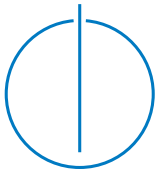
FAKULTÄT FÜR INFORMATIK

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master Thesis

Relation Extraction of Protein Localizations from the Biomedical Literature

Shrikant Vinchurkar



FAKULTÄT FÜR INFORMATIK

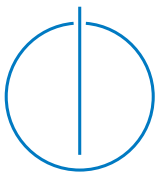
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master Thesis

Relation Extraction of Protein Localizations from the Biomedical Literature

Extraktion von Protein Lokalisierungen aus der Biomedizinischen LiteraturRelation

Author:	Shrikant Vinchurkar
Supervisor:	Prof. Dr. Burkhard Rost
Advisor:	Juanmi Miguel Cejuela
Submission Date:	March 15, 2015



I assure the single handed composition of this master thesis only supported by declared resources.

Munich, March 15, 2015 Shrikant Vinchurkar

Acknowledgments

Abstract

Contents

1 Introduction

1.1 Explosion of Biomedical data

In recent times, there has been phenomenal rise in the amount of textual information. At the time of writing, the database of Biomedical articles - PubMed boasts of more than 24 million articles. The amount of data available in the Biomedical research community is also increasing due to high throughput experiments. With huge amount of information in Biomedical texts, it has been impossible for researchers to keep track of all developments in the field. The Information Extraction (IE) techniques are helping the researchers to extract relevant information from this gold mine. As a result, there has been tremendous increase in research in the area of natural language processing and text-mining. These techniques when applied to Biomedical text is found to be quite helpful to the researchers who cannot manually scavenge thousands of articles for information.

1.2 Need for automatic extraction of information

Need for Biomedical research community

The biomedical research community is interested in wide variety of problems like protein-protein interaction, protein-ligand interaction, protein-subcellular location, protein mutation, protein structure prediction, protein function prediction etc. All these different cellular activities have a far reaching effect on human health and plays phenomenal role in causing diseases. Since a lot of researchers across the globe are working simultaneously on some of these problems, it is quintessential for an researcher to understand the recent advances and start building from there.

Conventionally, the best method to keep oneself updated about recent scientific advances is by reading the scientific literature. This remains to be the best way in some areas of science where the scientific throughput is less and harder to achieve. However in some fields like biomedical sciences, due to recent advances in the technology and construction of high throughput machines, it has become possible to do science at a faster rate. Since human health is an issue of paramount importance, the research funding in biomedical sciences has also remained sufficient which added to more

amount of research throughput. With huge amount of scientific literature available, it has therefore become very difficult for an researcher to stay updated about recent work being done in his area of interest.

The automatic extractors of information has played a key role in gathering information. The automatic extractors are mainly text-mining methods which can extract useful information from literature with some amount of confidence. These information extractors, therefore, have started playing a key role in biomedical sciences.

1.3 Protein-subcellular localization extraction

Why it is important ?

The subcellular localization of proteins is one of the most studied topics in the field of biomedical sciences. The proteins are a chain of amino acids. The proteins are created in the cytoplasm by a flow of genetic information from DNA to RNA and from RNA to Proteins. This process is also called as central dogma of molecular biology. The genetic information present in the DNA (Deoxyribonucleic acid) is converted to RNA (Ribonucleic acid) by a process called as transcription. Transcription takes places in the nucleus of the cell. The RNA strand so created travels from nucleus to cytoplasm through the nuclear pores. In the cytoplasm, the RNA strands are used to create protein with the help of cellular machinery called as Ribosomes. The process of converting RNA to a chain of amino acids called as proteins is called translation. Therefore, the creation of proteins consists of two main phases viz. transcription and translation.

Although the proteins are created in the cytoplasm, they either perform their biological functions in the cytoplasm or some cellular organelle or exit from the cell. Therefore, the biological functions of the proteins largely depend on their location and knowledge of the subcellular location of the protein is key to understanding its biological function.

