# UniGe at CLEF 2009 Robust WSD Task

**Jacques Guyot, Gilles Falquet, Saïd Radhouani**

**Computer Science Center, University of Geneva - Route de Drize 7, 1227 Carouge, Switzerland**

**{Jacques.Guyot, Gilles.Falquet, Said.Radhouani}@unige.ch**

For our second participation to the Robust Word Sense Disambiguation (WSD) Task, we focused on performing a deep analysis of the ambiguity issue in the field of Information Retrieval. During the 2008 edition, we noted that although the WSD corpus allowed lifting **lexical ambiguities**, our results based on the corpus' WSD were not clearly better than those based on words only. We showed that lexical ambiguity was an issue only when queries included only one or possibly two words, but whenever the query was "longer", its words created a context that implicitly decreased lexical ambiguities. We thought we had a **domain ambiguity** problem, *i.e.* the retrieved documents did contain some of the query's words but they turned out to be irrelevant. Thus, we tried to expand the query's vocabulary in the following way:

1) On the basis of the query's titles, we queried the Web (using Google's Search Engine) and selected the 50 top retrieved documents;
2) We downloaded those documents and kept only the text;
3) Then we trained a supervised classifier by associating the document classes to the query numbers;
4) Finally we extracted the 50 most classifying words for each query.

Thus, for each query, we produced a list of related words; for example, in query148, which deals about the hole in the ozone layer, we got the following words:

"ozone layer stratosphere cfcs ultraviolet depleting chlorofluorocarbons depletion atmosphere antarctic chlorine montreal chemicals cataracts stratospheric atmospheric rays gases hole deplete molecules damaging harmful compounds substances protocol phased cfc aerosols antarctica troposphere climate hcfc nitrogen molecule protective volcanic bromide halons temperatures arctic hydrochlorofluorocar warming bromine conditioners dioxide thinning refrigerants wmo volcanoes."

This word list was used (after withdrawing the words in the query's title) to re-formulate a query (with an OR operator between the words) on the CLEF corpus. We did not apply any "feedback relevance" method on the results. The answers (called DOM) contained about 60% of correct documents but the average precision fell to a level slightly above 10%. Then we ranked again those retrieved documents by directly querying the CLEF query on the corpus. The documents included in the DOM answers were promoted to the top of the list while the other ones were pushed down. This new ranking was meant to eliminate the documents which are outside the domain, thus improving the precision. The results showed that this process had virtually no impact, as almost all the documents on top of the list belonged to the query domain. Therefore the hypothesis of a domain ambiguity must probably be rejected. In another experience (WEB), we used the list of words which define the domain to expand the CLEF query. This had a negative impact on the precision: the additional words seem to "dilute" the original question.

For an easier analysis of the answers, we converted the results into hypertext, thus allowing for a quick access to the text of the referenced documents. A detailed analysis of the answers showed that the ambiguity was in fact of semantic nature. The "right" or "wrong" documents were not differentiated by the words they contained: both included words from the query and from the domain. Thus the WHAT aspect (the topic) was equivalent. However, the HOW aspect (how people talked about the topic) was different and required a semantic "understanding" of the text. For instance, in the query dealing with the fourth victory of Indurain in the Tour de France (we were looking for documents relating to the *reactions* to this victory), all the answers were linked to the victory but some of them related to its anticipation while others were referring to it after it occurred. A human being can easily tell the "right" answers because of their experience of reactions to a victory in a bike

contest. Therefore, in order to significantly improve the performance, we believe the problem should be addressed with methods allowing to introduce semantic elements.