON [Aone et al., 1993] ou du système bilingue M-LaSIE [Gaizauskas et Humphreys, 1997], une adaptation française de LaSIE. Cependant, à notre connaissance, il n'existe que peu de systèmes développés parallèlement pour traiter le français et l'anglais [Ecran, 1998, Tree, 1998, Mietta, 1998].

3 Le système EXIBUM

Le but du notre recherche est de développer un système bilingue d'extraction d'information en ré-utilisant autant que possible des outils existants, en particulier, ceux développés au RALI, à l'Université de Montréal. Dans le cadre de cette recherche, nous avons développé le système expérimental EXIBUM[1].

3.1 *Le domaine*

EXIBUM est un système d'extraction d'information oeuvrant dans le domaine des attentats terroristes. Le corpus est composé de communiqués de presse pris du World Wide Web. Le corpus français porte sur des attentats en Algérie alors que le corpus anglais traite du terrorisme aux États-Unis. Les dépêches utilisées proviennent des agences de presse Reuters, Algérie-Presse-Service (APS), Agence France-Presse (AFP), Associated Press et EmergencyNet NEWS Service. Actuellement, le corpus ne contient que 40 textes (20 en français et 20 en anglais), mais nous comptons l'élargir prochainement. Le domaine des attentats terroristes a été choisi pour plusieurs raisons :

– Ce domaine a été retenu au MUC-4 [MUC-4, 1992] ; nous fournissant ainsi des points de repère et de comparaison intéressants.

[1] EXIBUM est un acronyme de système d'EXtraction d'Information Bilingue de l'Université de Montréal.

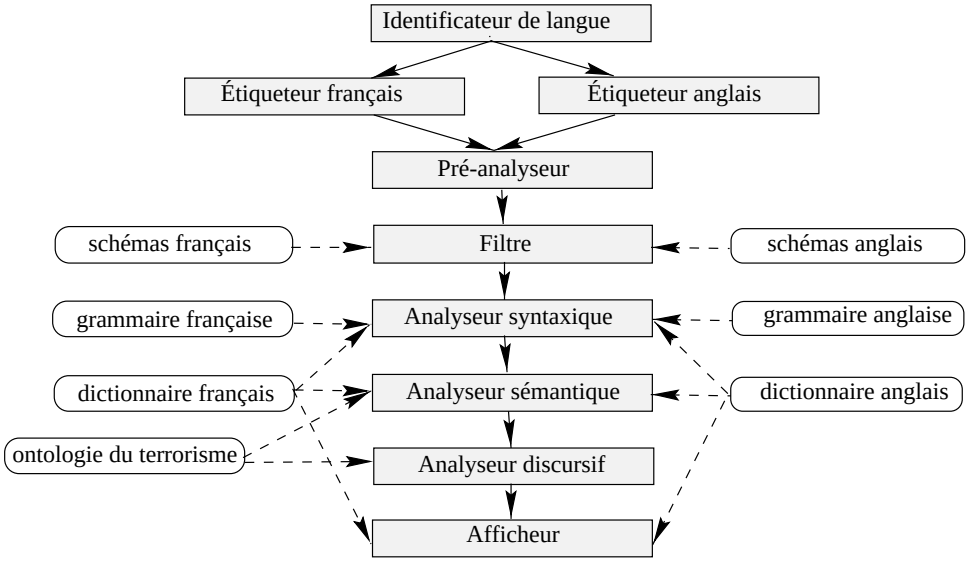

FIG. 1: *Architecture de* EXIBUM

- Un grand nombre de dépêches dans ce domaine sont disponibles sur le Web.
- Ces dépêches forment des textes réels, qui ne sont pas rédigés dans le but d'y faire de l'extraction d'information.
- Les dépêches typiques sont de longueur raisonnable ; en moyenne, les textes comptent 400 mots.

