Technische Universität Berlin, Institut für Sprache und Kommunikation
Humboldt-Universität zu Berlin, Wissensmanagement in der Bioinformatik

Peter Palaga

Extracting Relations from Biomedical Texts Using Syntactic Information

Magister Thesis Exposé

2008-10-03

Author: Peter Palaga, Matrikelnummer: 228182
Supervisors: Prof. Dr. See-Young Cho, Prof. Dr. Ulf Leser

Der Sachverhalt ist eine Verbindung von Gegenständen (Sachen, Dingen).
Ludwig Wittgenstein (1921)

1 Entities and Relationships

Well-formed sentences of natural language usually state facts. Intuitively, such facts can be seen as consisting of two different types of items: firstly, there are some things and secondly there are some associations that are claimed to hold between that things. In the field of natural language processing (NLP), such things are called *entities* and the associations between them are called *relationships*. Consider the following sentence as an example:

(1) Alice spends more money on eBay than Bob.

In this sentence, (at least) two entities were mentioned, namely persons Alice and Bob. Furthermore, the sentence asserts that there is a particular relationship between them: the former spends more money on eBay then the later.

To state which relationships hold between which entities is a typical task of researchers in natural sciences. So, biologists are observing cells to say (among other things) which genes cause which diseases, which genes inhibit other genes or which proteins influence/communicate which other proteins.

Scientists mostly publish their findings in form of free text, e.g. papers, books, etc. With raising number of known facts, it becomes necessary to systematize the knowledge in a more structured form, e.g. a database. Structured data sources allow richer queries than free text. For example, it is not effectively possible to find out which genes are inhibited by gene SOX9 only using free text queries on a big collection of scientific text.

Indeed, there exist attempts to create databases storing entities and their relations by hand. However, such projects are far from being complete (see e.g. Chatr-aryamontri et al., 2006).

It is a task of Information Extraction to develop methods for creating such databases automatically. This task is commonly seen as having three phases: (1) entity recognition (2) coreference resolution and (3) relation extraction (Culotta and Sorensen, 2004). In this thesis, the last phase will be focused.

2 Methods for Relation Extraction

Several approaches have been applied to extract relations from free text. The simplest method is based on the assumption that a mere co-occurrence of two entities implies that there is a relationship between them. Trivially, such methods reach 100% recall, but their precision stays low (Pyysalo et al., 2008).

Pattern and rule based methods try to use context information for finding relations between entities. They usually look for certain words occurring near entity names or use part-of-speech (POS) and/or syntax information. They usually exhibit high precision, but recall is low, i.e. many of the relations in the text are left uncovered by them. The patterns used by such approaches may be constructed by hand or learned automatically from an annotated corpus (Hakenberg et al., 2005; Fundel et al., 2007).

In this thesis, a statistical approach will be pursued. Statistical methods typically use a statistical classifier to predict the presence or absence of a relation between a given pair of entities in a sample sentence. The decisions of such a classifier rest upon a statistical model

which is usually produced by a training on a text corpus containing positive and negative examples. (Donaldson et al., 2003)

Several statistical classifiers have been used in the field of relation extraction, among others nearest neighbor (Fukunaga, 1990), naïve Bayes, sparse regularized least-squares (RLS) (Airola et al., 2008) and support vector machines (SVM) (Cortes and Vapnik, 1995; Culotta and Sorensen, 2004). SVM will be used in this thesis.

Internally, classifiers use some special representation of sentences to learn from or to classify. Such representations can be based either on feature vectors or on kernel functions.

Feature vectors are ordered *n*-tuples of binary (true/false) values. To use them as an input for a classifier one needs to define which positions in the vector characterize which properties of a given sentence. Searching for an optimal feature set is a substantial task for a user of this method. It can be accomplished either manually or automatically. Examples for features one could choose may be based on word *n*-grams, POS *n*-grams, or parse tree substructures like *has an NP-VP subtree* (Culotta and Sorensen, 2004), or similar.

Kernel function can be seen as similarity measure of two given data instances – in our case sentences. Given a set of labeled instances, a kernel function based classifier determines the label of a novel instance by comparing it to the labeled training instances using this kernel function (Culotta and Sorensen, 2004).

Lets us mention “bag-of-words” as an example for a kernel function. It simply characterizes the similarity of two sentences through the number of words they have in common.

