Du linguistique au conceptuel : identification de relations conceptuelles à partir de textes

Nathalie Aussenac-Gilles et Nathalie Hernandez

IRIT, Université de Toulouse,
{aussenac, hernande}@irit.fr
118, route de Narbonne - 31062 Toulouse cedex 9

1 Introduction

Le projet DAFOE4App propose de développer une plate-forme assurant un support pour la construction d'ontologies à partir de texte (Charlet *et al.*, 2008). Une de ses originalités est de reprendre les niveaux intermédiaires de représentation de la méthode Terminae (Aussenac-Gilles et al., 2008) pour assurer un passage progressif du texte à une ontologie. Nous avons donc revu le processus de repérage, représentation et validation de représentations conceptuelles à partir de leurs traces en corpus au regard des niveaux de la plate-forme DAFOE. Ce travail se veut une actualisation des principes retenus dans le logiciel Caméléon pour définir un module d'extraction de relations à partir de texte associé à DAFOE sous forme de greffon. Les différents niveaux de représentation de DAFOE situent les contributions possibles de ce greffon pour enrichir le modèle en cours de construction. Ils permettent également de considérer clairement des approches alternatives et complémentaires à l'approche par patrons (apprentissage de patrons spécifiques au domaine, réutilisation de relations trouvées dans des ontologies existantes, etc.) qui sont prévues pour des développements futurs.

Dans cet article, nous définissons les principes de ce module, ses fonctionnalités et les étapes clés du processus d'identification de relation qu'il supporte. Après un rappel de différentes techniques possibles pour le repérage de relations à partir de textes, nous présentons les niveaux de modélisation de DAFOE, les fonctionnalités du greffon et ses liens avec la plateforme.

2 Extraction de relations : degrés de conceptualisation

Différentes approches, statistiques ou linguistiques, ont été proposées pour extraire des relations à partir de textes dans la perspective de construire une ressource termino-ontologique. Ces méthodes et les outils associés assistent plus ou moins loin l'identification et la représentation des relations. Elles peuvent soit juste localiser des contextes en corpus, soit aider à identifier les termes en relation et la nature des

relations, soit encore, définir des concepts et des relations sémantiques formalisées. Ainsi, elles participent à la conception de ressources de niveaux formels différents (terminologie, thesaurus, réseau de concepts, ontologie formelle). Elles peuvent également être vues comme intervenant à différentes étapes de la conception d'une ontologie à partir de textes.

Un premier ensemble d'outils (Hearst, 1992) (Morin, 1999) vise à identifier des relations lexicales entre couples de termes. Proches des textes, ces approches effectuent une analyse linguistique des documents et proposent des relations candidates qui pourront soit être intégrées dans une terminologie ou un thesaurus, soit servir de point de départ à une conceptualisation de la connaissance du domaine. Un deuxième ensemble d'approches (Aussenac & Séguéla, 2000) ont pour objectif d'extraire des relations qui pourront participer à une modélisation sous la forme d'un réseau termino-conceptuel. D'un niveau d'abstraction supérieur, ce type de relations définit les interactions qui existent entre les concepts ou termes désambiguïsés du domaine. Finalement, plus récemment, des approches ont été proposées pour extraire des relations entre concepts formellement définis dans une ontologie. Par exemple, le système Scarlet (Sabou *et al.*, 2008) exploite les relations préalablement définies dans des ontologies accessibles en ligne pour proposer des relations entre concepts d'une ontologie en cours de construction.

Pour chacun des types d'approches, il s'agit dans un premier temps d'extraire les relations candidates puis de les valider et de les intégrer dans la ressource. L'interprétation d'un humain s'impose pour identifier la nature de cette relation lexicale (sa sémantique et les termes reliés). En effet, des variations dans la formulation des relations et des termes apparaissent en corpus, y compris dans des domaines spécialisés (Condamines, 2007). Pour superviser ce processus, les outils implémentant ces approches doivent fournir un environnement et des interfaces facilitant cette tâche. A notre connaissance, ces aspects sont peu développés.

3 Extraction de relations par patrons : Caméléon

Nous nous intéressons à l'extraction de relations à partir de patrons, qui repose sur le postulat selon lequel une relation peut *a priori* être inférée à partir d'analyses textuelles. Un patron correspond à une caractérisation abstraite de toutes les réalisations langagières associées à la relation, et peut être composé d'éléments lexicaux, grammaticaux ou sémantiques (Auger & Barrière, 2008). Dans ce type d'approche, le corpus doit préalablement être traité par un étiqueteur. L'extraction s'organise ensuite en trois étapes : une phase d'acquisition pendant laquelle les patrons sont créés, une phase de projection sur le corpus pendant laquelle les patrons servent à extraire des phrases contenant potentiellement des relations entre couples de termes, couples de termino-concept ou couples de concepts et une phase de modélisation pendant laquelle chaque relation candidate est analysée et intégrée dans la ressource.