Important role in drug research

The proteins are called as the building blocks of life and they play a very important role in functioning of cell machinery. Therefore, an abnormality in its functioning can result in adverse effects even leading to a disease. The study of proteins and their subcellular location is important stage in the development of a drug and the ability to predict the subcellular location of a protein is helpful to identify suitable drug, vaccine and diagnostic treatment [CITE:LIU].

Current sources of information extraction

Currently, the important sources of extraction of information about protein subcellular location are manual curation from the literature, high-throughput microscopy-based screens and prediction from primary sequence [CITE:COMPARTMENT].

The text-mining methods help in automatic extraction of this information from the literature and these methods have started playing a key role extraction of protein subcellular location extraction.

1.4 Existing work

There hasn't been so much of work done in the area of protein-subcellular localization relation extraction. There has been some projects which have tried to work on the same problem and there are a lot other projects that are working on an allied problem. This section presents an overview of the recent work done in the related area.

Using rich syntactic information for relation extraction - Liu.et.al

Lie et.al [CITE:LIU] proposed a method for extraction of protein-organism (PO) relations and protein-location (PL) relations from the text using sytanctic parse trees. The protein-organism (PO) and protein-location (PL) relations are then merged to predict a ternary protein-organism-location (POL) relation.

The corpus used in this method is composed of MEDLINE titles and abstracts annotated by domain expert biologists and parsed by Charniak-Johnson parser.

The focus of this work is to extract relations present in the same sentence and does not considers the inter-sentence relations. Two models are developed for extraction of relations viz. Semantic Role Labeling based relation extraction (SRL) and Tree kernel based relation extraction (TRK). Both models use features extracted from syntactic parse trees.

The model based on SRL uses manually extracted features from syntactic parse trees. These features are extracted along the path from protein to location/organism. This model is then trained by a Binary SVM using default linear kernel from Joachim's SVM light [CITE:JOACHIM SVMLIGHT].

The model TRK uses entire parse tree as input for tree kernel. The model is trained by default tree kernel from Moschetti's SVM-light-TK-1.2 [CITE:MOSCHETTI'S TK].

Table 1.1 shows the results for combination of both models.

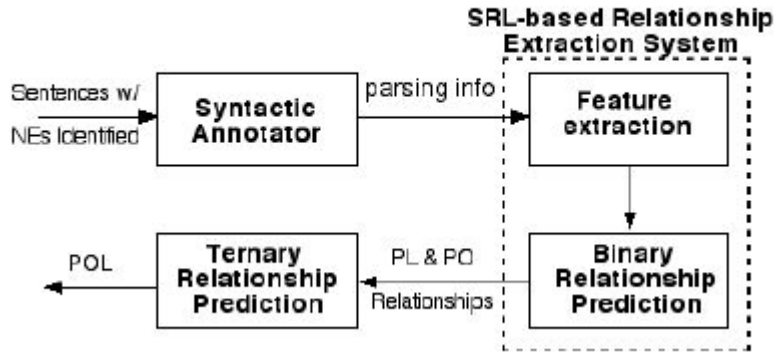


Figure 1.1: Flow of information in relation extraction models

Relation	Precision	Recall	Fscore	Accuracy
PL	74.9	79.4	77.1	72.8
PO	73.9	78.1	75.9	72.6
POL	75.3	74.5	74.9	71.8

Table 1.1: Results of SRL+TRK

BioNLP shared task and localization events

The BioNLP shared task [CITE:BIONLP SHARED EVENT] events in 2009 and 2011 focused on the aspects of event extraction from Biomedical literature. The BioNLP events provided an annotated corpus to the the text-mining community and asked for submission of innovative solutions for the tasks. One of the tasks was shared task in which the events present in the text have to be extracted. The data for shared task was a subset of GENIA event corpus [CITE:Genia event corpus]. Although, the original GENIA event corpus consists of 35 events, the shared task event corpus consisted of 9 events. One of the events common to both BioNLP event corpus as well as the main GENIA event corpus is localization event and therefore, the methods employed in these BioNLP events can be an important resource of study for the purpose of protein-subcellular location extraction.