3.2 *L'architecture générale d'*EXIBUM

Comme l'illustre la figure 1, EXIBUM possède une architecture typique des systèmes d'extraction d'information [Hobbs, 1993]. Pour illustrer le processus d'extraction, voyons comment EXIBUM extrait l'information de la figure 2 à partir de la phrase :

> *Les quotidiens algériens ont publié samedi un résumé de ces pamphlets et ont rapporté que des rebelles du GIA avaient attaqué un poste militaire à Larbaa à la fin du mois d'avril, tuant 80 soldats et faisant 40 otages.*[2]

3.2.1 *L'identificateur de langue*

Idéalement, un système d'extraction d'information multilingue est en mesure d'analyser une suite continue de textes écrits en différentes langues, sans interruption humaine. Malheu-

[2]extrait de la dépêche Reuters du 2 mai 1998, *Algérie-Le GIA appelle à l'intensification de la lutte armée*

ÉVÉNEMENT		CIBLE PHYSIQUE	
date :	à la fin du mois d'avril, samedi	**nom** :	un poste militaire
lieu :	Larbaa	**type** :	un poste militaire : bâtiment
type :	personne(s) tuée(s), enlèvement	**nombre** :	un poste militaire : 1
état :	accompli	**effet** :	un poste militaire : détruit ou endommagé
INSTRUMENT		CIBLE HUMAINE	
nom :	-	**nom** :	-
type :	-	**description** :	80 soldats
		type :	80 soldats : militaire, 40 otages : civil
AUTEUR		**nombre** :	80 soldats : 80, 40 otages : 40
individu :	-	**effet** :	80 soldats : mort
organisation :	des rebelles du GIA		40 otages : mort ou blessé
responsabilité :	des rebelles du GIA : responsable		

FIG. 2: *Exemple d'information extraite*

reusement, certaines composantes du systèmes sont particulières à chaque langue (par ex. la grammaire, le dictionnaire). Pour éviter qu'une source extérieure (voire un humain) ne doive indiquer au système dans quelle langue le texte est écrit, nous avons doté EXIBUM d'un module d'identification de langue.

Le module d'identification de langue identifie dans quelle langue le texte est écrit, avec quel jeu de caractères il est encodé et, si nécessaire, le recode dans le jeu de caractères compris par EXIBUM. Le cœur de ce module est le système SILC[3]. SILC utilise un modèle de Markov (trigramme) qui postule que la probabilité d'observer un caractère est fortement conditionnée par ses deux caractères précédents. En français, par exemple, après les caractères *to*, il est beaucoup plus probable de trouver le caractère *u* que le caractère *o*. Par contre, en anglais, la séquence *too* est plus probable que la séquence *tou*. Grâce à ces observations, en rencontrant la séquence *tou* dans un texte, SILC en déduira qu'il est plus probable que le texte soit écrit en français qu'en anglais. Grâce à cette technique, SILC peut identifier pas moins de 25 langues, chacune encodée avec deux ou trois jeux de caractères différents. La performance de SILC est presque sans faille lorsqu'on lui soumet des textes raisonnablement propres et suffisamment longs (plus de 50 caractères).

Une fois que SILC a identifié la langue et le type d'encodage du texte, l'identificateur de langue détermine si EXIBUM est en mesure d'analyser le document. Si le texte n'est ni en français, ni en anglais, un message approprié est affiché et le texte ne sera pas soumis au reste du traitement. Si EXIBUM reconnaît la langue, le texte est recodé en `ISO Latin-1`, jeu de

[3]SILC est un acronyme de Système d'Identification de la Langue et du Codage (http ://www-rali.iro.umontreal.ca/ProjetSILC.fr.html)

```
phrase   ([J0, J1, J2, J3, ..., J36, J37, J38, J39]) :-
         ⋮

         J15 = [can : du, lex : des, cat : dete, type : arti, def : oui, genre : masc, nomb : plur],
         J16 = [can : rebelle, lex : rebelles, cat : nomc, genre : masc, nomb : plur],
         J17 = [can : du, lex : du, cat : dete, type : arti, def : oui, genre : masc, nomb : sing],
         J18 = [can : 'GIA', lex : 'GIA', cat : nomp, genre : masc, nomb : sing, type : pgeo],
         J19 = [can : avoir, lex : avaient, cat : verb, temps : imp, mode : ind, nomb : plur, pers : 3],
         J20 = [can : attaquer, lex : attaqué, cat : verb, mode : part, temps : pass, genre : masc, nomb : sing],
         ⋮

         J36 = [can : faire, lex : faisant, cat : verb, mode : part, temps : pres],
         J37 = [can : '40', lex : '40', cat : quan, genre : masc, nomb : plur, def : oui],
         J38 = [can : otage, lex : otages, cat : nomc, genre : masc, nomb : plur],
         J39 = [can : '.', lex : '.', cat : punc, type : finP].
```