Convolution kernels represent a special type of kernel functions. They are intended for cases when the structure of instances is important. The main idea is to qualify the similarity of two structures through summing the similarities of their substructures. In this way the similarity of two strings can be characterized through the number of their common substrings, which are weighted by their length. Or analogically, the similarity of two trees can be determined as the number of their common subtrees. The matching substructures can be effectively found using dynamic programming (Culotta and Sorensen, 2004).

The approach with dependency tree based convolution kernel function will be applied in this thesis.

3 Pipeline

In this thesis, a statistical extraction system will be implemented. It will comprise a set of modules, which will be ordered in a pipeline where each module will use the output of its predecessor as its input.

For the most of the modules there are implementations available which only need to be integrated. The following modules are supposed to occur in the pipeline. They are presented together with available implementations.

- Tagger (optional)
  - MedPost
- Parser
  - Stanford Lexicalized Parser (Klein D, Manning Ch. D., 2003a, 2003b)
  - Bikel Parser: http://www.cis.upenn.edu/~dbikel/software.html

- Full parser: Charniak and Lease (2005)
- Collin's Parser: http://people.csail.mit.edu/mcollins/code.html

- Parse tree postprocessing (optional)
  - fnTBL Noun Phrase Chunker: http://www.cs.jhu.edu/~rflorian/fntbl/index.html
- Classifier
  - SVM Light: http://svmlight.joachims.org/

4 Corpora

The chosen method presupposes a resource with annotated entities and relevant relations on which the statistical model will be trained. There are several PPI-annotated resources available. Pyysalo et al. (2008) have transformed five corpora to a common format. These corpora will be used in this thesis.

5 Evaluation method and metrics

The same method may perform very differently on different corpora. Pyysalo et al. (2008) show that the choice of corpus may have a stronger impact on the result than the choice of the extraction method. They propose to use the trivial co-occurrence approach as a baseline.

Precision, recall and F-measure are standard measures for evaluating relation extraction systems. They can be applied in different ways, thus delivering substantially different results. Here again the proposition of Pyysalo et al. (2008) and Airola et al. (2008) will be followed: relations are considered as untyped and undirected pairs of specific protein mentions; reflexive relations (self-relations) are not considered as relations.

F-measure has been severely criticized by Airola et al. (2008). Their argument is that F-measure is very sensitive to the underlying positive/negative pair distribution of the corpus. For a classifier, it is much more probable to reach high recall on a corpus which has more positive examples than negative examples.

Airola et al. (2008) propose the *area under the receiver operating characteristics curve* (AUC) measure (Hanley and McNeil, 1982). Unlike F-measure, AUC is invariant to the class distribution of the used dataset.

6 References

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (20069): MINT: the Molecular INTeraction database. Nucleic Acids Research, 2007, Vol. 35, Database issue D572-D574

Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T (2008): A Graph Kernel for Protein-Protein Interaction Extraction, BioNLP 2008.

Culotta, A, Sorensen, J (2004): Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting on Association For Computational Linguistics (Barcelona, Spain, July 21 - 26, 2004). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 423. DOI= http://dx.doi.org/10.3115/1218955.1219009

Charniak, E., Lease M (2005): Parsing biomedical literature. In Proceedings of IJCNLP’05, pages 58–69.

Cortes C, Vapnik V (1995): Support Vector Networks, Machine Learning 273-297

Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CW (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4(1): 11.

Fukunaga K (1990): Introduction to Statistical Pattern Recognition. Academic Press.

Fundel K, Küffner R, Zimmer R. (2007): RelEx--relation extraction using dependency parse trees. Bioinformatics. 2007 Feb 1;23(3):365-71. Epub 2006 Dec 1. PMID: 17142812

Han J, Kamber M. Data Mining (2000): Concepts and Techniques. Morgan Kaufmann.

Haussler, D. (1999): Convolution kernels on discrete structures, UCSC-CRL-99-10

Jörg Hakenberg, Conrad Plake, Ulf Leser, Harald Kirsch, and Dietrich Rebholz-Schuhmann (2005): LLL'05 Challenge: Genic Interaction Extraction with Alignments and Finite State Automata. Proc Learning Language in Logic Workshop (LLL'05) at ICML 2005, pp. 38-45. Bonn, Germany.

Klein D, Manning Ch. D. (2003a): Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.

Klein D, Manning Ch. D. (2003b): Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Pyysalo S, Airola A, Heimonen J, Björne J, Ginter F, Salakoski T (2008): Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics. 2008 Apr 11;9 Suppl 3:S6. PMID: 18426551

Wittgenstein, L. (1961 [1921]): Tractatus-Logico Philosophicus. Routledge & Kegan Paul.