Advanced Parser for Biomedical Texts

Anton Karazeev

Moscow Institute of Physics and Technology, Institutskiy per., 9, Dolgoprudnyy, Moscow Region, Russia,

anton.karazeev@phystech.edu

Maxim Holmatov

Saint-Petersburg State Pediatric Medical University, St. Petersburg, Russia,

maksim.holmatov@gmail.com

Anatolii Zaikovskii

Saint-Petersburg State University, St. Petersburg, Russia, st010379@student.spbu.ru

Ilia Korvigo

Moscow Institute of Physics and Technology, Institutskiy per., 9, Dolgoprudnyy, Moscow Region, Russia,

ilia.korvigo@gmail.com

Mikhail Skoblov

Moscow Institute of Physics and Technology, Institutskiy per., 9, Dolgoprudnyy, Moscow Region, Russia

mskoblov@gmail.com

Large amounts of biomedical data available to us today from various sources make it at least impractical and in many cases impossible to analyze by hand even if confined within a specific problem. On the other hand most of these data are stored in a natural language form which makes it hard to process automatically. Fortunately a vast experience gained in the field of natural language processing (NLP) can be utilized to automate this process. We developed an advanced parser for biomedical texts that should simplify both data retrieval and analysis.

We considered the following problems:

1. parsing of informative multiword phrases
2. parsing and detection of chemical names written in different notations - trivial notation and IUPAC and SMILES-like
3. assigning word embeddings for parsed words and phrases
4. analyzing complex syntactic dependencies between them

After preliminary analysis of proteins' descriptions from UniProt the existence of multiword biomedical terms was noticed. We decided to learn to extract informative n-grams to improve parsing quality (e.g., instead of [‘amino’, ‘acid’, ...] we want to get [‘amino_acid’, ...]).

The simplest method was based on coincidence statistics (frequencies) of words - we took top-n most frequent words and phrases (but there were too many false-positives). This simple frequency-based method doesn’t evaluate the loss of contextual information caused by misparsing of multiword phrases. Quite obviously, this is not an optimal solution, because the total number of n-grams is combinatorially huge. Therefore we introduce a method which involves information theory and Bayesian inference to handle such a hard task.

Inspired by the word2vec package [Mikolov et al.] we apply neural networks written in TensorFlow for building representation of words - word embeddings. We are working on solving problems with computation of KL-divergence between multinomial distributions. Since contextual distributions are empirical, we approximate them with multidimensional prior distributions — using Gaussian mixture model — that’s the key point of KL-divergence calculation between multidimensional distributions. One of the tools that we use is Variational Autoencoder [Kingma et al.] — it helps us to reconstruct contextual distributions by given word. There are two steps: firstly, we build word embeddings and secondly, assign multivariate Gaussian distribution to every word [there are two vectors of parameters — vector of means and vector of standard deviations]. VAE is used as Bayesian inference at the step of parameterization our empirical distributions.

Our goal is to measure informativeness of n-grams using KL-divergence between distributions of contexts by given word. We compute (1) context distribution over texts containing both words ‘A’ and ‘B’ separately and (2) context distribution over texts containing only phrases (e.g. ‘A_B’). Then we calculate KL-divergence between (1) and (2) distributions to determine whether given words form a multiword biomedical term.

To better identify informative n-grams and give a numerical estimate of their validity two main approaches were used. First one relies on finding the most important edges in word collocation network for analyzed text. Word collocation networks are weighted directed graphs with each vertex corresponding to a word in the text and edge weights equal to the bigram frequency in the document. The most important edges are found by calculating centrality measures of network (degree, closeness, betweenness, etc.) or with the PageRank algorithm [Lahiri et al.]. This process can be applied to analyze documents separately or to generate a custom dictionary of n-grams from a large corpus of texts. Second approach uses term frequency–inverse document frequency (tf-idf) statistic. It rewards frequent terms inside a document but punishes words that are frequent in the whole corpus which helps to filter out the words that are just commonly used in a language.

A major problem with natural language data is that it consists of sentences with convoluted syntax and complex word dependency tree. Thereby our aim is to bring those sentences to such simple structure that could be understandable by computer. In order to gain this we, firstly, replace all names of chemical compounds by one token. It is crucial, because there are a lot of unique chemical terms in mutagenesis data, which make difficulties with training neural networks, especially in lack of data. We reach it using LSTM and take pubchem as source of chemical data for training this network. Then we use seq2seq framework for TensorFlow to bring sentences in same syntax structure. We have chosen seq2seq instead of word2vec because in our case we have too much of rare tokens in data. And finally we use The Stanford Parser to simplify the sentences.

Site-directed mutagenesis data from uniprot was used as a benchmark. A portion of these data was analyzed by hand and brought to uniform format easily parsed by computer. These data are really interesting because it is almost the only way to get information about changes of protein caused by mutations in useful for computer format.

1. Diederik P Kingma, Max Welling (2013) Auto-Encoding Variational Bayes
2. Shibamouli Lahiri et al. (2014) Keyword and keyphrase extraction using centrality measures on collocation networks. arXiv:1401.6571v1
3. Wei Liu, Bo Chuen Chung, Rui Wang, Jonathon Ng, Nigel Morlet (2015) A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters
4. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013) Efficient Estimation of Word Representations in Vector Space
5. Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, Christoph M. Friedrich (2008) Detection of IUPAC and IUPAC-like chemical names