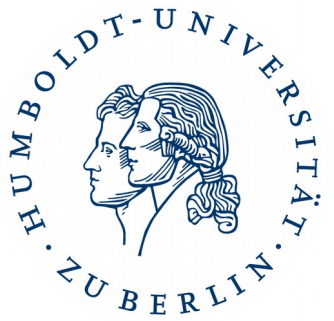
Humboldt-Universität zu Berlin
Department of Computer Science
Knowledge Management in Bioinformatics
Exposé - Study Project



Neural Biomedical Named Entity Normalization

by Christopher Schiefer

19th of May 2019

Motivation

Methods for automatic information extraction from vast amounts of unstructured text become highly necessary due to the rapid growth of the biomedical literature (Zhiyong, 2011) (Mura et al., 2018). It is essential to identify biomedical entities in text documents to enable tasks like searching for specific entities, extracting document background information and linking similar documents (Rzhetsky et al., 2008).

This task is difficult because of the huge variability of names used for biomedical entities in the literature (Erhardt et al, 2006). For instance genes have multiple spelling conventions, equivalent abbreviations or synonyms and mostly consist of many words. It is non-trivial even for a human to find relevant identifiers for biomedical named entities and inter-annotator agreement hardly exceeds 90% (Leser and Hakenberg, 2005) (Morgan et al., 2008).

The task of identifying specific entities in an unstructured text is called *named entity normalization* (NEN) which extends the *named entity recognition* (NER) task. In NER the aim is to find entities and determine their general types (i.e. gene, chemical or disease in the biomedical domain), whereas normalization aims at determining a unique identifier from a dictionary for each specific entity. Typically, NER is performed as the first step to find entities in a text which are then to be normalized.

Previous approaches for normalization focus on a single entity type (e.g. chemicals, diseases or genes) and are based on sets of rules, large dictionaries, and pre-defined features that are expected to capture the knowledge of experts. This not only takes a lot of effort for transforming the knowledge into a machine-usable format but also leads to highly specialised solutions for one specific entity type which are costly to maintain and still cannot find all possible mentions of each entity. However, this is far from perfect and on cross-species corpora the performance given as F-score hardly exceeds 50% (Wei et al., 2015). For genes, the lack of naming conventions and the large number of variants make it hard to map an ambiguous gene name to identifier. Moreover, many names are attributed to multiple identifiers or are used for various species. The Entrez Gene dictionary alone contains more than 22 million genes with more than 4.5 million synonyms and is subject to continuous maintenance.

An approach by (Habibi et al., 2017) for NER shows that generic methods based on deep learning are able to outperform the state-of-the-art without relying on any external knowledge base. Since NER is an integral part of NEN, this sets the motivation for this study project to create a generic approach to the task of entity normalization and then compare its performance to established methods. Since the resources, e.g annotated corpora and dictionaries, are most comprehensively available for genes, the evaluation is performed on gene names, but the method should also be easily applicable to other entity types. Although the implementation is completely generic, the results will depend on how sophisticated the data is to learn from.

Goals

The aim is developing a workflow for entity normalization independent of pre-defined features which encode background knowledge on a particular entity type. Instead, the input will consist only of labeled data that can be leveraged to learn the genes' representations in texts, as well as a list of gene identifiers with their names and synonyms. In the scope of this study project, I will implement the proposed approach and apply it along with the baseline solutions described below on three gene normalization corpora to obtain comparable test results.

As described above, the entity extraction process consists of NER and NEN, and one can separate the second step again in two subtasks, candidate generation and disambiguation (Hachey et al., 2013). These three steps build on one another but can also be developed and evaluated in isolation. My approach will be based on the work by (Habibi et al., 2017) on NER with deep learning, and will extend it to the whole normalization process while remaining as generic as possible. The experiments will only be performed on gene named entities, by comparing this generic method with three baselines with the highest NER and NEN performances: *GNAT, GNormPlus* and *GeneTUKit.* A fourth tool, *TaggerOne*, will be included in the comparison since it follows a similar generic approach as the one described here, however, no results on its performance on genes have been reported yet.

Previous Work

GNAT applies background knowledge from several sources with a mixture of dictionaries, CRFs (Conditional Random Fields) and pre-defined features to disambiguate gene mentions. It includes an extra step to identify the correct species with a set of rules based on the context of the mention (Hakenberg et al., 2008).

GNormPlus is an open source tool from 2015 which uses CRFs and implements manually constructed features for different aspects as general linguistics as well as semantic and case pattern features. For normalization, different matching strategies were implemented, as well as abbreviation resolution and compose mention simplification (Wei et al., 2015).

