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

Tilia Renate Ellendorff, Fabio Rinaldi

University of Zurich, Institute of Computational Linguistics

ellendorff@cl.uzh.ch

University of Zurich[UZH]

OntoGene
www.ontogene.org

Motivation:

Multi-word Entities in the Biomedical Domain

- Multi-word entities (MWEs) are very common for most frequent biomedical entity types
- Processing multi-word entities is normally only treated as a by-product in biomedical textmining
- Multi-word entities have different characteristics across different entity types

⇒ **An awareness of differences between entity types is helpful**

Overview

Approaches

- Experiments with a large database of biomedical terminology
- Experiments with Multi-word word embeddings (Word2Vec Phrases)

Included Entity Types

Genes/Proteins, Organisms/Species, Diseases, Psychiatry, Chemicals/Drugs

Experiments with a large Database of Biomedical Entities

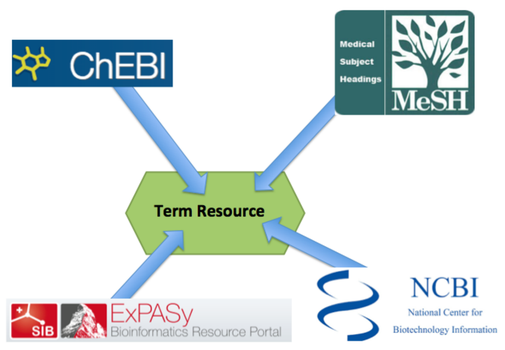

Figure 1: A large database of biomedical terminology compiled from different resources [2]

Distribution of Multi-word Entities

	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 (Numbers as of June 2016)

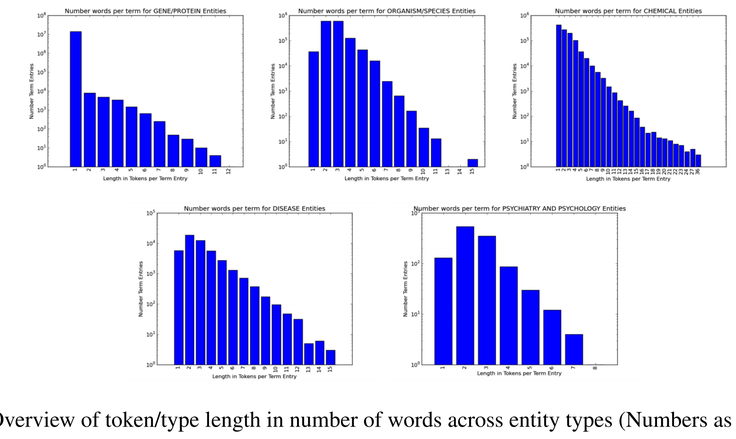

Figure 2: Overview of token/type length in number of words across entity types (Numbers as of June 2016)

Annotation Experiments: Results

Annotation Experiments using 10,000 random PubMed abstracts with a dictionary look-up.

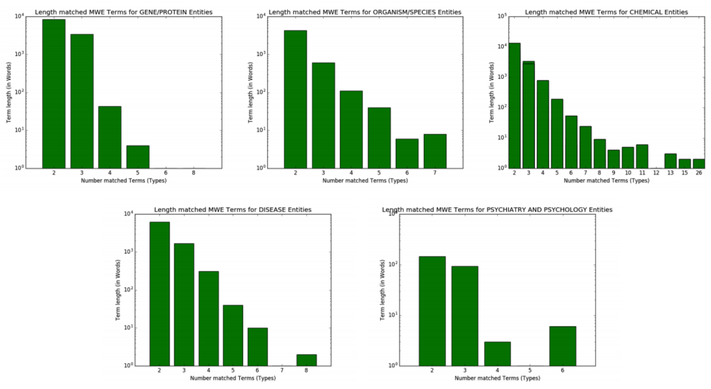

Figure 3: Overview of mwe length in number of annotated words (types) across entity types (Numbers as of June 2016)

Experiments with Word Embeddings

- Word2Vec Phrases[5] model trained over the whole of PubMed (as of January 2016)
- Word2Vec Phrases includes collocation detection: $\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$ [5]
- Bi-, tri- and quad-gram models trained (word2vec skip-gram)
- Concept and relation-matching against the Unified Medical Language System (UMLS) [1]

