Métaphores : extraction automatique de lexiques métaphoriques à partir de corpus

Résumé

Le traitement de la métaphore en langage naturel pose encore de nombreux problèmes. Les précédents travaux dans le domaine se sont essentiellement situés au niveau sémantique, sans proposer d'outil de détection robuste de cette figure.
Dans ce papier, nous exposons les résultats d'une étude des indices de surfaces liés à la métaphore. Certains de ces indices, des marqueurs textuels composés d'un marqueur lexical et d'une structure syntaxique l'accompagnant, permettent en effet non seulement de détecter la présence d'une éventuelle métaphore avant tout traitement sémantique, mais encore de préciser, au niveau phrastique, quels éléments interviennent dans la « source » et la « cible » de la figure.
Nous proposons une méthodologie pour l'extraction automatique de relations métaphoriques à partir de corpus thématiques fondée en partie sur ces indices. STK, un prototype de traitement de corpus mettant en œuvre cette méthode, est en cours de développement. Il permet actuellement le balisage de corpus, l'extraction de lexiques ainsi qu'une évaluation semi-automatique de la méthode proposée.

Abstract

In natural language understanding, metaphor processing is still problematic. The classical studies in this domain mainly focussed on the semantic level, without proposing any robust detection method to spot the words that bear the figure.
In this paper, we expose the results of an analysis concerning the surface regularities related to the metaphor. Indeed, some of them, basically textual clues including a lexical marker and a surrounding syntatic structure, allow us to find possible metaphors, as well as to spot the terms that are involved, at the sentence level, in the "source" and in the "target" of the figure.
We propose a methodology for automatically extracting metaphoric relations from large domain-specific corpora, using these clues. STK, a prototype implementing the method, is currently under development. It includes marking up, tokenizing and tagging of corpora, extraction of lexica, and is used for evaluating the method semi-automatically.

Métaphores : extraction automatique de lexiques métaphoriques à partir de corpus

Stéphane Ferrari

LIMSI-CNRS
BP 133
F-91403 Orsay cédex, France

1. Introduction

La métaphore linguistique est une figure de style reflétant un processus cognitif pouvant s'appliquer à d'autres domaines d'expression que celui de la langue, écrite ou parlée. Ce processus consiste en quelque sorte à mettre en relation deux objets, l'un, la « source », exprimant un aspect de l'autre, la « cible ».

Les précédentes études sur les métaphores en traitement automatique de la langue, présentées en section 2., se sont essentiellement orientées vers l'étude de régularités sémantiques afin de permettre une interprétation de cette figure. Elles offrent de ce fait différents processus de traitement sémantique, inspirés de la nature de la relation introduite entre la source et la cible de la métaphore, et aboutissant sur une ou plusieurs représentations du sens au niveau de la phrase. Cependant, de telles approches éludent ou complexifient le problème de la détection de la métaphore. C'est pourquoi nous proposons d'étudier dans quelle mesure une analyse précédant le traitement sémantique peut faciliter la mise en œuvre des processus de constructions de représentation du sens issus des travaux précédents.

L'analyse manuelle d'un corpus explicatif nous a permis de mettre en évidence l'existence de régularités dans l'énonciation des métaphores en langue française. Nous en exposons brièvement les résultats dans la section 3.. Ces régularités sont caractérisables par l'emploi de termes lexicaux spécifiques auxquels peuvent être associées des structures syntaxiques récurrentes. L'adjonction d'une description syntaxique à un marqueur lexical constitue ce que nous nommons un *marqueur textuel*, permettant le repérage des éléments de la phrase impliqués dans la relation métaphorique : source et cible.

Les régularités issues de cette première analyse ont été généralisées, selon un protocole que nous

décrirons, et l'ensemble des marqueurs textuels en résultant font l'objet d'une évaluation sur un corpus électronique de grande taille, de manière à déterminer leur pertinence. Nous proposons en conséquence, section 4., une méthode pour l'extraction automatique de lexiques métaphoriques en partie fondée sur le marquage textuel[1]. Nous décrirons un prototype, STK, qui implémente la méthode. STK permet notamment le balisage de corpus de grande taille, leurs segmentation et étiquetage en vue d'une extraction de lexiques spécialisés. Il est en outre actuellement utilisé pour la validation de la méthodologie proposé.

