Ressource MAR-REL (MARqueurs de RELations)

MAR-REL est une liste candidats-marqueurs français pour trois relations : l'hyperonymie, la méronymie et la cause, dont la précision est évaluée sur des corpus variant du point de vue du domaine et du genre textuel.

DESCRIPTION

La ressource MAR-REL a été réalisée dans le cadre du projet ANR Contint CRISTAL (Contextes Riches en connaissances pour la TrAduction terminoLogique,). Projet ANR- 12-CORD-0020 dont les partenaires étaient le LINA (-Laboratoire d'Informatique de Nantes), le laboratoire CLLE-ERSS (Cognition, Langues, Langage, Ergonomie-Equipe de Recherche en Syntaxe et Sémantique), la Société Linga et Machina et la Faculté de Traduction et d'Interprétation de l'Université de Genève.

Un des objectifs du projet a consisté à recenser un certain nombre de marqueurs-candidats de relations conceptuelles en français et à évaluer leur **précision** dans des corpus variant en fonction du domaine et du genre textuel (voir ci-dessous).

Les marqueurs-candidats étudiés concernent trois relations : l'hyperonymie, la méronymie et la cause.

Dans cette ressource, la précision est le rapport entre le nombre de contextes dans lesquels les marqueurs-candidats apparaissent et le nombre de contextes dans lesquels la relation recherchée est bien avérée. Elle correspond au calcul suivant :

$$Précision = \left(\frac{\text{Nombre d'occurrences des marqueurs−candidats dans lesquelles la relation est présente}}{\text{Nombre d'occurrences total des marqueurs candidats}}\right) \times 100$$

Par marqueur-candidat, nous entendons des éléments lexicaux, syntaxico-sémantiques ou typographiques qui, hors contexte, sont susceptibles d'être interprétés avec le sens d'une relation déterminée (dans notre cas, hyperonymie, méronymie ou cause).

ELABORATION DE LA RESSOURCE (LABORATOIRE CLLE-ERSS, UMR 5263)

Le recensement des marqueurs-candidats a été fait à partir de travaux existants et par ajout d'autres marqueurs (recours à l'introspection ou identification en corpus). Au total, 37 marqueurs-candidats d'hyperonymie, 99 marqueurs-candidats de méronymie et 316 marqueurs-candidats de cause ont été recensés.

La présence et la capacité de chaque structure-candidate recensée à « marquer » la relation attendue a été évaluée sur des corpus variant de deux points de vue :

- le domaine : cancer du sein vs volcanologie
- le genre : scientifique vs vulgarisation

soit quatre sous-corpus de taille à peu près similaire.

Chaque sous-corpus traitant du cancer du sein comporte environ 200 000 mots, tandis que les sous-corpus traitant de la volcanologie comportent environ 400 000 mots.

	Cancer du sein	**Volcanologie**
Corpus scientifique	Environ 200 000 mots	Environ 400 000 mots
	2002 – 2008	1980 - 2002
Corpus vulgarisé	Environ 200 000 mots	Environ 400 000 mots
	2001 - 2008	1980 - 2012

PRESENTATION DES RESULTATS

Les résultats sont présentés dans trois fichiers : un fichier qui liste l'ensemble des candidats marqueurs, un fichier qui donne la traduction de chaque candidat-marqueur en format UIMA (Unstructured Information Management Applications ; réalisée par la société Lingua et Machina), enfin un tableau qui concerne l'évaluation de la précision de chaque candidat-marqueur dans chacun des quatre sous-corpus.

1) Liste de tous les candidats-marqueurs

Les marqueurs-candidats sont présentés par sous-types des trois relations (hyperonymie, méronymie, cause). Un code d'identification alpha-numérique est attribué à chaque candidat-marqueur. Ce même code est repris dans les autres fichiers.

2) Traduction UIMA (réalisation : Kevin Coustot, *Société Lingua et Machina*, Paris)

Le codage linguistique est à interpréter de la façon suivante :

DET : N'importe quel déterminant,

Det défini : déterminant défini : le/la/les

Dét Indéfini : déterminant indéfini : Un/une/de

Adj : adjectif

Adv : Adverbe

X : premier élément de la relation (soit pour les trois relations, générique, holonyme, élément causateur). Il se réalise linguistiquement sous la forme d'un nom ou d'un goupe nominal.

Y second élément de la relation (soit, pour les trois relations, spécifique, méronyme, effet de la cause). Il se réalise linguistiquement sous la forme d'un nom ou d'un goupe nominal.

3) Tableau Excel (réalisation : Luce Lefeuvre, Laboratoire CLLE, Toulouse)

L'entrée dans le tableau se fait par type de relations voire de sous-relations. Pour chaque sous-relation, les marqueurs-candidats sont listés et chaque marqueur-candidat se voit attribuer sa précision en fonction des deux paramètres : domaine (cancer du sein vs vulcanologie) et genre (scientifique vs vulgarisation).

RESPONSABLE RESSOURCE

Anne Condamines (anne.condamines.univ-tlse2.fr)

REALISATION

Luce Lefeuvre (CLLE-ERSS), Kevin Coustot (Lingua et Machina), Anne Condamines (CLLE-ERSS), Josette Rebeyrolle (CLLE-ERSS).

La constitution des candidats-marqueurs et l'analyse de leur fonctionnement en corpus a fait l'objet de la thèse de Luce Lefeuvre, dirigée par A. Condamines et J. Rebeyrolle :

Analyse des marqueurs de relations conceptuelles en corpus spécialisé : recensement, évaluation et caractérisation en fonction du domaine et du genre textuel,

dont la soutenance est prévue le 5 septembre 2017.