Extraction d'informations multilingues utilisant des paraphrases

François-Régis Chaumartin (1, 2)

(1) Alpage – Université Paris 7, UFRL, Case 7003, 75251 Paris cedex 5
fchaumartin@linguist.jussieu.fr

(2) Proxem – 7 impasse Dumur, 92110 Clichy
frc@proxem.com

Résumé Cette démonstration présente un composant d'extraction d'informations multilingues qui permet (1) d'associer un ensemble de paraphrases à un prédicat, puis (2) de rechercher sur le Web des instances de ce prédicat. La méthode d'extraction d'informations utilisée s'appuie sur une traduction des paraphrases en patrons syntaxiques ; les phrases susceptibles de contenir l'information recherchée font l'objet d'une analyse syntaxique en dépendances, puis d'un appariement de formes avec les graphes syntaxiques des patrons. Cette méthode fournit des résultats précis, au prix d'un temps de calcul élevé.

Abstract This demonstration introduces a multilingual Information Extraction component. This component makes it possible (1) to associate a set of paraphrases with a predicate, then (2) to search on the Web instances of this predicate. The method used during the Information Extraction stage relies on a translation of the paraphrases as syntactic patterns; the sentences likely to contain the right information are parsed and produce a dependency output; a pattern matching is then achieved against the dependency graph of the paraphrases. This method provides precise results, at the price of high CPU usage.

Mots-clés : extraction d'information, paraphrases, analyse syntaxique, lexique sémantique
Keywords: Information Extraction, paraphrases, parsing, semantic lexicon

1 Introduction

La démonstration présente un composant d'extraction d'informations multilingues utilisé par une application de veille économique. Ce composant est orienté vers la tâche de reconnaissance de formes (*template filling*) ; il détecte des prédicats dans un texte en anglais ou en français, et remplit automatiquement les valeurs des arguments de ces prédicats. L'application de veille économique automatise des scénarios complexes d'extraction d'informations, et permet de les rejouer avec des paramètres différents ; ainsi, à partir du nom d'une société, on peut chercher les événements marquants (rachat d'autres sociétés, nomination de dirigeants, dépôt de brevets, événements juridiques…). Notre objectif est triple : exprimer aussi simplement que possible les informations cherchées ; privilégier des réponses précises ; être indépendant d'une langue donnée ou d'un analyseur particulier.

2 Evaluation sur un exemple d'acquisition de sociétés

Le prédicat `acquisition(buyer:`*`properNoun`*`, company:`*`properNoun`*) est associé, comme montré en Figure 1, à onze réalisations linguistiques en anglais et deux en français[1,2]. L'application interroge des moteurs de recherche (Google, Yahoo! et Microsoft Live Search) avec les mots clés de chaque paraphrase et collecte une liste de documents, qui sont segmentés en phrases. Celles contenant tous les mots clés (dans le même ordre) sont retenues en tant que phrases candidates, puis testées par le composant d'extraction d'informations.

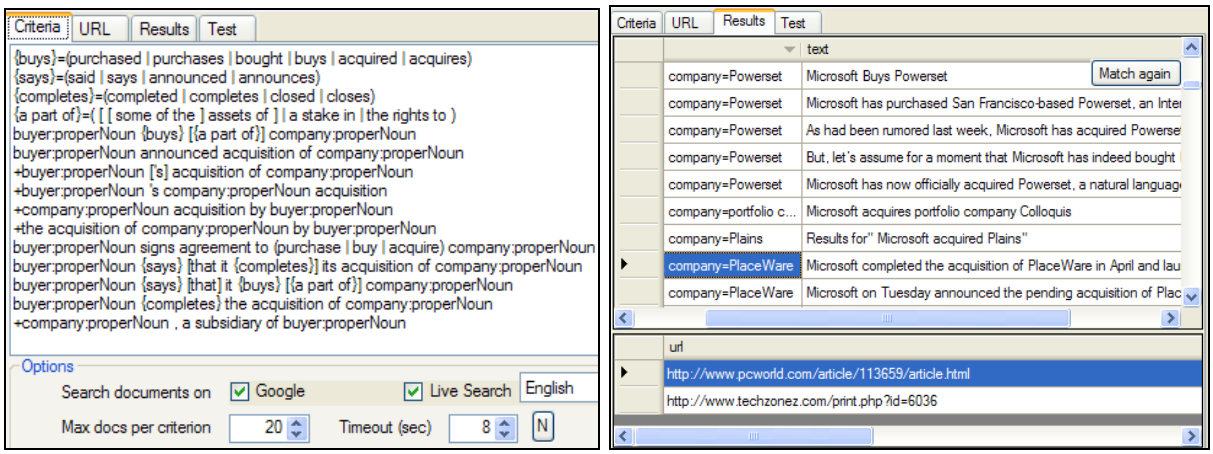

Figure 1 : Interfaces de saisie des critères de recherche et d'affichage des résultats

Nous avons évalué le composant en le testant sur les rachats de sociétés et prises de participations effectués par Microsoft. L'application trouve 2 160 documents à partir du Web, contenant 4 367 phrases candidates, et extrait une information à partir de 1 353 phrases (en un peu moins d'une heure de traitement). À peu près 10% de ces résultats sont erronés (segmentation ou analyse syntaxique incorrecte). Une fois regroupés, les résultats extraits restants concernent 245 instances distinctes de sociétés. 46 instances ne sont pas pertinentes[3]. Nous avons comparé les 199 autres résultats obtenus avec ceux d'un site d'informations financières de référence (AlacraStore), qui énumère 195 sociétés (132 rachats et 63 prises de participation). L'application trouve 182 de ces 195 sociétés, plus 17 autres (Finjan Software, Green Button, Changhong…) ayant effectivement fait l'objet d'une opération par Microsoft.

