

A Combination of Powerful Tools for Entity Extraction in Specific Domains (e.g. Biology)

Extracting meaningful content from very large quantities of unstructured text is an ever growing need. It calls for advanced computational text analysis.

Entity extractors (e.g. for company names, person names, locations, dates) are now commercially available in several languages (including Chinese and Japanese) as a result of previous R&D conducted at Xerox Research Centre Europe (XRCE). However, given their complexity, rapid evolution and the lack of naming standards, some domain-specific types of entities (see Figure 1) require more sophisticated tools.

An examination of *Drosophila melanogaster* from natural populations revealed genetic variation for **dipeptidase-A (DIP-A)** and **dipeptidase-B (DIP-B)** activities within sets of lines that differed only in the second or the third chromosome.

Fig. 1: Example of Entity Extraction in Biology

Xerox has developed a complete Entity Extraction solution for the Biology domain. The lexical resources included in the license consist of an English dictionary and a Biology terminology dictionary.

Besides biology, Xerox can develop entity extractors customized to any specific domain, language or customer.

Furthermore, Xerox offers its entity extractor development environment for licensing.

Xerox Entity Extraction is done through three complementary methods. All three leverage linguistic pre-processing of documents, which includes state-of-the-art tokenization, normalization, morphological analysis and Part-of-Speech tagging.

1. Manual Rules

Entity extraction rules are based on lexical, contextual and syntactic properties. Rules are managed through the Xerox *XIP* Parser which offers high computational performance and a superior rule formalism. As an example, Xerox licenses a set of more than 250 manual rules for Biological Entity Extraction, all of which have been validated by biology experts.

2. Rule Induction

In this method sets of rules are automatically generated and tested on annotated data by the Xerox *ALLiS* rule induction system. *ALLiS* learns rules that identify terms representing a biological entity. Only rules with precision above a set threshold are kept. For each precision target

a different set of rules with associated exceptions is induced.

Induced rules can directly work in conjunction with manual ones, or be manually edited, thus offering a very powerful tool to develop domain-specific entity extractors. Furthermore, because of the threshold on precision provided by *ALLiS*, the precision vs. recall trade-off of an entity extractor can be tuned according to the user's needs. This is definitely a key differentiator.

3. Statistical Classification

Entity extraction can also be handled as a classification problem: does a term represent a specific type of entity or not? Xerox combines various classification methods based on probabilistic models and large margin classifiers (e.g. Support Vector Machines), which brings key competitive advantages when annotated data is sparsely available. The idea behind combining several state-of-the-art supervised and unsupervised Machine Learning methods is to use a limited amount of annotated documents while leveraging the availability of large collections of unannotated data, such as PubMed which contains more than 10 million records of articles related to Biology.

A Very Powerful Combination

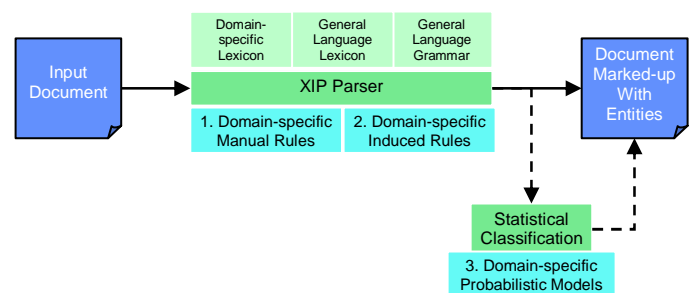


Fig. 2: Three Methods for Domain-specific Entity Extraction

By combining three complementary approaches to Entity Extraction in a modular way (see Figure 2), Xerox offers a breakthrough solution in the field of Information Extraction based on world-class underlying linguistic technologies.

Xerox offers its Entity Extraction tools for evaluation, research and commercial (incl. OEM) licensing. Xerox can help to select the best approach for any given application.