Chapter XII
Discourse Processing for Text Mining

Nadine Lucas
GREYC CNRS, Université de Caen Basse-Normandie Campus 2, France

ABSTRACT

This chapter presents the challenge of integrating knowledge at higher levels of discourse than the sentence, to avoid "missing the forest for the trees". Characterisation tasks aimed at filtering collections are introduced, showing use of the whole set of layout constituents from sentence to text body. Few text descriptors encapsulating knowledge on text properties are used for each granularity level. Text processing differs according to tasks, whether individual document mining or tagging small or large collections prior to information extraction. Very shallow and domain independent techniques are used to tag collections to save costs on sentence parsing and semantic manual annotation. This approach achieves satisfactory characterisation of text types, for example reviews versus clinical reports, or argumentation-type articles versus explanation-type. These collection filtering techniques are fit for a wider domain of biomedical literature than genomics.

INTRODUCTION

In this chapter we address higher-levels of text processing, as related to text mining. The domain of biomedical language processing (BLP or bio-NLP) "encompasses the many computational tools and methods that take human generated texts as input, generally applied to tasks such as information retrieval, document classification, information extraction, plagiarism detection, or literature-based discovery" (Hunter & Bretonnel Cohen, 2006 p. 589).

Access to biomedical literature itself (primary sources) is provided since 2004 through PubMed Central established by the American National Library of Medicine (NLM) as a repository of free access articles. The search system Entrez PubMed offers abstracts from Medline along with on line

full-text indexed by Mesh and access to databases (see NLM site). This has fostered a new circular situation where data and text bases feed literature in turn feeding databases and ontologies.

Related events are first, advances in genomics and the information deluge. A double exponential growth of published material is recorded in the biomedical field, creating in turn an increased amount of facts to be stored (Shatkay & Craven, 2007). Second, text mining techniques for specific purposes were developed in particular to help in database curation. Automats now directly fill a growing part of databases (Hunter & Bretonnel Cohen, 2006). Third, a new field called systems biology emerged at the frontier between data and text mining (Krallinger & Valencia, 2005). Text mining is used to back data interpretation. Computational processes are ubiquitous and the frontier between text and data mining is blurred as well as the frontier between human and automated processes. Integrated text mining systems inherit from expert systems (nomenclatures linked with inference rules) and from statistical data mining. They rely on what might be called tertiary sources of knowledge: unified nomenclatures, hand curated interaction databases and hand annotated corpora. These are sometimes grouped under the term "ontologies" (Ananiadou *et al*., 2006). Last, users now take it for granted that raw information is quickly translated into secondary and tertiary sources, and rely on computer-manageable "concepts" (Rebholz-Schumann *et al.*, 2005). Recent developments can be watched by consulting the Biomedical Literature Mining Publication portal (Blimp) (2008).

Success in the genomics field opened the way for less specific purposes. As text mining is advertised in more publications, not only "omics" researchers, but also clinicians, general practitioners and medical librarians call for text processing (Fluck *et al*., 2005; Hunter & Bretonnel Cohen, 2006; Mizuta *et al*., 2006). One emerging trend in research is to take patients into consideration to best respond to users' needs (Leroy *et al*., 2006). This implies a change in the way to produce results. While researchers can do with highly specialised words, evoking for them research trends, a wider public need full explanations, therefore lengthier passages of text. Robust text processing is needed but it is still in its infancy.

Another trend calls for semantic characterisation of texts in a collection. Most semantic oriented tasks, such as characterizing original findings on a topic, or eliciting hypotheses, require a wider context than the sentence. Valuable meta-information that could be used to qualify texts is still lacking. Some attempts at qualifying parts of them, like conclusions, e.g. tentative or definite are on the way. Yet, very few studies address text at a global level as a semantic unit. The gap between expectations and realisations is blatant.

The approach explained here relies on combining text mining characterisation techniques for collections and robust high-level discourse parsing techniques. We advocate a shift of paradigm from word-level description to text-level description. In the domain of biomedical text processing, text is characterized as "unstructured data" as compared to databases (Hunter & Bretonnel Cohen, 2006), or at best as semi-structured data (Hakenberg *et al.,* 2005). Yet, texts are structured by layout, a feature that has been overlooked. Academic articles in particular are highly structured. Scale issues are seldom addressed, although they are important for information retrieval and knowledge integration in biomedicine.

We consider layout structures in relation with rhetorical structures and discourse segments. A survey of the state of the art is provided in the background section. In the third section, original research on multi-scale text descriptors is thoroughly explained. Experiments for complex tasks in text mining are then introduced and future trends including evaluation is discussed in the following section. We conclude on these experiments and broaden perspectives in the last section.