Statistics and Characteristics of Multi-word Entities in the Biomedical Domain across Entity Types

Tilia Renate Ellendorff[1]*, Fabio Rinaldi[1]

Abstract
We investigate statistics of the occurrence and linguistic properties of multi-word entities in the biomedical domain. For this we focus on a selection of frequent and representative biomedical entity types, including genes/proteins, chemicals, diseases and species/organisms. We use a previously developed large resource of controlled domain vocabulary as basis for our research. Our aim is to understand the differences between biomedical entity types in terms of numbers and distribution of multi-word entities and the reasons behind. In a second step we apply a dictionary look up using the multi-word entity resource. A thorough examination of linguistic properties will be presented for each entity type. This error analysis includes experimenting with an unsupervised approach for multi-word extraction in order to get further insight into multi-word mechanisms as well as an estimate of the differences in performance for each entity type.

Relevant WGs: WG 3 (Statistical, Hybrid and Multilingual Processing of MWEs)

[1] *Institute of Computational Linguistics, Universität Zürich, Zurich, Switzerland*

1. Introduction and Motivation

Multi-word expression remain one of the main challenges to natural language processing (NLP). Even though research has advanced in the recent years, a big amount of work still has to be invested. The work presented with this poster investigates biomedical entity names as a category that is central to biomedical textmining but to a very large degree consists of multi-word entities (MWEs). A deeper understanding of the distribution and characteristics of biomedical MWEs across entity types supports the development of approaches that are able to deal with MWEs in a more efficient way. Therefore, knowledge of linguistic features of MWE has the power to improve biomedical textmining at its core.

Past research on multi-word entities in the biomedical domain has focused on term extraction and named entity recognition [1, 2], where processing of MWEs is usually treated as a byproduct but only rarely as the main focus of research. Furthermore, research on biomedical MWEs usually does not consider different entity types but either focuses on one entity type in separation or, even more frequently, tries to find a solution that fits for all possible biomedical entities. However, it is crucial to consider multi-word behavior and characteristics separately for each entity type.

We investigate the distribution and the characteristics of MWEs over a large collection of controlled domain vocabulary. Presented statistics include multi-word counts per entity type and distribution of word numbers per MWEs for each entity type in focus. A dictionary look-up allows us to map the entities of the resource to text. Thereafter, a thorough error analysis provides further evidence concerning the individual multi-word properties of the entity types. As part of the linguistic analysis, we perform a set of experiments combining the dictionary look-up with an unsupervised multi-word expression extraction approach. This approach uses word embeddings which we generated using the *Phrases* module which is part of the word2vec distribution by gensim. We used the data of the whole Pubmed release of January 2016 for MWEs consisting of two or three tokens. Apart from reaching a deeper linguistic understanding, our purpose is to get insights in the efficiency of unsupervised methods for the extraction of multi-word entities across the different entity types.

2. Materials and Methods

For studying multi-word entities in the biomedical domain we use a previously published large database of biomedical terminology [3]. This resource was compiled from the terminology of a selection of biomedical databases and ontologies. The chosen terminologies represent the most widely used controlled vocabularies for the biomedical domain. We consider a subset of the originally compiled terminology resource. Resources included in the subset are NCBI[2] (Entrez Gene, Taxonomy), MeSH[3] (Chemicals, Diseases, Psychiatric terms, Organisms), ChEBI[4] (Chemicals). We calculated statistics of multi-word entity types for this subset. Furthermore, the dictionary of multi-word entities used in further experiments consists of all MWEs in the subset. As the resource is regularly updated with the newest versions of the original

[1] All Numbers as of June 4, 2016.

[2] http://www.ncbi.nlm.nih.gov/

[3] http://www.ncbi.nlm.nih.gov/mesh

[4] https://www.ebi.ac.uk/chebi/

	Tokens			Types		
Entity Type	**Number Entries (Tokens)**	**Number MWEs (Tokens)**	**Percent MWEs**	**Number Entries (Types)**	**Number MWEs (Types)**	**Percent MWEs**
Gene/Protein	14,199,460	18,639	0.13 %	11,272,054	16,289	0.14 %
Organism/Species	1,407,664	1,371,300	97.4 %	1,359,712	1,342,423	98.7 %
Diseases	47,617	41,904	88.0 %	37,774	32,829	86.9 %
Psychiatry	1,149	1,019	88.7 %	1,134	1,008	88.9 %
Chemicals/Drugs	1,074,680	653,205	60.8 %	834,886	501,373	60.1 %
All Entity Types	16,730,570	2,086,067	12.5 %	12,285,560	1,894,022	15.5 %

