Approximate dictionary matching for biomedical information extraction

György Móra

Information extraction techniques are widely used in various text mining fields like scientific literature mining and Web mining. Named entity recognizers often make use of word lists, that they subsequently map to the text. There are efficient methods available for mapping large dictionaries to texts. The species name recognizer Linnaeus, for instance, uses regular-expression-like patterns generated from a simple list of organism names [3]. The patterns are generated in such a way that the system can recognize spelling variants, but their flexibility is limited. Biomedical entities display a great spelling variability, hence the tagging of gene and protein names is essential for good information extraction in the biomedical domain. We developed a method for effectively matching large dictionaries containing millions of entity names using a flexible matching procedure. Our system was developed for gene and protein name matching, but it can also be used for other tasks where flexible matching is required.

Our approach tries to overcome the difficulties of biomedical entity matching. The order of the words in these entities may be different and extra punctuation characters and words may appear inside entity names. For instance, *member 3 of the solute carrier family 6* can be written as *solute carrier family 6, member 3*. Numbers can be written as Roman or Arabic numerals, which increases the number of possible variations. Greek letters are often transcribed or a similar Latin letter is used. For instance, *nuclear factor kappa beta* can be written as *NF-kappa B* or *NF-*$\kappa\beta$. To match all linguistic forms of a word we determined the base form of all words used in all dictionary entries and during the matching the base forms of the words in the texts are matched against the base form of the entries in the dictionary. Biomedical entities often have nonconventional plural or adjective forms. In our experiments, we used the LVGTools [1] linguistic package to determine the possible base forms of the words in the dictionary and in the text. Most words have one possible base form (like *dogs -> dog*), but there are ambiguous cases where only a syntactic or semantic analysis can determine which base form belongs to the word form in a given context. The LVGTools also handle the above-mentioned issues with Greek letters and Roman numerals as it generates all possible 'interpretations' of the ambiguous word.

Our approach utilizes the Lucene search engine [2] for indexing. The dictionary entries are split into words and the possible base forms of the words are determined. The entities are indexed with respect to the base forms and the number of words in the entity name. When the text is tagged, the possible base forms of the words are also determined. If one of the base forms of a word is contained in the index, the word is marked as a potential part of an entity name. Based on these annotations, a value for each word in each sentence is calculated which tells us how many words the longest entity name can have, which the current word may be a part of. This value is used as a maximum value when the possible names are queried from the index. This restriction speeds up the process because an entity name may be as long as 22 words. The tagging speed is approximately one document per second on biomedical full articles with over 7 million different entity names using a computer with a 2 GHz processor. The spans of the hits may overlap and a rule-based system filters out the matches to produce the longest non-overlapping annotations.

We developed a dictionary lookup tagger, which can be effectively applied in biomedical information extraction pipelines, to perform string matching with the necessary flexibility. We showed that this task can be completed in reasonable time with alternative indexing and searching methods.

References

[1] *SPECIALIST Lexicon and Lexical Tools*. National Library of Medicine (US), September 2009.

[2] Cutting, D. Lucene, 2001.

[3] Gerner, Martin, Nenadic, Goran, and Bergman, Casey M. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85, January 2010.