2. Approches classiques du T.A.L.

Les précédents travaux en Traitement Automatique de la Langue (T.A.L.) consacrés à l'étude des métaphores se situent pour la plupart d'entre eux au niveau sémantique. En effet, leur principal objet est la recherche d'une représentation satisfaisante du sens des phrases dans lesquelles apparaît une métaphore. De ce fait, comme nous l'avons précédemment exposé [Ferrari, 1996a], ces travaux ont essentiellement étudié la nature de la relation entre la « source » et la « cible » d'une métaphore, de manière à pouvoir en proposer une interprétation en contexte. En reprenant la classification proposée par Dan Fass [Fass, 1991], les approches par comparaison (*comparison view*), interaction (*interaction view*) ou anomalie (*selection restrictions violations view*), mettant respectivement l'accent sur les aspects analogiques, nouveaux et anormaux de la relation métaphorique, semblent ignorer la difficulté qui réside dans la détection des termes métaphoriques. Le sens métaphorique y est systématiquement opposée à la notion de sens « littéral », lequel est supposé représenté à la suite d'un traitement sémantique générique dont le principe fondamental est d'associer un concept à une entrée lexicale.

L'incohérence au niveau phrastique d'une représentation du sens *littéral* constitue cependant un indice peu robuste de détection d'un écart de sens de type métaphorique. Ainsi, il arrive qu'une telle représentation reste cohérente malgré la présence d'une métaphore, comme dans l'exemple :

(1) Mc Enroe killed$_S$ Connors (*Mc Enroe a* achevé *Connors*)[2]

1. Ces travaux s'inscrivent dans une action de recherche concertée financée par l'AUPELF-UREF, Association des Universités Partiellement ou Entièrement de Langue Française, sur l'ingénierie de la langue, « Linguistique, Informatique et Corpus écrits », dans le thème « Construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus »

2. dans [Martin, 1992]

En outre, il se peut qu'une « anomalie » soit due à un autre type d'écart qu'une métaphore, comme dans :

(2) Il a terminé Zola et commence Stendhal pour une autre fiche de lecture.

D'autres types d'approches ont considéré la possibilité de traitements sémantiques multiples et concurrents. Ainsi, Fass envisage une recherche de sens métonymiques ou métaphoriques en même temps que celle d'un sens littéral [Fass, 1991], mais n'indique pas comment choisir parmi les multiples représentations possibles du sens celle qui convient le mieux au contexte. James Martin propose un système prenant en considération des connaissances spécifiques relatives à des métaphores conventionnelles [Martin, 1992]. Il est alors possible de considérer certains emplois métaphoriques communs au même niveau que le sens littéral. Une méthodologie de désambiguïsation permet en outre de décider quelle interprétation choisir. Il reste malgré tout que des métaphores nouvelles peuvent être ignorées par une telle approche. De plus, la base de connaissance nécessaire à la mise en œuvre de ce système ne paraît pas couvrir l'ensemble des cas possibles. Le modèle s'inspire en effet des travaux de Georges Lakoff [Lakoff et Johnson, 1980], répertoriant manuellement ce type de métaphores. Une automatisation de cette recherche paraît ici nécessaire si l'on veut obtenir une couverture intéressante.

Dans le but d'améliorer les approches précédentes, nous proposons d'étudier dans quelle mesure des analyses ne se situant plus uniquement au niveau sémantique peuvent fournir des indices soit pour la détection des métaphores, soit pour l'extraction de termes métaphoriques, soit encore pour la désambiguïsation dans le cas de multiples représentations du sens.

3. Indices de surface liés aux métaphores

La notion d'indices de surfaces, telle que l'introduit Pery-Woodley [Pery-Woodley, 1990], concerne des régularités d'énonciation pouvant être liées à un type de discours, à un style linguistique ou cognitif sous-jacent, ou encore à un auteur particulier. L'existence de tels indices a été prouvée dans le cadre du discours explicatif, et utilisé par exemple pour la modélisation du profil de l'utilisateur dans le cadre d'un système d'enseignement par ordinateur [Daniel et al., 1992].

Dans le cadre d'une collaboration avec des psychologues de l'équipe « Cognition Humaine »