Method flowchart

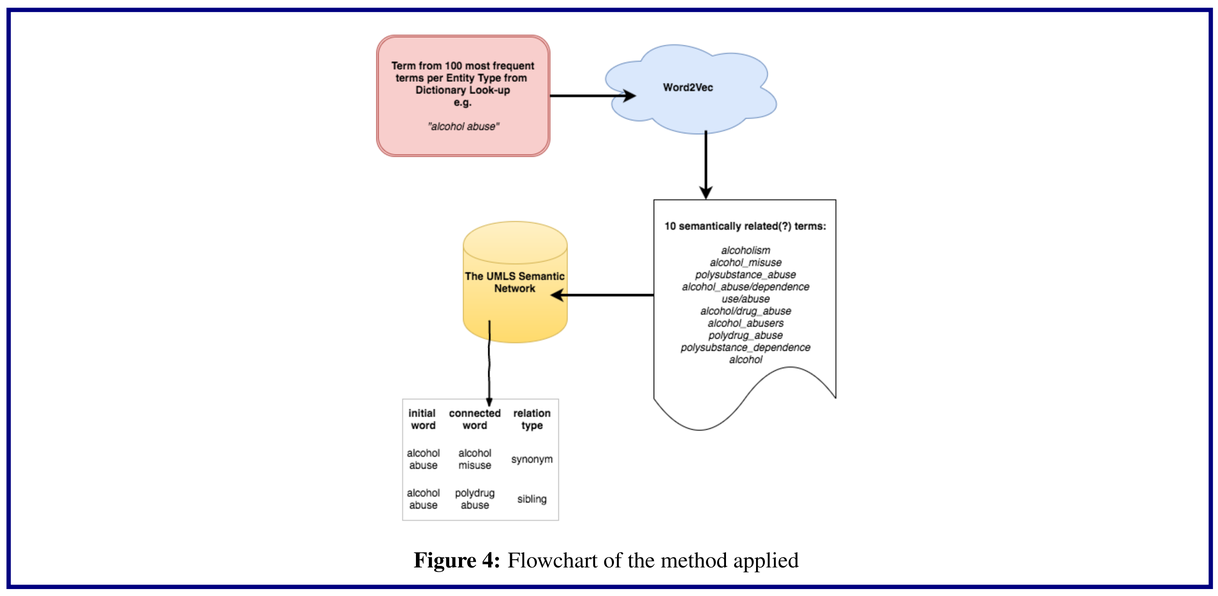

Figure 4: Flowchart of the method applied

Results

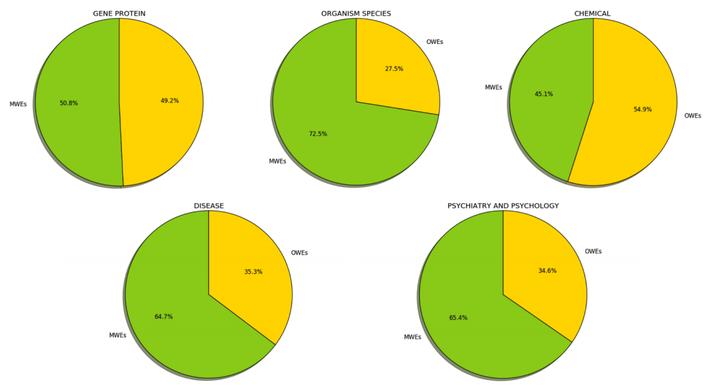

Figure 5: Percentages of semantically related one-word terms as compared to multi-word terms per entity type

	Relation Types										
Entity Type	Synonyms	SIB	CHD	PAR	broader	narrower	similar	related	other	not found	no relation
Gene/Protein	24	9	2	4	2	1	0	0	19	11 + 179	49
Organism/Species	62	53	5	27	20	1	0	6	30	5 + 223	275
Chemical	72	58	33	23	21	38	0	11	29	4 + 369	118
Disease	39	72	23	16	13	19	4	30	48	0 + 208	42
Psychiatry and Psychology	13	48	8	22	19	7	1	26	25	0 + 120	66

Table 2: Overview of relations (from the UMLS) that were detected between semantically related words from word embeddings
Note: it is possible that there more than two relation types between two concepts.

	Relation Types										
Entity Type	Synonyms	SIB	CHD	PAR	broader	narrower	similar	related	other	not found	no relation
Gene/Protein	20	8	0	0	0	1	0	0	9	11 + 82	21
Organism/Species	55	46	1	3	2	1	0	2	22	5 + 138	222
Chemical	44	21	11	4	5	13	0	2	11	4 + 178	43
Disease	26	43	15	5	2	11	3	16	18	0 + 145	26
Psychiatry and Psychology	10	24	3	8	6	4	0	13	15	0 + 86	42

Table 3: Overview of relations (from the UMLS) that were detected between semantically related words from word embeddings - Relations between MWEs only

Future Work

- Experiments with different methods of collocation detection
- Include additional entity types
- Systematic discussion/interpretation of results

References

[1] Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology, 2004.

[2] 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.

[3] 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.

[4] Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, and Hae-Chang Rim. Biomedical named entity recognition using two-phase model based on {SVMs}. *Journal of Biomedical Informatics*, 37(6):436 – 447, 2004. Named Entity Recognition in Biomedicine.

[5] 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.

[6] L. Venkata Subramaniam, Sougata Mukherjea, Pankaj Kankar, Biplav Srivastava, Vishal S. Batra, Pasumarti V. Kamesam, and Ravi Kothari. Information extraction from biomedical literature: Methodology, evaluation and an application. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 410–417, New York, NY, USA, 2003. ACM.

[7] 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.