Some of the issues with BioNLP corpus

Although localization task is quite similar to protein-subcellular location extraction task, there are subtle differences which disallows using the same methods as it is for

our task of protein-subcellular location extraction.

One of the issues is that not all the locations found in the GENIA event corpus are subcellular locations. Some of them are even specific cells or tissues in the body. Some localization events were annotated without any mention of actual location but with the presence of a clue which hints that the text contains a localization event. In addition, the location mention contains extraneous words that are a part of the phrase but not absolutely essential for location mention. Owing to all these problems, we decided not to use GENIA event corpus as it is, develop our own corpus and develop a method using it.

State of the art in BioNLP'09

The method developed by Björne et.al. [CITE:BJORNE PAPER] registered the state of the art performance in BioNLP Shared Task Event 2009.

The shared task of event extraction consisted of three stages viz. Trigger detection, argument detection and event extraction. Triggers are the tokens or set of tokens that indicate the presence of the event. Once the triggers are detected, the triggers can be joined to other triggers or proteins as a part of predicate-argument structures. Finally, the triggers and their arguments are suitable combined to form an event.

This method extract events from the same sentence and do not look for inter-sentence event. The method developed by Björne et.al. represents a sentence as a graph where nodes are the tokens of the sentence and dependency relations are the edges between them. Every entity (Protein/DNA/RNA) forms a node in the graph as well. The event triggers are recognized and they also form a node in the graph. The edge between the trigger and argument is classified mainly as theme, cause, negation etc. The model is trained by rich set of features extracted from the graph. For every stage of event extraction, somewhat different set of features are used.

The sentences are tokenized and parsed before processing. The task of event extraction is accomplished in 3 steps:

1. **Trigger Recognition**: Each token is passed into multi-class SVM which predicts the event class depending on feature set. Each event class recognized forms a node in the final graph.

2. **Argument/Edge detection**: Every edge between event-event nodes and event-entity nodes is classified into theme, cause or negation.

3. **Semantic post-processing**: It is a rule based step to remove irrelevant connections.

The method uses BioNLP event corpus as the dataset and Joachim's SVM Light [CITE:

JOACHIM MULTICLASS] for multiclass classification. Different multiclass classifier are built for classifying triggers and edges between triggers and arguments.

Following are the results of the method developed by Björne et.al.:

Event Class	# Events	Recall	Precision	Fscore
Localization	174	49.43	81.90	61.65
Total	3182	46.73	58.48	51.95

Table 1.2: Results of Graph based event processing

Using MLN for event extraction (Yashikawa et.al)

Extending the results of Björne et.al., Yashikawa et.al. [CITE:YASHIKAWA] used coreference resolution to extract relations that span over multiple sentences in addition to extracting relations from same sentence.

Yashikawa et.al.uses GENIA event corpus instead of BioNLP'09 corpus since GENIA corpus provides proper cross-sentence coreference. They show that the SVM multiclass method proposed by Björne et.al. gives better results if it makes use of coreference annotation. In addition, they also developed a new model based on joint Markov Logic Network (MLN) which improved the results significantly.

The property of transitivity is used such that if there is a relation between event trigger & entity and the entity has antecedent which can be resolved through anaphora/-coreference resolution, then there is a relation between event & antecedent.

The MLN model achieves a F1 score of 53.8 with naive coreference resolver and 56.7 with gold coreference annotation.

COMPARTMENTS

Binder et.al. [CITE:COMPARTMENT] designed the COMPARTMENTS resource which integrates protein-subcellular location relations from different sources such as databases and prediction tools. To create a uniform representation of the information, the proteins and localizations are mapped to common protein identifiers and GO Ontology terms. Confidence scores are assigned to the localization evidence to enable comparison of different types and sources of evidence. The unified location evidence for a protein is then visualized on a schematic cell to provide a simple overview.

There are different channels or sources from which the information is collected. The first channel is a *knowledge* channel which contributes information based on annotation from databases like UniProtKB, MGI, SGD, FlyBase and WormBase. The second channel of information called *experiments* is based on HPA - ongoing effort to experimentally

validate the tissue expression and subcellular localization for entire set of human proteins. The third channel which is of particular interest to us is *automatic text mining* and the fourth channel of information is *predictions* which depend on prediction of subcellular location depending on protein sequence.

