Natural Language Processing in the Medical and Biological Domains: a Parallel Perspective

Pierre Zweigenbaum
LIMSI - CNRS
BP 133, F-91403 Orsay Cedex, France
http://www.limsi.fr/~pz/
pz@limsi.fr

Abstract

Natural Language Processing (NLP) has been active in the medical domain for more than thirty years, with pioneering projects such as the Linguistic String Project. 'BioNLP', the application of Natural Language Processing methods to the analysis of the biological literature in the genomics era, has undergone a fast development in little over ten years.[1] It rapidly attracted Medical NLP and Computational Linguistics researchers, especially through challenges and evaluation initiatives. We examine here to which extent medical NLP prepared the ground for BioNLP. Conversely, we study the ways BioNLP influenced the practice of medical NLP.

1 Medical NLP: Specificities and Contributions to BioNLP

A growing community of researchers applies NLP to the medical domain and develops new methods for that purpose. Medical NLP has seen important breakthroughs, such as routine, machine analysis of clinical reports (MedLEE), but it is probably fair to say that it has had until now only a moderate direct impact on clinical applications. It has been mostly concerned with the clinical domain (clinical notes, etc.), but also with the analysis of the scientific literature (MEDLINE titles and abstracts).

1.1 Contributions

We wish to stress nevertheless that medical NLP prepared the ground for BioNLP by providing resources and tools which could be reused in that domain. A large effort was spent on the creation of lexical (*e.g.* the UMLS Specialist Lexicon) and unified terminological resources (*e.g.* the terminologies which can be found in the UMLS Metathesaurus). These resources are used for instance in automatic term recognition, *e.g.* in MetaMap, a widely used component in BioNLP systems. Medical ontologies have also seen a continuous stream of work since GALEN (*e.g.* Foundational Model of Anatomy, SNOMED CT), whose methods could help the design of the Gene Ontology. Indexing and Information Retrieval from the medical literature (*e.g.* SAPHIRE, MTI) and from health records (*e.g.* ICD and SNOMED coding) are long-standing research topics. They aim at concept-based indexing (*e.g.* MetaMap), a task similar to the gene normalization task of the BioCreAtIvE challenges. Text analysis was an early target of Medical NLP. It led to successful systems which were then applied to the biomedical domain (*e.g.* GENIES[2], SEMREP[3]), porting from one sublanguage to the other (Friedman et al., 2002). Finally, literature-based discovery as started by Swanson (inasmuch as it uses NLP) was applied to medical problems long before it was geared towards biomedical knowledge.

1.2 Specificities

The medical domain has specific features which bear consequences on medical NLP research. First, the requirement for *privacy* of clinical records has had a strong impact on clinical NLP. It prevents researchers from sharing text corpora (NLP based on the medical literature does not have this limitation). Deidentification methods

[1] For an introduction, see *e.g.* (Ananiadou and McNaught, 2006).

[2] http://www.cat.columbia.edu/genies.htm

[3] http://skr.nlm.nih.gov/papers/

have been investigated to overcome this barrier, but human review is generally still needed. Second, *localisation* of clinical records and associated functionality is necessary. Clinical records must use the language of the user, which entails a need to develop resources for each natural language (terminologies and ontologies, being designed as concept representations abstracted from natural languages, are an exception: they can be shared across languages). This has created strong constraints on the sharing of resources and tools, which led to the dispersion of concrete efforts in medical NLP. Besides, medical NLP has tackled *several specific sublanguages* beyond that of the biomedical literature: that of clinical reports, often with short phrases and terse style, and more recently those of practice guidelines and patient-oriented documents, with constraints of readability and understandability. Finally, it has addressed *diverse user needs*, mainly those of a variety of health care professionals and administrative staff in hospitals (clinical documents) and those of researchers (articles). This dispersed market segmentation also tends to disperse research.

2 BioNLP: Specificities and Contributions to Medical NLP

2.1 Contributions

BioNLP promotes a dynamic, shared way of conducting research within a community. This can be seen in the organization of challenges (*e.g.* TREC Genomics or BioCreAtIvE). These depend on shared annotated corpora (*e.g.* GENIA[4]), a key component in such initiatives. In contrast, medical NLP researchers have had to overcome the above-mentioned strong limitations on the development of clinical corpora to launch such challenges. This has recently started with the i2b2 de-identification challenge (2006)[5] and the Cincinnati Medical Center ICD-9-CM coding challenge (2007)[6], and the AMIA NLP working group is striving to foster this process. Sharing in BioNLP also applies to information extraction pipelines, *e.g.* LingPipe or the JULIE tools.

BioNLP has had a faster and wider attraction of 'mainstream' Computational Linguistics researchers. More attractive funding of the genomics domain probably played a role here, but the availability of corpora and resources and the organisation of challenges were certainly very important factors too. Directly usable training and testing corpora have also allowed many researchers to test Machine Learning methods (*e.g.* to recognize gene mentions) with only minimal investment in the specificities of the domain.

2.2 Specificities

BioNLP has the great advantage of working mostly on common input documents: the biomedical literature. There is no need for privacy here (but access rights are enforced on the majority of full-text articles), and documents are written in one language: 'bio English'. The biomedical sublanguage inherits that of scientific, experimental literature; it also has specific components, *e.g.* for gene and protein names and interactions. The development of lexical, terminological and ontological resources for these components has therefore been the subject of much work in BioNLP. The emphasis of BioNLP is on text mining, and it has more focused targets, namely, researchers and database curators. This may form a more homogeneous user base than that of medical NLP.

3 Conclusion

Based on the above comparison, hypotheses can be formulated to explain differences between the attractivity and development speed of medical- and BioNLP. Not yet mentioned is the intrinsic scientific attractivity of the biomedical domain, with a promise of more fundamental outcomes. Funding is indeed an important factor too. Nevertheless, we believe the importance of enabling factors must be stressed: a shared input language facilitates shared resources and tools, no requirement for privacy enables shared corpora and the organisation of challenges, which have been a driving force in BioNLP.

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[5] https://www.i2b2.org/NLP/Workshop.php

[6] http://www.computationalmedicine.org/challenge/

References

Sophia Ananiadou and John McNaught, editors. 2006. *Text mining for biology and biomedicine*. Artech House Publishers.

Carol Friedman, Pauline Kra, and Andrei Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform*, 35(4):222–235.