FIG. 3: *Résultat de l'étiqueteur lexical en Prolog*

caractères utilisé par EXIBUM, et passé aux prochains modules.

Dans notre exemple, SILC identifie qu'il est très probable qu'il s'agisse d'un texte en français encodé en `ISO Latin-1`. EXIBUM sera donc en mesure d'analyser le texte sans recodage et, lorsque nécessaire, il activera ses composants français.

3.2.2 *Les étiqueteurs lexicaux*

L'étiquetage lexical est effectué par deux étiqueteurs stochastiques : un pour le français (TAG-F) et un pour l'anglais (TAG-E) [Foster, 1991]. Ces étiqueteurs se basent aussi sur un modèle de Markov pour déterminer la probabilité qu'un mot possède un ensemble de traits morpho-syntaxiques et sémantiques en fonction du résultat des mots précédents. TAG-E et TAG-F ont tous deux été entraînés sur un extrait d'environ 100 000 mots du corpus Hansard (actes du Parlement du gouvernement fédéral canadien). Bien qu'ils aient été entraînés sur un domaine différent du nôtre, ces étiqueteurs sont assez robustes pour être utilisés dans le cadre de notre recherche.

La figure 3 illustre le type de résultat de l'étiquetage lexical avec les actes terroristes de notre corpus. Cette figure sera expliquée en détails à la prochaine section.

3.2.3 *Le transcripteur*

Le transcripteur a pour but de créer une passerelle entre l'étiqueteur lexical et le reste du programme, écrit en Prolog. Le but du transcripteur est donc de réécrire le résultat de l'étiqueteur lexical et de classifier les traits en une liste d'attributs/valeurs, de façon à générer

une structure Prolog. Le même module transcrit le résultat de TAG-F et celui de TAG-E.

Dans notre exemple, le transcripteur génère la structure Prolog de la figure 3. Cette structure indique qu'il n'y a qu'une phrase dans le texte (1 seul prédicat `phrase`) qui contient 40 jetons[4] (`J0` à `J39`). Pour chaque jeton, une liste d'attributs est spécifiée. Par exemple, le mot *rebelles* (`J16`) est étiqueté comme ayant la forme canonique `rebelle`, la forme lexicale (fléchie) `rebelles`, est un nom commun masculin, pluriel (`cat : nomc, genre : masc, nomb : plur`).

3.2.4 *Le filtre*

Le but du filtre est d'identifier les phrases du texte qui font référence à des événements pertinents. Pour effectuer cette tâche, EXIBUM utilise deux bibliothèques de schémas basés sur des mots-clé (verbes ou noms) pertinents au domaine. Les schémas étant basés sur la forme lexicale des mots, EXIBUM a besoin d'une bibliothèque pour les schémas français et d'une seconde pour les schémas anglais. Ces schémas ont été identifiés par une analyse manuelle de corpus. Les cas de polysémie (verbale ou nominale) sont minimisés car seul un sous-langage est considéré. En effet, les schémas ont été développés pour un domaine particulier et pour un type d'information bien précis à extraire. De plus, comme nous verrons à la section 3.2.6, l'analyseur sémantique s'efforce à son tour de restreindre le sens des mots-clé en imposant des attributs sémantiques à ses arguments. Si un attribut imposé par l'analyseur sémantique entre en conflit avec un attribut déjà identifié, l'analyseur syntaxique tentera de trouver un autre argument, sinon le schéma sera rejeté. Prenons comme exemple le verbe *souffler*. *Souffler* possède plusieurs sens en français général ; il peut signifier *expirer*, *respirer avec peine*, *dire à voix basse*, ... Dans notre domaine, le verbe *souffler* est intéressant uniquement dans son sens *détruire par l'effet du souffle* (ex. *l'autobus est soufflé par l'explosion*). Comme EXIBUM n'analyse que des textes portant sur des attentats terroristes, les probabilités que ce verbe soit utilisé dans un autre sens sont assez faibles, mais pas nulles. L'analyseur sémantique restreindra ensuite l'objet de *souffler* à être un bâtiment ou une automobile, ce qui devrait filtrer un certain nombre de sens qui ne nous intéressent pas.