The automatic text mining channel contributes information extracted by a text-mining method. The text-mining method which extracts information from MEDLINE abstracts works on the fact that more the protein and cellular compartment are co-mentioned, the more likely the protein is localized to that compartment. This leads to calculation of a score that determines the probability/confidence of localization of a protein to a cellular compartment. This text-mining method can be useful for comparing the performance with our method.

1.5 Need for a dedicated method

As mentioned previously, the GENIA event corpus or the BIONLP shared task corpus cannot be directly used for extracting protein-subcellular location relations. Therefore, there is a need to have a dedicated corpus designed for the purpose of training the text-mining method.

In addition, using the nuances of the new corpus, the new method can be trained on the corpus annotated and can produce better results.

Although, the model developed by Yashikawa et.al. extracts relations spanning over multiple sentences, extracting coreference information from the text at run-time is pretty slow. The method of Yashikawa et.al could work since their intention was not to create a runtime relation extraction.

We had decided to develop a dedicated text-mining method to extract protein location relations present in the corpus. We did not focused just on the same sentence relations but also on difference sentence relations in which the participating entities are in different sentences. In our annotated corpus, we also found out that 40% of total relations cross sentence boundaries and therefore, developing a method that would consider different sentence relations would help us to extract more relations from the data.

2 LocText Corpus

2.1 Need for a separate corpus

As mentioned in the previous chapter, the GENIA event corpus or its subset, the BIONLP shared task corpus is not the best corpus that can be used for training a model to extract protein- subcellular location relations. As pointed out previously, following are some of the issues related to the corpus:

1. Not all the locations found in the GENIA event corpus are subcellular compartments. Some of the locations are the names of cells or tissues in the body.

2. In some mentions of subcellular compartment, the actual mention contains extraneous words in addition to the mention of subcellular compartment. These extracellular words takes away the preciseness of the mention of subcellular compartment.

3. Some localization event does not contain an actual mention of subcellular compartment but the context just points out the clue leading to hypothesis that a localization event may have been mentioned.

To summarize, the GENIA event corpus have some serious concerns and cannot be directly used if we are trying to train a classifier for protein-subcellular compartment relation extraction. There was a need to create a separate corpus dedicated for this task.

2.2 LocText

2.3 Corpus annotation process and guidelines

2.4 Inter-annotator agreement

2.5 Important conclusions from the dataset

2.6 Corpus statistics

2.7 Linked Annotation

3 Protein Location Relation Extractor: Materials & Methods

3.1 Repplication of Bjorne's work ?

3.2 Method pipeline

3.3 Support Vector Machines

3.4 Machine Learning Models

3.4.1 SameSentenceModel

3.4.2 DiffSentenceModel

3.5 Graph representation

3.5.1 Graph representation for SameSentenceModel

3.5.2 Graph representation for DiffSentenceModel

3.6 Feature extraction

3.6.1 Feature extraction for SameSentenceModel

3.6.2 Feature extraction for DiffSentenceModel

3.7 Feature Selection

3.8 Training, Cross validation and Classification

3.9 Experiments that worked and that did not work

3.9.1 Experimentation for SameSentModel

- Removing sentences which had no proteins improved the performance - Using different kernels did not work

3.9.2 Experimentation for DiffSentModel

- Using extra information from the document such as shortform longform pairs decreased the performance since it added more FP's (15) than TP's (1). It seems that this rule was nicely followed during annotation

3.9.3 Experimentation with different kernels

3.10 Tools used

4 Results

4.1 Evaluation Criteria

4.2 Results for SameSentenceModel

4.3 Results for DiffSentenceModel

4.4 Results for CombinedModel

4.5 Assessment of Performance evaluation results

5 Conclusion

5.1 Conclusions from results

5.2 Future work and directions

5.2.1 Use of coreference

5.2.2 Use of different methods like MLN

List of Figures

List of Tables