Semantic relations: new challenges in a world of linked data

Nathalie Aussenac-Gilles

Keynote speech

Nathalie Aussenac-Gilles

http://www.irit.fr/~Nathalie.Aussenac-Gilles/index_en.php

Semantic relations: new challenges in a world of linked data

Today, connecting two entities with a labelled, meaningful relation is a task that is required in such a large variety of situations that there are thousands of papers, research works and approaches about relation identification. Depending on the application using or producing "relations" or "links", this notion covers actually different realities:
from a natural language phrase to a formal formula, from a relationship between 2 entities to one between classes, from domain specific to universal relations, from contextual and possible relations to validated and proved links ... Terminology is one of these research and application domains where relations play a major role to structure and organise language. They contribute to account for meaning and illustrate the difficulty to fix a dynamic and moving reality in a static model.

In my talk, I shall first draw an overview of these diverse realities behind the notion of relation in research areas close to linguistics, terminology, knowledge modelling, natural language processing and semantic web. Then I shall connect these definitions with tools, methods and approaches to identify these relations. I shall finally comment how terminology can play a key role in this context, either to change its practices by integrating results and tools from other research fields, or by contributing to improve these approaches with the studies carried out during the last 20 years to build terminologies and terminological knowledge bases.

Detection of the Disease-Symptom Relation in Medical Texts in Catalan: a Corpus-Based Approach

Experts in terminology state that in specialized texts there are many terms connected among them, which form nodes and are distributed in conceptual structures. Following this idea, the study of conceptual and semantic relations in specialized texts is relevant. Specifically, the automatic or semiautomatic detection of these relations in real texts is a great challenge. In medicine, this challenge is especially important, since currently the developing of this kind of systems is useful for automatic construction and updating of ontologies, thesaurus, and definitions included in medical dictionaries and specialized manuals.

To carry out this kind of research, two main strategies are used: strategies based on linguistic patterns (e.g. Feliu, 2004) and machine learning strategies (e.g. Hearts, 1999; Riloff & Jones, 1999; Turney, 2006). Machine learning strategies are useful when a big annotated corpus is available. In Catalan there is a lack of some Natural Language Processing (NLP) tools. Specifically, there is not any annotated corpus with conceptual relations among terms and any system for detecting conceptual relations. The only work about this subject is the research by Feliu (2004), where she describes several linguistic patterns that would allow detecting some specific conceptual relations in specialized texts in Catalan.

In this context, our work has two main goals:

- to elaborate a list of linguistic patterns that show the disease-symptom relation between terms extracted from medical texts in Catalan,
- to propose a methodology for the automatic creation of these linguistic patterns.

The methodology that we propose consists of several stages. First, to select corpus including specialized texts in Catalan from the medical domain. Second, to choose some specific diseases to work with. Third, to extract contexts automatically from the corpus. Fourth, to analyse manually contexts including terminological units that express symptoms, in order to find explicit linguistic marks showing the disease-symptom relation. Fifth, by using these linguistic marks, to produce a list of general and lemmatized linguistic patterns. This list could be used by an automatic or semiautomatic system to detect this specific relation among terms in medical texts. In our research, we use the medical subcorpus of the IULA Technical Corpus and the online interface BwanaNet (Vivaldi, 2009) to extract the contexts automatically. Also, in our work, neurodegenerative diseases are chosen, taking into account that the study of their symptoms has a social interest, since their early diagnosis is nowadays necessary.

This research has obtained three main results:

- A list of 24 linguistic patterns that are used in medical texts in Catalan to express the disease-symptom relation are obtained, by using the proposed methodology.
- 31 different terms expressing symptoms of neurodegenerative diseases have been found in our corpus. From them, only the 22% appears in the definitions of these diseases in specialised dictionaries and manuals. For example, the term *estrès oxidatiu* ["oxidative stress"] is a symptom of the disease *esclerosi lateral amiotròfica* ["amyotrophic lateral sclerosis"]. We have found this term in our corpus, but it is not possible to find it in the definitions of *esclerosi lateral amiotròfica* included in the consulted terminological resources. This means that Specialized Texts Processing (STP) is an important tool for updating medical resources.
- Although in the terminology field it is stated that terms are the prototypical units that show specialized knowledge in texts, we have found that, in our corpus, symptoms are not only expressed by means of monolexical and polilexical terms, but also by means of specialized phraseology or collocations. For example, the disease *Alzheimer* ["Alzheimer disease"] is associated with the symptoms *estrès oxidatiu* ["oxidative stress"] (a prototypical terminological unit), but also with *aparició de dipòsits de plaques amiloides* ["appearance of beta-amyloid deposits"] (a specialized collocation formed by the term *plaques amiloides* and the collocative *aparició de dipòsits de.* Our research shows that, in our corpus, the 42% of symptoms are expressed by terms and the 58% by collocations.

The results obtained in this work are promising, because the list of generated linguistic patterns will be the basis for the implementation of a system to detect the disease-symptom relation. Also, the proposed methodology will allow us to generate more lists of patterns for the detection of other different conceptual relations in the medical field. Moreover, results are interesting because they help to understand how medical information is shown in specialized texts, regarding terms expressing new symptoms in texts, and symptoms expressed by collocations and not only by prototypical terminological units.

References

Feliu, J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [PhD. Thesis]

Hearst, M. A. (1992). «Automatic acquisition of hyponyms from large text corpora». In *Proceedings of COLING*. 539-545.

IULA Technical Corpus [on-line]. Barcelona: Institut Universitari de Lingüística Aplicada. http://bwananet.iula.upf.edu/

Riloff, E.; Jones, R. (1999). «Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping». In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*. 474-479

Turney, P.D. (2006). «Similarity of semantic relations». In *Computational Linguistics*, 32(3). 379-416.

Vivaldi, J. (2009). «Corpus and exploitation tool: IULACT and bwanaNet». In Cantos, Pascual; Sánchez, Aquilino (ed.). *A survey on corpus-based research.* Murcia: Asociación Española de Lingüística de Corpus. 224-239.