Nous avions proposé un premier outil, Caméléon, qui implémente la mise au point et la projection de patrons sur corpus pour identifier des relations et des concepts pour la construction d'ontologies (Aussenac-Gilles & Séguéla, 2000). L'utilisation de Caméléon a confirmé qu'ajuster manuellement des patrons et filtrer les phrases correspondantes pour identifier des relations conceptuelles était coûteux et fastidieux. De plus, le logiciel situe mal les validations requises et le degré de formalisation des représentations obtenues (Aussenac-Gilles & Jacques, 2008). D'autres travaux mettent également en avant les efforts demandés lors de la phase de création des patrons et de la validation des résultats.

Notre objectif est de partir de ces expériences pour préserver la robustesse de l'extraction tout en améliorant le processus d'acquisition et en réduisant la charge cognitive demandée à l'utilisateur. Nous voulons également intégrer l'utilisation de l'outil d'extraction de relations à d'autres logiciels d'analyse du langage au sein d'un environnement de construction d'ontologies.

4 La plate forme DAFOE

La plateforme DAFOE proposée dans le cadre du projet ANR DaFOE4app[1] a pour but de définir un cadre méthodologique et un environnement de conception d'ontologies. La plateforme a été conçue pour être capable d'accueillir et d'intégrer les outils permettant d'aller de l'analyse d'un corpus à la définition d'une ontologie formelle du domaine. Elle repose sur une architecture orientée service facilitant l'accès et le stockage des modélisations et des corpus par le biais d'un noyau (Charlet *et al.*, 2008). La plateforme est organisée en couches :

- La couche 0 (corpus) permet à l'utilisateur de travailler à partir de textes. S'il possède des outils de TAL, il pourra visualiser les résultats de ses outils dans DaFOE. Sinon le texte brut est découpé en phrases.
- La couche 1 (terminologique) rend compte de l'analyse linguistique. Elle consiste à identifier des termes et des relations entre termes.
- La couche 2 (termino-conceptuelle) permet de représenter les termes désambiguïsés appelés termino-concepts et les relations désambiguïsées.
- La couche 3 (ontologique) formalise l'ontologie en OWL.

[1] http://dafoe4app.fr/

5 Un greffon d'extraction de relations dans DAFOE

Au niveau des différentes couches de la plateforme, les relations sont considérées et peuvent être saisies à la main. Nous proposons un greffon qui vise à faciliter leur acquisition à partir de textes et ce, à l'aide de patrons. Il implémente les trois étapes nécessaires au repérage de relations potentiellement présentes dans le corpus. Dans un premier temps, le greffon facilite la mise au point du patron. Il permet ensuite d'effectuer la projection du patron sur les phrases de corpus. Finalement, en présentant les contextes des phrases retrouvées par la projection, il assiste l'utilisateur dans l'identification de la relation et des arguments sur lesquels elle porte. L'utilisation de ce greffon suppose que le corpus a préalablement été traité dans la plateforme par un greffon permettant d'effectuer une analyse morphosyntaxique. Deux greffons de ce type sont en cours de construction, basés sur les l'analyseur syntaxique Syntex (Bourigault & Fabre, 2002) et le concordancier Yatea.