du LIMSI-CNRS[3], nous avons recueillis un corpus de 26 textes explicatifs, d'en moyenne 200 mots chacun, de manière à étudier plus particulièrement les éventuels indices de surfaces liés à la métaphore et à l'analogie [Prince et Ferrari, 1996]. Les régularités que nous avons pu trouvé dans ce corpus sont de différentes natures : marqueurs lexicaux et structures syntaxiques récurrentes, introduisant une métaphore au niveau phrastique, et extensions de métaphores conceptuelles[4], sur un paragraphe ou un texte, à différents termes. Nous considérons plus particulièrement dans cet article le premier type d'indices. La présence d'un marqueur lexical ou d'une structure syntaxique constitue en effet un indice facilement repérable par une analyse ne faisant pas intervenir de connaissances d'ordre sémantique. Les marqueurs lexicaux relevés dans le corpus peuvent être regroupés en fonction de leur capacité à introduire soit une comparaison, soit une emphase. La catégorie grammaticale de ces marqueurs influence la structure syntaxique des phrases dans lesquelles ils apparaissent, permettant ainsi d'identifier quels sont les éléments comparés et comparants. La plupart des marqueurs que nous avons relevés fonctionnent selon ce principe, les comparaisons implicites pouvant être liées soit à une véritable structure comparative (ex. 3), soit à une structure d'identification (ex. 4) ou d'opposition (ex. 5)

(3) *comparaison via comme* **Comme**$_{LM}$ le tracteur$_S$ est un outil qui permet de réaliser plus rapidement ce que l'homme faisait pour cultiver et traiter la terre, l'ordinateur^C est un outil qui permet de traiter des informations que l'homme [lui] donne [...].

(4) *identification via apposition + emphase typographique* Un ordinateur [...] est généralement composé [...] d'une mémoire, et d'un processeur^C,$_{LM}$ le «$_{LM}$ cerveau$_S$ »$_{LM}$ de la machine.

(5) *opposition via verbe d'état* c'est-à-dire un cerveau^C qui **ne serait pas**$_{LM}$ humain$_S$ car incapable d'effectuer une tâche pour laquelle il n'est pas conçu.

Il devient alors possible de déterminer, par la syntaxe accompagnant ce type d'indice, la source$_S$ et la cible^C de la métaphore. Nous proposons de ce fait une représentation informatique orientée objet des marqueurs textuels de manière à rendre compte de leurs propriétés[5]. Les marqueurs emphatiques ne permettent pas quant à eux la détection de la cible de la métaphore. Il s'agit principalement d'averbes ou adjectifs (*littéralement*, *véritable*) pouvant apparaître conjointement avec l'un des précédents types (ex. 4).

3. Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
4. au sens donné par Lakoff [Lakoff et Johnson, 1980]
5. Représentation détaillée dans [Ferrari, 1996b]

4. Extraction de lexiques métaphoriques

4.1. Prérequis

Les marqueurs issus de la première analyse permettent de repérer, au niveau phrastique, un ensemble de termes pouvant être employés de manière métaphorique. Cependant, leur usage ne garantie en aucun cas la présence d'une métaphore, et ces marqueurs n'ont de plus pas une couverture optimale. Afin de pouvoir automatiser l'extraction de lexiques métaphoriques à l'aide de marqueurs, il est donc nécessaire, d'une part, de généraliser les marqueurs afin d'optimiser la couverture, d'autre part, de valider leur pertinence en tant que marqueur de métaphores, voire de leur attribuer une valeur de qualité pouvant servir de probabilité de présence d'une métaphore, et d'aide à une désambiguïsation du sens.

Nous avons dans un premier temps généralisé de deux manières les marqueurs lexicaux : synonymes ou termes de sens proche dans une même catégorie grammaticale (ex. : *véritable* donne *vrai*, *réel* et *littéral*), formes fléchies issues d'une même racine terminologique (ex. : *littéral* donne *littéralement*). Chaque marqueur résultant est en outre actuellement évalué sur un corpus d'environ 450 000 mots de manière à établir sa pertinence en tant que marqueur de métaphore. Ainsi, *littéralement* y est employé 36 fois, dont 35 en présence d'une métaphore (éventuellement lexicalisée). Le contre-exemple concerne l'explication du sigle SICAV, ce qui nous a permis de considérer le marqueur textuel (incluant la description syntaxique) afin d'obtenir une pertinence plus fine, fonction de la catégorie de la source : ici, *littéralement* est d'une meilleure qualité pour l'introduction de métaphores verbales que pour celles nominales.

L'extraction proprement dite ne nécessite plus à ce stade que la mise en place des outils adéquats de traitement de corpus. Un prototype, nommé STK, est actuellement en cours de développement selon une approche incrémentale : segmentation de texte, étiquetage[6], balisage[7], extraction de lexique. Le prototype actuel n'incluant pas d'analyseur syntaxique, les marqueurs textuels sont repérés par leur attribut *marqueur lexical*, puis les termes co-occurrents sont extraits en fonction de leur catégorie grammaticale. Ainsi, par exemple, lorsque l'adverbe *littéralement* est trouvé, si un verbe le suit ou le précède, il est extrait, sinon, le nom ou groupe de noms consécutif. Cette phase expérimentale devrait être améliorée à l'avenir par l'utilisation d'une analyse syntaxique des phrases, totale ou partielle, les résultats étant fortement bruités.

