# Combining NLP and ML for Processing Sublanguages and Domain-Specific Corpora

Davy Weissenbacher[1]

## Abstract

The fact that from one domain to another, the language and textual structure vary is well acknowledged in corpus linguistics (Harris et al., 1989) but it is hardly taken into consideration in natural language processing (NLP). A corpus gives only local and partial view on the linguistic phenomena of a given domain, whereas those phenomena vary from a domain or type of text to another.

These variations have a strong impact on the performances of the NLP systems when they are based on linguistic rules. These systems are designed and tuned on a reference corpus and are often inadequate for processing different sublanguages. A solution consists in using machine learning (ML) to adapt the systems automatically (Evans, 2001) from a sample of the corpus to analyze. However the ML-based systems do not exploit rich linguistic rules but only surface clues (Clemente Litran J.C., 2004).

In this article we propose a probabilistic approach based on the formalism of the Bayesian Networks. This model offers a great expression capacity to integrate heterogeneous pieces of knowledge (deep or shallow) in a single representation as well as an elegant mechanism to adapt them to the domain or the kind of the corpus. This new approach, which is promising for NLP, will be exemplified on a specific NLP task (anaphora resolution).

We will study the robustness of our system w.r.t the corpus domain and type by analyzing its performances on four different corpora, respectively made of abstracts of genomic and management articles, technical manuals and journal articles. The performances of our system are better than those of the state of the art systems with which we have compared it.

## References

Satou K. and Torisawa K. Clemente Litran J.C. 2004. Improving the identification of non-anaphoric it using support vector machines. In Actes d'International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 58–61.

R. Evans. 2001. Applying machine learning toward an automatic classification of it. Literary and linguistic computing, 16:45–57.

Zellig Harris, M. Gottfried, T. Ryckman, Jr P. Mattick, Anne Daladier, T.N. Harris, and S. Harris. 1989. The Form of Information in Science, Analysis of Immunology Sublanguage, volume 104. Kluwer Academic Publisher.

[1] Laboratoire d'Informatique de Paris-Nord, Universite Paris-Nord, Villetaneuse, France
*e-mail*: davy.weissenbacher@lipn.univ-paris13.fr