Les schémas sont générés dynamiquement à partir d'une classification lexicale de comportement syntaxique. Ainsi les verbes *attaquer* et *tuer* généreront les mêmes schémas, alors que

[4]Un jeton (ou *token*) est un occurrence d'un mot ou d'une marque de ponctuation dans un texte.

mort générera des schémas différents. Parmi les schémas français on retrouve :

schéma 1a :	X **attaque**	schéma 2a :	X **tue**
schéma 1b :	X **attaque** Y	schéma 2b :	X **tue** Y
schéma 1c :	X **attaque** Y **avec** I	schéma 2c :	X **tue** Y **avec** I
schéma 1d :	X **attaque avec** I	schéma 2d :	X **tue avec** I
schéma 3a :	**mort de** Y	schéma 4 :	X **fait** Y **otages**
schéma 3b :	X **revendique la mort de** Y		

où X fait référence à l'auteur de l'attentat, Y aux victimes et I, à l'instrument.

Le filtre ne s'occupe pas d'instancier les arguments des schémas ; il ne fait qu'identifier les phrases pertinentes et les schémas activés. Il passe ensuite cette information à l'analyseur syntaxique, qui est en mesure d'instancier les arguments des schémas activés.

Dans notre exemple, le filtre identifie que la phrase est pertinente en activant les schémas 1b, 2b et 4.

schéma 1b : *... des rebelles du GIA avaient* **attaqué** *un poste militaire ...*
schéma 2b : *... à la fin du mois d'avril,* **tuant** *80 soldats et faisant 40 otages.*
schéma 4 : *... tuant 80 soldats et* **faisant** *40* **otages.**

Les schémas sont identifiés grâce à une analyse de corpus manuelle. Il s'agit d'une étape longue, pénible et sujette à des erreurs. Pour faciliter cette tâche, certains outils automatiques ou semi-automatiques ont été proposés [Riloff, 1998]. Pour accélérer ce processus et s'assurer d'une couverture plus large, nous nous penchons pour l'instant sur l'utilisation du WordNet [Miller, 1990]. Le WordNet est un système de référence lexicographique où les mots anglais sont organisés dans des classes de synonymie. De plus, le WordNet a l'avantage d'être disponible en version Prolog ; il est donc facile de l'intégrer à EXIBUM. Le WordNet nous permettrait de ne définir qu'une classe générique de schémas et grâce à ses relations de synonymie et d'hyponymie, des instances de schémas pourraient être générées automatiquement. Malheureusement, à notre connaissance le WordNet français n'est pas encore disponible[5].

3.2.5 *L'analyseur syntaxique*

L'analyseur syntaxique prend les phrases pertinentes identifiées par le filtre et y effectue une analyse syntaxique partielle pour identifier le syntagme verbal dans lequel le mot-clé se trouve (s'il s'agissait d'un mot-clé verbe) et les syntagmes nominaux autour de celui-ci. Pour