GeneTUKit was developed for gene normalization in 2010. It utilises several tools to make use of dictionary-based approaches as well as the local and global context of a mention. It applies CRFs and dictionary look-ups to identify entity mentions, Lucene to generate candidates, a ranking algorithm to disambiguate genes, and, lastly, a SVM (Support Vector Machine) to generate confidence scores (Huang et al., 2011).

TaggerOne was the first machine learning approach that combines NER and normalization in a joint model, meaning both training and prediction are done in a single step. It defines scoring functions to assign a score to each text segment for each NER class and each possible normalization entity with simple text features. Using semi-Markov models, it aims to optimise the segmentation for a text so that the sum

of scores given to each segment is optimal. TaggerOne can be trained for each entity type and achieved state-of-the-art results for disease and chemical entities, but has not yet been applied to genes (Leaman et al., 2016).

(Habibi et al., 2017) implemented a generic method for recognising any entity type which outperformed other state-of-the-art methods by about 2 to 3 percentage points, depending on the entity type. The method combines word embeddings (vector representations of a word by taking into account its context), long-short-term-memory-networks (to learn a non-linear combination among features) and CRFs (sequential classifier considering a sequence of word labels for labelling a new word) without having to rely on the typical background knowledge. It only requires an annotated gold standard and a large, entity-independent corpus in the considered domain to compute word embeddings on.

So far, all of the approaches on gene normalization depend on features defined by the developers. In contrast, methods using word embeddings and neural networks were able to outperform previous approaches significantly, as shown e.g. by (Francis-Landau et al., 2016), (Sun et al., 2015) and (Zhao et al., 2018). By utilising word embeddings, these approaches capture more subtle signals on the syntactic as well as the semantic level of a word and its context. With such complex tasks as NER and NEN, this is an essential property to be able to distinguish very similar entities and entity types.

There are also approaches that do not separate entity recognition from normalization, e.g. (Lou et al., 2017), (Leaman et al., 2016) and (Zhao et al., 2018). Their advantage is that errors from NER are not propagated to the normalization step. By using feedback from normalization for NER it is possible to make corrections to the number of tokens that are combined to a mention because tokens could be falsely attributed to it. A comparison between these approaches might be interesting, yet it is outside the scope of this study project.

Approach

Method

As mentioned above, the normalization process is typically approached with these three separate steps: a) named entity recognition, b) identifiers retrieval, and c) re-ranking identifiers.

- *Named entity recognition*: This step identifies entity mentions. I will use the implementation stated in (Habibi et al., 2017).
- *Identifiers retrieval*: In this step, the mentions are checked against the list of all known entities (from the given dictionary) to find similar ones. The aim is to reduce the number of candidates drastically in a cost-efficient manner so that the remaining candidates can be ranked with a more complex analysis in the

next step. Here, the open-source search framework *Lucene* will be used to find syntactically similar names. It is fast and reliable and can be easily customised with different preprocessing steps. An approach by (Sennrich et al, 2015) that breaks up words into subunits based on byte pair encoding seems promising for finding meaningful and suitable tokens as basis for this search. These might otherwise be too short to capture meaningful patterns or too long to be valuable for generalisation. Another option is to take n-grams of different lengths, however, these are not calculated based on the contents of the text and might be less expressive.

This step is similar to how other tools approach normalization (except for applying byte pair encoding), however, here it is only used to filter the candidates for the next step.

- *Re-ranking identifiers*: In this third step, the most likely candidate is picked out of the remaining candidates. Similar to the first step, word embeddings will be used to calculate the likelihood of a match given the document context. The idea behind this is to compare the typical context of a candidate entity (from the training corpus) with the context of the found mention, based on the assumption that it will be most similar for the right match. Word embeddings are further explained in (Mikolov et al., 2013) and there are easily accessible tools as *word2vec* or *sent2vec* for calculating these embeddings from a corpus. The NCBI maintains a curated list of links between Entrez gene identifiers to PubMed document ids which allows calculating embeddings from these documents and associating these with the corresponding gene identifiers. The list provides more than 11 million links between genes and articles (Maglott et al., 2005).

Evaluation

Evaluation of NEN is complicated due to the scarcity of gold standard corpora, since it requires a lot of effort creating them. Methods are rarely tested on more than one corpus and usually without any cross-validation. This leads to a high risk of overfitting to the training documents and overestimating performance scores.

For evaluation, three freely available gold standard corpora from BioCreative tasks will be used. They were used for the second, third and fifth challenge respectively and all contain gene annotations which link mentioned entities to gene identifiers (Morgan et al., 2008) (Lu et al., 2011) (Pérez-Pérez et al., 2017). While the annotations for the second and fifth challenge refer to specific mentions, the third challenge only provides annotations on document-level.