Table 1. Overview of Token, Type and Multi-word counts for each Entity Type[1]

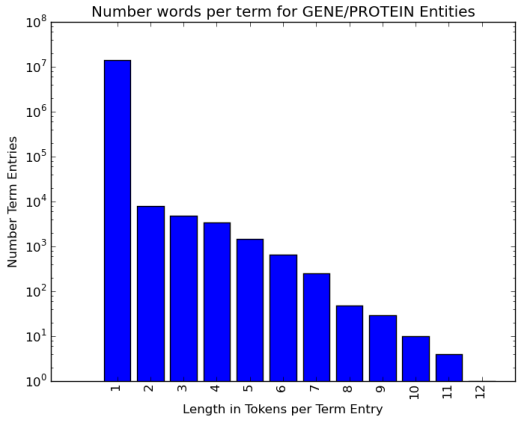

Figure 1. Number of Words per Term for GENE/PROTEIN Entities

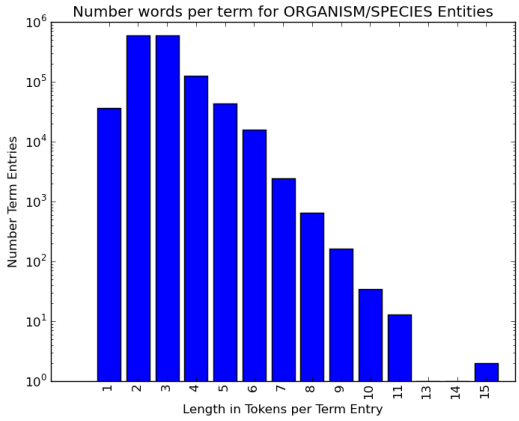

Figure 2. Number of Words per Term for ORGANISM/SPECIES Entities

databases and ontologies, we were able to work with the most current version of the domain vocabulary.

We counted multi-word entities for each entity type present in the terminology subset and calculated statistics accordingly. An overview of these statistics can be seen in Table 1. Furthermore, we counted the distributions of numbers of words per entity name for each entity type.

For building the word embeddings, we used the well-known word2vec Python package[5], described in [4]. We train 3 phrase-based models for bi-, tri- and quadgram recognition. These models are trained over the whole of PubMed[6] which includes more than 26 Million abstracts[7]. We pre-processed these abstracts by applying sentence-splitting and tokenization. Finally, we use the models to discover bi-, tri- and quadgram MWEs which we compare to the multi-word expressions found by the dictionary look-up. These are then analyzed by their specific entity type in order to provide a list of linguistic features which are responsible for being able to discover an entity type by one of the two approaches. Linguistic features give important clues as to which aspects need to be taken into account in future approaches for named entity recognition in the biomedical domain.

[5]https://radimrehurek.com/gensim/models/word2vec.html
[6]As of January 2016
[7]http://www.ncbi.nlm.nih.gov/pubmed

3. Results and Discussion

Two graphs representing the differences between gene/protein names and organisms/species as the two extreme entity types can be seen in Figures 1 and 2. Whereas MWEs are not as common for genes/proteins, they exist and can contain up to 11 words. On the other hand for organisms and species, MWEs are by far more common than one-word entities. These two examples give an idea of how different MWE distribution is across biomedical entity types. For reasons of space, graphs for the other examined entity types had to be excluded from this abstract but they are part of the poster presentation. Concerning the experiments with word embeddings for MWE extraction, our findings suggest that unsupervised methods for extracting MWEs, have a different degree of efficiency depending on the specific type of an entity. This is due to a range of characteristics as a in-depth linguistic analysis reveals. For reasons of space, presentations of results for the experiments with the dictionary look-up and unsupervised multi-word entity discovery using word-embeddings are not presented here, but are part of the poster presentation.

References

[1] Joachim Wermter and Udo Hahn. Effective grading of termhood in biomedical literature. In *AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA, October 22-26, 2005*, 2005.

[2] Lejun Gong, Ronggen Yang, Jiacheng Feng, and Geng Yang. A combined approach for the extraction of the multi-word and nested biomedical entity. In *DSP*, pages 708–711. IEEE, 2015.

[3] Tilia Renate Ellendorff, Adrian Van der Lek, Lenz Furrer, and Fabio Rinaldi. A combined resource of biomedical terminology and its statistics. In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence, Universidad de Granada, Granada, Spain, November 4-6, 2015.*, pages 39–49, 2015.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.