[5] *cf.* http ://www.let.uva.nl/ ewn

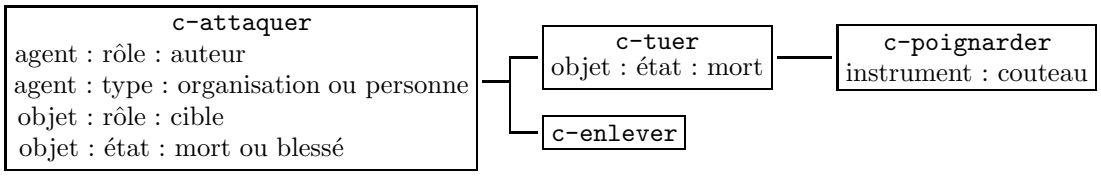

FIG. 4: *Extrait de l'ontologie*

faire ceci, l'analyseur syntaxique utilise, pour l'instant, une grammaire simple développée empiriquement pour chaque langue. Nous comptons bientôt formaliser ce module soit en utilisant deux analyseurs syntaxiques plus robustes, soit en utilisant des grammaires plus formelles.

Dans notre exemple, l'analyseur syntaxique détermine que les jetons J19-J20 (*avaient attaqué*) forment un syntagme verbal, les jetons J16-J18 (*des rebelles du GIA*), son sujet et J21-J23 (*un poste militaire*), son complément d'objet direct. EXIBUM est ainsi en mesure d'instancier les schémas à :

schémas 1b : *(X= des rebelles du GIA)* **attaque** *(Y= un poste militaire)*
schémas 2b : *(X = ?)* **tue** *(Y = 80 soldats)*
schémas 4 : *(X = ?)* **fait** *(Y = 40)* **otages**

EXIBUM sait maintenant que l'auteur de l'attaque est *des rebelles du GIA* et la cible est *un poste militaire*. L'auteur de la tuerie est inconnu ; par contre, les victimes sont *80 soldats*, et finalement l'auteur de la prise d'otages est inconnu, mais les victimes sont *40 otages*.

3.2.6 *L'analyseur sémantique*

Le but de l'analyseur sémantique est d'effectuer des inférences générales et particulières au monde du terrorisme. Ce module est indépendant de la langue, car il se base sur une ontologie du monde du terrorisme. En effet, chaque schéma est associé à un concept de l'ontologie qui permet de préciser ou de compléter l'information trouvée par l'analyseur syntaxique. Par exemple, le schéma français <agent> **tue** <objet> et le schéma anglais <agent> **kills** <objet> sont tous deux associés au concept `c-tuer`. De son côté, le concept `c-tuer` impose à son agent d'être une organisation ou une personne, auteur d'un attentat et impose à son objet d'être une personne morte, cible (ou victime) d'un attentat. La figure 4 illustre un extrait de l'ontologie.

Dans notre exemple, le résultat de l'analyseur sémantique, illustré à la figure 5, contient

événement	**type**	**date**	**lieu**	**objet**	**agent**
ev1	c-attaquer	ent1	ent2	ent3	ent4
ev2	c-tuer	ent1	ent2	ent5	ent6
ev3	c-enlever	ent1	ent2	ent7	ent8

entité	**forme lexicale**	**type**	**nombre**	**rôle**	**état**
ent1	*à la fin du mois d'avril, samedi*	-	-	-	-
ent2	*Larbaa*	-	-	-	-
ent3	*un poste militaire*	bâtiment	1	cible	détruit ou endommagé
ent4	*des rebelles du GIA*	criminel	>1	auteur	-
ent5	*80 soldats*	militaire	80	cible	mort
ent6	*Groupe Islamique Armé*	organisation	-	suspect	-
ent7	*40 otages*	civil	40	cible	mort ou blessé
ent8	*Groupe Islamique Armé*	organisation	-	suspect	-

FIG. 5: Résultat de l'analyseur sémantique

3 événements ; un événement pour chaque schéma activé. Les entités impliquées dans ces événements ont été qualifiés soit par l'analyseur syntaxique, soit par l'analyseur sémantique en consultant le dictionnaire ou l'ontologie. Par exemple, c'est est consultant le dictionnaire qu'EXIBUM a déterminé que *un poste militaire* est un bâtiment, c'est l'ontologie qui a imposé que le poste soit détruit ou endommagé (mort ou blessé) et c'est l'analyseur syntaxique qui a déterminé le nombre de postes attaqués.