6. assignation de catégories grammaticales
7. repérage des unités logiques *textes, paragraphes*, etc...

4.2. Ressources et contraintes

Le prototype STK exploite des ressources déjà existantes en plus de ressources propres. Ainsi, l'étiquetage se fait par appel à l'étiqueteur de Brill [Brill, 1992] spécialement entraîné pour le français. Le lexique utilisé est issu de BDLEX, développé par G. Perennou et M. De Calmès à l'IRIT, Toulouse, France. Il contient environ 23 000 formes canoniques résultant en 250 000 formes fléchies.

L'adaptation de l'étiqueteur de Brill au français a nécessité un formatage spécifique du lexique : à chaque forme fléchie doit correspondre la liste des catégories possibles, avec en tête de liste la plus fréquemment utilisée. Cette dernière information n'apparaissant pas dans le lexique initial, nous avons mis un ordre sur les étiquettes, fonction de la catégorie, et corrigeons actuellement les erreurs résiduelles à la main, après analyse des résultats de l'étiquetage. Ainsi, à l'entrée *son* correspond la liste ordonnée *(Déterminant Masculin Singulier, Nom Masculin Singulier)* (le trait *adjectif possessif* n'étant pas exploité ici).

Une contrainte supplémentaire est imposée au format des textes issus de la phase de segmentation : une phrase par ligne. Nous avons de ce fait développé une procédure spécifique de recherche de l'unité phrastique n'exploitant pas de ressource autre que le texte brut.

Le corpus actuellement utilisé est issu du journal « Le Monde sur CD-ROM ». Il est constitué d'articles d'économie, d'environ 450 000 mots au total. Nous l'utilisons actuellement pour déterminer les valeurs de qualité —ou pertinence— des différents marqueurs textuels.

5. Conclusion, perspectives

Nous avons proposé une méthode de repérage des métaphores n'exploitant pas de connaissances sémantiques, facilement implémentable pour le traitement de textes de grande taille (de type corpus). L'existence de marqueurs textuels a été démontré par une première analyse sur un petit corpus. Ces marqueurs ont été généralisés et sont désormais évalués sur un corpus de grande taille, de manière à établir leur pertinence.

Les outils nécessaires à l'extraction de lexiques métaphoriques à partir de corpus sont en cours de développement. Un prototype, STK, permet déjà le balisage, la segmentation et l'étiquetage de gros corpus. Les premiers résultats nous ont permis d'étendre le jeu de marqueurs initial.

Nous nous engageons maintenant dans la mise en œuvre d'une analyse syntaxique suffisante pour un repérage optimal des termes métaphoriques co-occurrents avec les marqueurs.

Références

Brill, E. (1992). A simple rule-based part of speech tagger. Dans *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento. ACL.

Daniel, M., Nicaud, L., Prince, V. et Pery-Woodley, M. (1992). Apport du style Linguistique à la Modélisation Cognitive de l'Élève. *Lecture Notes in Computer Sciences*, 608:252–260. Proceedings of the International Conference on Intelligent Tutoring Systems (ITS-92), Montréal.

Fass, D. (1991). met : A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics*, 17(1):49–90.

Ferrari, S. (1996a). A corpus-based approach for metaphor processing. Dans *Proceedings of the AISB96 2nd Tutorial and Workshop Series - LEDAR Workshop.* University of Sussex, Brighton, UK. 31 March - 2 April 1996.

Ferrari, S. (1996b). Using textual clues to improve metaphor processing. Dans *Proceedings of the Student Sessions at the 34th Annual Meeting of the Association for Computational Linguistics.* Santa Cruz, CA, USA. June 23-28 1996. to be published.

Lakoff, G. et Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press, Chicago, U.S.A.

Martin, J. (1992). Computer Understanding of Conventional Metaphoric Language. *Cognitive Science*, 16:233–270.

Pery-Woodley, M. (1990). Textual clues for user modeling in an intelligent tutoring system. Mémoire de maîtrise, University of Manchester, England, U.K.

Prince, V. et Ferrari, S. (1996). A textual clues approach for generating metaphors as explanations by an intelligent tutoring system. Dans *Proceedings of TALC96, Conference on Teaching And Language Corpora.* Lancaster University, U.K., 9-12th August. to be published.