Resolving Structural Ambiguity of Medical Terms with Statistical Model Fitting
(Computational Linguistics, Corpus Linguistics) (15 min presentation)

Abstract

We address a specific problem in Natural Language Processing (NLP) of Biomedical texts. Most medical concepts are expressed via a domain specific terminology that can either be explicitly agreed upon (i.e. Systematized Nomenclature of Medicine (SNOMED)) or extracted empirically from large amounts of domain specific text. A typical medical term is a noun phrase that is often structurally ambiguous, which is illustrated in (a) below:

a. $small_1$ $bowel_2$ $obstruction_3$

The example in (a) can have at least two interpretations depending on the analysis:

a.1 [[$small_1$ $bowel_2$] $obstruction_3$]
a.2 [$small_1$ [$bowel_2$ $obstruction_3$]]

Unlike truly ambiguous general English cases such as "American history professor" where the appropriate interpretation depends on the context, medical terms such as in (a) have only one appropriate interpretation within the medical subdomain regardless of the context. Ambiguity resolution requires domain knowledge (of human anatomy in the case in (a)). Noun phrase parsers are often constructed to exploit domain knowledge encoded in various terminologies and ontologies such as SNOMED to resolve the ambiguity [1, 5]; however, domain specific databases are often incomplete and not up-to-date. Statistical methods are also used represented by probabilistic grammars such as PCFG or dependency grammars [3]. We present a novel structural ambiguity resolution method that uses Log Likelihood (LL) [2] computed for three word SNOMED terms (trigrams) also found in a 10M word corpus of clinical notes. The method is based on fitting a trigram's LL score to one of two models of independence where either the first or the last word are hypothesized to be independent of the other two:

1. Model1: $w_1 w_2 \_ w_3$
2. Model2: $w_1 \_ w_2 w_3$

If the trigram like "small bowel obstruction" fits Model1, then the interpretation consistent with the analysis in (a.1) is correct; otherwise, (a.2) would be chosen.

We present experimental results showing that out of 700 SNOMED 3-word terms, 86% correctness in identifying the correct analysis using our model fitting technique compared with 37% achieved with Metamap parser [1, 5]. The main limitation of this method is the exponential growth in complexity of model fitting with ngrams where N is greater than 3. We plan to continue extending this methodology to 4 and 5 grams, as 5 is the average number of words in a medical term. This methodology can be used in conjunction with available domain knowledge in order to improve parsing of medical terms and their mapping to controlled terminologies..

Bibliography:

1. Aronson, A. R. 1996. MetaMap: Mapping Text to the UMLS Metathesaurus.
2. Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1):61-74.
3. Manning, C. and Shütze, H. 2001. Foundations of Statistical Natural Language Processing. *MIT Press.*
4. Moore, R. 2004. On Log-Likelihood Ratios and the Significance of Rare Events. *In Proc. EMNLP'04.*
5. Rindflesch, T. C. *et al.* 2000. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature", *Proceedings of Pacific Symposium on Bio-informatics (PSB'2000)*.

Word Count: 479