

Une proposition pour l'extraction de relations non prédicatives

Mouna Kamel

IRIT-Université Paul Sabatier-31069 Toulouse Cedex
kamel@irit.fr

Résumé. Les relations sémantiques généralement reconnues par les méthodes d'extraction sont portées par des structures de type prédicats-arguments. Or, l'information recherchée est souvent répartie sur plusieurs phrases. Pour détecter ces relations dites complexes, nous proposons un modèle de représentation des connaissances basé sur les graphes conceptuels.

1 Problématique

L'Extraction d'Information consiste à identifier de l'information bien précise d'un texte en langue naturelle et à la représenter sous forme structurée. Les composantes de l'information recherchée sont généralement prédéfinies et circonscrites à un domaine spécifique, et les principaux travaux réalisés en matière d'identification de relations sémantiques ont essentiellement concerné les relations portées par une structure de type prédicats-arguments. Les principales approches d'identification de ces relations ont été basées sur l'analyse syntaxique (identification du verbe et ses arguments) (Khélif, 2006), ou sur la définition de patrons lexi-co-syntaxiques (Aussenac et al., 2000).

L'étude de corpus de domaines différents montre que bonne partie de l'information pertinente peut aussi être distribuée sur plusieurs phrases, par le biais de relations exprimées à l'aide de variations linguistiques, comme la coréférence, l'anaphore ou l'ellipse. Les méthodes classiques d'extraction de relations ne sont alors plus adaptées.

La résolution de relations non prédicatives, et plus particulièrement d'une certaine forme elliptique (formes passives où l'argument agent est effacé) utilisée fréquemment, nous a conduits à proposer une représentation des connaissances du domaine considéré, à l'aide du modèle des graphes conceptuels, car ce modèle est doté d'opérations et offre des procédures de raisonnement (Salvat, 1997).

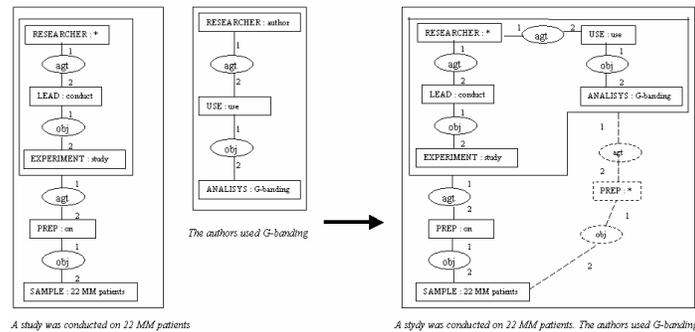
2 Identification de relations non prédicatives

Nous avons étudié un corpus de résumés d'articles scientifiques décrivant des expériences génétiques menées par des chercheurs sur un ensemble de patients porteurs d'une même maladie génétique, le but étant de localiser les régions chromosomiques affectées. Une des relations pertinentes identifiées est *Conditions Expérimentales* qui relie l'ensemble des patients observés au type d'analyse subie. Dans l'exemple "A study was conducted on 22 MM patients. The authors used G-banding", la relation *Conditions Expérimentales* ne peut être détectée par une approche classique. La mise en œuvre d'une procédure de raisonnement qui établit un lien entre les auteurs (*authors*) et l'étude (*study*) menée par les auteurs (*authors*)

Extraction de relations non prédicatives

permet d'inférer le lien existant entre l'échantillon (22 MM patients) et l'analyse (G-banding).

Chacune des phrases est traduite à l'aide d'un graphe dit de référence (GRef) qui modélise le domaine de l'étude et auquel des règles d'inférence ont été associées :



Pour cet exemple, l'opération de jointure entre les deux graphes fournit un graphe résultat, auquel sont appliquées les règles de déduction.

3 Conclusion

Ce travail, qui dans un premier temps est dédié à une forme d'ellipse particulière, doit être étendu aux relations exprimant des liens de cohésion et de cohérence entre différents éléments du texte, pour une meilleure compréhension du texte.

Références

- Aussenac N. et P. Séguéla (2000). *Les relations sémantiques : du linguistique au formel*. Cahiers de grammaire, Numéro spécial sur la linguistique de corpus. A. CONDAMINES (Ed.). Toulouse : Presse de l'UTM. 25 175-198.
- Khelif K. (2006). *Web sémantique et mémoire d'expériences pour l'analyse du transcriptome*. Thèse de doctorat dirigée par Rose Dieng, Université de Nice –Sophia Antipolis/INRIA.
- Salvat E. (1997). *Raisonnement avec des opérations de graphes : graphes conceptuels et règles de référence*. Thèse de Doctorat, Université des Sciences et Techniques du Languedoc.

Summary

Semantic relations generally recognized by information extraction tasks are of a predicative form. The useful information is often distributed over several sentences. To detect this kind of complex relations, we propose a knowledge representation based on the conceptual graph model.