3 Architecture technique du composant

Le composant d'extraction d'informations est un ajout récent à la plateforme de traitement linguistique Antelope (Chaumartin, 2009)[4]. Conçue pour traiter différentes langues, elle

[1] Ces différentes paraphrases ont été créées manuellement par examen d'un corpus de dépêches financières.

[2] Le préfixe + indique un groupe nominal. Les mots entre accolades sont des macros qui seront substituées.

[3] 16 rumeurs de rachat non avérés (Yahoo!, Disney…), 15 plaisanteries de 1er avril (rachat d'IBM, de l'église catholique…), 12 noms de logiciels appartenant aux sociétés rachetées, et 3 des dirigeants de ces sociétés.

[4] Cette plateforme, en partie inspirée de la Théorie Sens-Texte, permet l'analyse syntaxique et sémantique de textes. Antelope intègre des données linguistiques à large couverture provenant de différentes sources (WordNet, VerbNet, SUMO…) et des composants d'étiquetage morphosyntaxique, d'analyse syntaxique, de résolution d'anaphores, de désambiguïsation lexicale et syntaxique et d'étiquetage de rôles sémantiques.

intègre pour l'instant des analyseurs syntaxiques et lexiques sémantiques pour l'anglais et le français. Elle formalise un modèle linguistique unifié permettant de manipuler les différents niveaux de représentation d'une façon homogène, quelle que soit la langue du texte.

Chaque paraphrase est transformée en une forme logique, qui permet de tester si une phrase donnée correspond au patron et d'en extraire la valeur des arguments. Une transformation intermédiaire convertit la paraphrase en un « exemple canonique » qui fait l'objet d'une analyse syntaxique en dépendances, comme indiqué à gauche de la Figure 2 (sur la partie droite figure une phrase où on reconnait cette forme). L'analyse syntaxique de la paraphrase est transformée en un programme PROLOG qui constitue sa forme logique ; ce programme cherche des mots reliés entre eux par certaines dépendances, en testant la partie du discours de chaque mot et d'éventuelles contraintes de sélection ; l'ordre des mots (consécutifs, mais non forcément contigus) est vérifié. L'extraction d'informations sur une phrase se ramène alors simplement à l'identification d'un sous-graphe au sein d'un autre graphe. Ce mécanisme tolère la présence de mots intercalés entre ceux qu'on cherche, y compris quand il s'agit de sous-phrases longues (appositions, relatives…). Grâce à la couche d'abstraction qu'apporte la plateforme Antelope, ce processus est largement indépendant de la langue et de l'analyseur syntaxique considérés. En termes de temps de calcul, l'analyse syntaxique représente une opération longue (quelques secondes par phrase). Il est toutefois possible de pré-calculer la forme logique des paraphrases, puis de stocker ce résultat, qui est ensuite prêt à l'emploi.

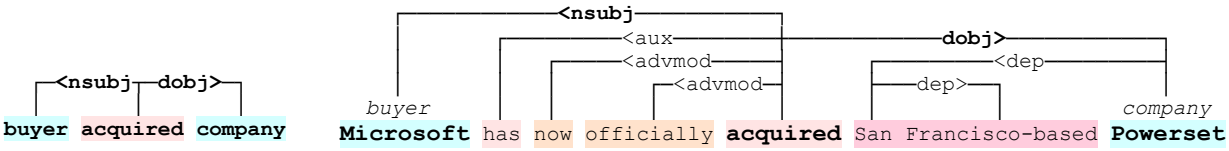

Figure 2 : Analyse en dépendances d'une paraphrase et d'une phrase lui correspondant

4 Conclusion

Les systèmes d'extraction d'informations utilisent classiquement des techniques d'analyse robustes (Tannier, 2006) : patrons morpho-syntaxiques, automates à états finis, collocations obtenues par des statistiques… Nous avons présenté un composant effectuant une analyse syntaxique, d'une façon indépendante de la langue. L'intérêt de ce composant vient d'une part de la précision de ses résultats, et d'autre part de la simplicité avec laquelle un utilisateur peut associer différentes paraphrases à un prédicat, avec une approche basée sur des exemples. Nous envisageons deux pistes d'amélioration des traitements linguistiques : un mécanisme de résolution d'anaphores permettra de dépasser les limites imposées par le traitement d'une seule phrase ; l'application de la forme logique sur une représentation des dépendances en syntaxe profonde (plutôt qu'en syntaxe de surface) devrait améliorer sensiblement le rappel du système, en permettant d'exploiter les relatives, constructions passives et noms prédicatifs.

Références

CHAUMARTIN F.R. (2009). Antelope : une plateforme industrielle de traitement linguistique. *TAL* volume 49, n°2.

TANNIER X. (2006). *Traitement automatique du langage naturel pour l'extraction et la recherche d'informations*. Rapport de recherche 2006-400-006. Saint-Etienne : ENSM.