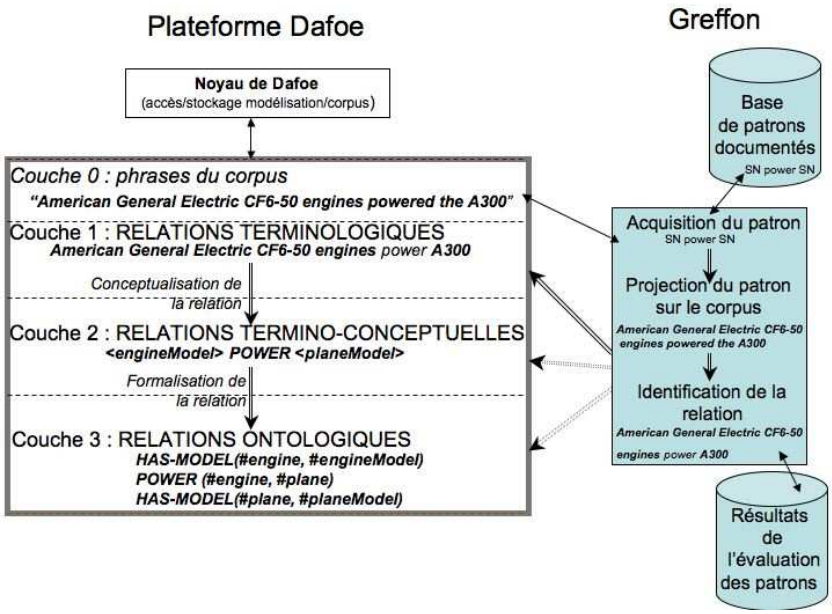

Figure 1 : Liens entre le greffon et la plateforme Dafoe

Les liens entre la plateforme Dafoe et notre greffon sont schématisés dans la figure 1. Le greffon projette le(s) patron(s) sélectionné(s) sur le corpus chargé dans la plateforme. Il exploite ensuite le modèle en cours de construction pour identifier les relations potentiellement présentes dans les phrases. Finalement les relations identifiées dans le greffon sont importées dans la plateforme et enrichissent le modèle en cours de construction. Dans l'exemple de la figure, le patron permettant d'extraire la relation «power » est sélectionné dans la base lors de l'étape d'acquisition. L'étape

de projection renvoie la phrase « American General Electric CF6-50 engines powered the A300 » avec comme contexte gauche du patron le terme « American General Electric CF6-50 engines » et comme contexte droit le terme «A300 ». L'utilisateur choisit ces deux termes présents dans le modèle en cours de construction dans la couche terminologique comme étant les arguments de la relation « POWER ». La relation terminologique ainsi identifiée est importée dans la plateforme pour enrichir le modèle. Cette relation est ensuite conceptualisée dans la plateforme sous la forme d'une relation termino-conceptuelle reliant les termino-concepts « engineModel » et « planeModel ». Cette nouvelle relation est finalement formalisée au niveau de la couche ontologique comme représentant trois relations conceptuelles HAS-MODEL entre les concepts formels #engine et #engineModel, POWER entre les concepts formels #engine et #plane et HAS-MODEL entre #plane et #planeModel.

6 Originalités

Tout d'abord, *cet outil facilite la réutilisation des patrons.* Pour cela, une base stocke des patrons déjà définis avec leur documentation, qui renseigne le type de la relation pour laquelle le patron a été écrit, le ou les corpus pour le(s)quel(s) la pertinence du patron a été vérifiée, l'étiqueteur de corpus nécessaire pour la projection du patron et l'auteur du patron. Ces informations facilitent le choix et l'adaptation du patron. L'utilisateur a la possibilité de rechercher un patron dans la base à partir de ces différents critères et de le réutiliser soit directement, soit en l'ajustant. Les résultats de la projection et l'évaluation de l'utilisateur sont également stockés de façon à pouvoir calculer par l'usage la pertinence du patron. Toujours dans un souci de réutilisation, une fonction d'import permet d'ajouter à la base des patrons conçus par d'autres applications, appris automatiquement ou définis par d'autres équipes.

Ensuite, *le greffon guide l'identification de la relation dans le contexte des phrases retournées par la projection,* en facilitant la sélection de la relation et de ses arguments. Pour cela, des interfaces d'interprétation des résultats de la projection présentent l'ensemble des informations nécessaires à l'analyse. Par exemple, pour identifier une relation terminologique, l'interface présente la phrase identifiée par le patron dans son contexte ; l'utilisateur peut accéder aux phrases qui la précèdent et qui la suivent dans le document. Ceci lui permet par exemple de résoudre une anaphore. L'interface met en valeur les contextes droit et gauche identifiés par le patron ainsi que les termes reconnus dans ces contextes ; l'utilisateur peut sélectionner ces termes comme étant arguments de la relation ou bien en renseigner de nouveaux. L'interface présente également le come de la relation associée à ce patron ainsi que la liste des relations déjà identifiées ; l'utilisateur peut choisir parmi celles-ci la relation reconnue dans la phrase, en proposer une autre, ou encore rejeter cette proposition.

Enfin, *ce greffon doit permettre de manipuler différents types de patrons,* définis avec des éléments lexicaux, syntaxiques et/sémantiques. Le jeu d'éléments composant

un patron est paramétrable, adaptable à la langue, aux étiquettes présentes dans les textes (grammaticales, dépendances syntaxiques, annotations sémantiques) et au niveau d'interprétation choisi. Ainsi, on envisage de pouvoir l'utiliser pour définir des patrons pour rechercher des relations lexicales (patrons lexico-syntaxiques classiques) ou des relations termino-ontologiques, grâce à des patrons utilisant des catégories sémantiques et des relations de dépendance syntaxique.