PubMed and PMC (PubMed Central, which contains full-text articles) corpora as well as a collection of 4 million English Wikipedia articles are available for calculating word embeddings, for which no annotations are necessary. However, the NCBI provides a mapping between gene identifiers and PubMed documents which can be used to

calculate embeddings for specific identifiers. Although only covering a fraction of the documents (see Table 3), it can help to prevent difficulties that otherwise come up with homonyms. *Table 1* summarises the sizes and number of annotations of corpora from the BioCreative tasks, and Table 2 lists the sizes of the unannotated corpora.

With these test sets, the performance of the approach will be compared with the baseline methods to assess the impact of word embeddings and deep learning on normalization tasks. In addition, each of the three steps will be evaluated in isolation to find error sources for misclassified entities. Comparing approaches for intermediate results is more difficult, since the baseline methods only return the normalized identifiers.

Corpus	**Split**	**Size**	**Number of Annotations**
BioCreative II	Training	281 abstracts	640
BioCreative II	Test	262 abstracts	785
BioCreative III	Training	32 articles	607
BioCreative III Gold	Test	50 articles	1,669
BioCreative III Silver	Test	457 articles	7,709
BioCreative V GPRO	Training	12,600 abstracts	2,483
BioCreative V GPRO	Development	2,100 abstracts	515
BioCreative V GPRO	Test	6,300 abstracts	1,177

Table 1: Overview of available corpora with annotations.

Corpus	**Size**
PubMed	23,000,000 abstracts
PMC	700,000 articles
Wikipedia	4,000,000 articles

Table 2: Overview of unannotated corpora for calculating word embeddings.

Gene2pubmed Statistics	
Number of unique genes	6,139,875
Number of documents	1,176,142
Average number of documents per gene	1.82
Average&median number of genes per document	9.50 & 1
Average number of species per document	1.15

Table 3: Overview of NCBI's gene2pubmed mapping between gene identifiers and documents.

References

Erhardt, Ramón AA, Reinhard Schneider, and Christian Blaschke. "Status of text-mining techniques applied to biomedical text." *Drug discovery today* 11.7-8 (2006): 315-325.

Francis-Landau, Matthew, Greg Durrett, and Dan Klein. "Capturing semantic similarity for entity linking with convolutional neural networks." arXiv preprint arXiv:1604.00734 (2016).

Habibi, Maryam, et al. "Deep learning with word embeddings improves biomedical named entity recognition." Bioinformatics 33.14 (2017): i37-i48.

Hachey, Ben, et al. "Evaluating entity linking with Wikipedia." Artificial intelligence 194 (2013): 130-150.

Hakenberg, Jörg, et al. "Inter-species normalization of gene mentions with GNAT." Bioinformatics 24.16 (2008): i126-i132.

Huang, Minlie, Jingchen Liu, and Xiaoyan Zhu. "GeneTUKit: a software for document-level gene normalization." Bioinformatics 27.7 (2011): 1032-1033.

Leaman, Robert, and Zhiyong Lu. "TaggerOne: joint named entity recognition and normalization with semi-Markov Models." Bioinformatics 32.18 (2016): 2839-2846.

Leser, Ulf, and Jörg Hakenberg. "What makes a gene name? Named entity recognition in the biomedical literature." Briefings in bioinformatics 6.4 (2005): 357-369.

Lou, Yinxia, et al. "A transition-based joint model for disease named entity recognition and normalization." Bioinformatics 33.15 (2017): 2363-2371.

Lu, Zhiyong, et al. "The gene normalization task in BioCreative III." BMC bioinformatics 12.8 (2011): S2.

Lu, Zhiyong. "PubMed and beyond: a survey of web tools for searching biomedical literature." *Database* 2011 (2011).

Maglott, Donna, et al. "Entrez Gene: gene-centered information at NCBI." Nucleic acids research 33.suppl_1 (2005): D54-D58.

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Morgan, Alexander A., et al. "Overview of BioCreative II gene normalization." Genome biology 9.2 (2008): S3.

Mura, Cameron, Eli J. Draizen, and Philip E. Bourne. "Structural biology meets data science: Does anything change?." *Current opinion in structural biology* 52 (2018): 95-102.

Pérez-Pérez, Martin, et al. "Evaluation of chemical and gene/protein entity recognition systems at BioCreative V. 5: the CEMP and GPRO patents tracks." (2017): 11-18.

Rzhetsky, Andrey, Michael Seringhaus, and Mark Gerstein. "Seeking a new biology through text mining." *Cell* 134.1 (2008): 9-13.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).

Sun, Yaming, et al. "Modeling mention, context and entity with neural networks for entity disambiguation." Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu. "GNormPlus: an integrative approach for tagging genes, gene families, and protein domains." BioMed research international 2015 (2015).

Zhao, Sendong, et al. "A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization." arXiv preprint arXiv:1812.06081 (2018).