Actuellement, l'analyseur sémantique n'est pas en mesure de trouver l'antécédent d'une référence anaphorique. Lorsque que la forme lexicale d'une entité n'est composée que d'un pronom, l'analyseur sémantique l'ignore. Par exemple, dans la phrase :

Le juge français, qui a entendu Zaoui en présence du procureur fédéral helvétique,
Mme Carla del Ponte, l'a mis en examen pour association de malfaiteurs ...

l'analyseur sémantique ne peut identifier *Zaoui* comme antécédent du pronom *l'*. L'objet de la mise en examen ne sera donc identifié comme étant Zaoui. Pour déréférencer assez bien une expression anaphorique, la recherche du groupe nominal précédent s'avère insuffisante. Comme la phrase précédente l'illustre, la sémantique et la pragmatique des verbes et des groupes nominaux du domaine de discours doivent être considérées.

3.2.7 *L'analyseur discursif*

L'analyseur discursif tente de regrouper l'information retrouvée à différents endroits du texte en une seule entité. Pour l'instant, seuls les événements mentionnés à différents endroits dans

événement	**type**	**date**	**lieu**	**objet**	**agent**
ev1	c-attaquer	ent1	ent2	ent3	ent4

entité	**forme lexicale**	**type**	**nombre**	**rôle**	**état**
ent1	*à la fin du mois d'avril, samedi*	-	-	-	-
ent2	*Larbaa*	-	-	-	-
ent3	*un poste militaire*	bâtiment	1	cible	détruit ou endommagé
	80 soldats	militaire	80	cible	mort
	40 otages	civil	40	cible	mort ou blessé
ent4	*des rebelles du GIA*	criminel	>1	auteur	-

FIG. 6: Résultat de l'analyseur discursif

le texte sont fusionnés. EXIBUM considère que 2 événements font référence au même attentat si et seulement si leurs concepts sont compatibles et leurs dates et lieux sont les mêmes. EXIBUM tente d'abord de fusionner les événements provenant de la même phrase, puis considère les événements du reste du texte. Cette stratégie donne priorité aux événements les plus proches textuellement, ainsi ayant plus de chance de faire référence au même attentat.

Dans notre exemple, l'analyseur discursif fusionne tous les événements identifiés par l'analyseur sémantique en un seul (*cf.* la figure 6). En effet, les 3 événements se sont produits à la même date (`ent1` = *à la fin du mois d'avril, samedi*) et au même endroit (`ent2` = *Larbaa*). De plus, les concepts `c-attaquer, c-tuer` et `c-enlever` sont tous les 3 compatibles. EXIBUM en conclue alors qu'un seul attentat s'est produit. Ce résultat est ensuite passé à l'afficheur qui remplit le formulaire (*cf.* la figure 2) à partir de l'événement extrait.

4 Conclusion

Dans cet article, nous avons présenté le système EXIBUM. Ce système est actuellement en développement ; la majorité des modules peuvent donc être nettement améliorés. Cependant, les résultats préliminaires obtenus démontrent l'avantage de la réutilisation de modules existants et la faisabilité du développement rapide d'un système d'extraction d'information multilingue.

À notre connaissance, EXIBUM est le premier système d'extraction d'information développé parallèlement pour le français et l'anglais. Ainsi, le traitement d'une langue n'est pas une adaptation du traitement de l'autre. Chaque langue est considérée à part entière et possède son propre module d'étiquetage lexical et ses propres ressources linguistiques. De plus, à notre connaissance, EXIBUM est le seul système a pouvoir changer dynamiquement son comportement

en identifiant lui-même dans quelle langue le texte est rédigé.

Parmi les travaux futurs, nous envisageons à court terme le l'utilisation d'un analyseur syntaxique plus robuste. À plus long terme, nous envisageons d'explorer l'utilisation du WordNet comme outil de génération de schémas anglais.

Remerciements

Nous tenons à remercier Jérôme Soucy pour son travail sur l'implantation d'EXIBUM, ainsi que les membres du RALI et les re-lecteurs anonymes pour leurs commentaires judicieux.