Une première version de *ce greffon est en cours de développement* à partir du logiciel Caméléon. Cette version permettra d'ici peu d'utiliser les patrons définis dans Caméléon pour importer des relations terminologiques au niveau de la couche 1 de Dafoe. A moyen terme, l'ensemble des fonctionnalités seront développées. A plus long terme, une nouvelle version permettra de définir et d'exploiter plus facilement des annotations sémantiques en s'appuyant sur une plate-forme standard d'analyse de textes (de type Gate ou LinguaStream). Surtout, des outils complémentaires, permettant d'alimenter la base de patrons par apprentissage, d'identifier des relations au niveau termino-conceptuel et au niveau ontologique (outil du type de Scarlet) pourront être définis comme d'autre greffons.

7 Références

AUBIN S., HAMON T.(2006), Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing (5th International Conference on NLP, TAL* 2006).

AUGER A., BARRIERE C. (2008), Pattern based approaches to semantic relation extraction : a state-of-the-art. *Terminology*, John Benjamins , 14-1,1-19.

AUSSENAC-GILLES N., DESPRES S., SZULMAN S. (2008), The TERMINAE Method and Platform for Ontology Engineering from texts, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. P. Buitelaar, P. Cimiano (Eds.), IOS Press, p. 199-223.

AUSSENAC-GILLES N., SEGUELA P. (2000), Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire. Numéro spécial linguistique de corpus*. A. Condamines (Ed.). Toulouse : Presse de l'UTM. 25 175-198

AUSSENAC-GILLES N., JACQUES M.-P. (2008), Designing and Evaluating Patterns for Relation Acquisition from Texts with CAMÉLÉON, Auger A. and Barriere C. (Eds.), *Terminology* 14-1, Pattern-based approaches to semantic relation extraction, Amsterdam/Philadelphia, John Benjamins Publishing Company, 14-1, p. 45-73.

D. BOURIGAULT, C. FABRE (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires*, (25):131–51, Numéro spécial « Sémantique et corpus ».

BUITELAAR P., CIMIANO P., MAGNINI B. (2005), *Ontology Learning From Text: Methods, Evaluation and Applications*. IOS Press.

CHAGNOUX M., HERNANDEZ N., AUSSENAC-GILLES N. (2008). An interactive pattern based approach for extracting non-taxonomic relations from texts. *Workshop on Ontology Learning and Population (associated to ECAI'08) (OLP 2008), Patras (Greece)*, Univ. of Patras, p. 1-6.

CHARLET J., AUSSENAC-GILLES N., PIERRA G., NADA N., SZULMAN S., TEGUIAK H.V. (2008), DAFOE : Une plateforme multi-méthodes et multi-modèles pour le développement d'ontolo-

gies de domaine. Dans : *Journées Francophones sur les Ontologies (JFO 2008), Lyon (F.), 01-03 dec 08*, D. Benslimane, C. Roche, S. Spaccapietra (Eds.), ACM, p. 1-12.

CONDAMINES A. (2007), L'interprétation en sémantique de corpus : le cas de la construction de terminologies, *Revue Française de Linguistique Appliquée*, Corpus : état des lieux et perspectives. Vol.XII-1. p. 39-52.

GRABAR N., HAMMON T. (2004), Les relations dans les terminologies structurées : de la théorie à la pratique, *Revue d'Intelligence Artificielle (*RIA), 18(1), Paris : Hermès, p. 57-85.

HEARST, M.A. (1992) Automatic acquisition of hyponyms from large text corpora. *Proc. of the 14th conference on Computational linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, 539–545.

MORIN E. (1999), Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques, *Traitement Automatique des Langues*, 40-1, 143-166.

SABOU M., D'ACQUIN M., MOTTA E. (2008), SCARLET: SemantiC RelAtion DiscoveRy by Harvesting OnLinE OnTologies, in The Semantic Web: Research and Applications Proceedings of ISWC 2008, Springer : Berlin / Heidelberg, LNCS Vol. 5021, 854-858.

SCHUTZ A., BUITELAAR P. (2005), RelExt: A tool for relation extraction from text in ontology extension, Gil,Y., MOTTA, E., BENJAMINS V.R., MUSEN, M., (eds), *The Semantic Web – Proceedings of ISWC 2005: 4th International Semantic Web Conference*, Galway, Ireland, Berlin: Springer Verlag, LNAI 3729, p. 593-606.