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

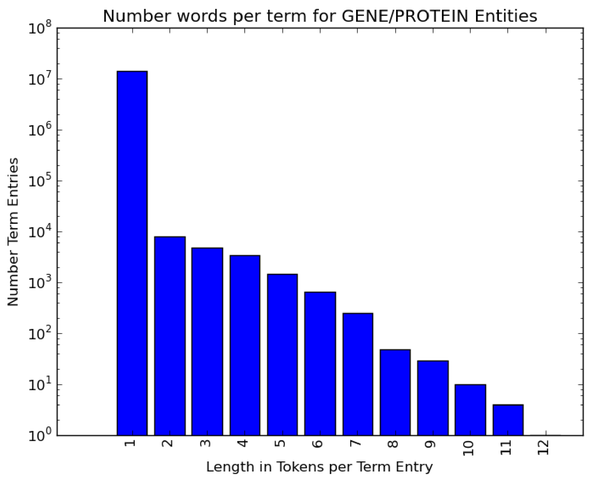
Tilia Renate Ellendorff, Fabio Rinaldi
Institute of Computational Linguistics, Universität Zürich

University of Zurich[UZH]

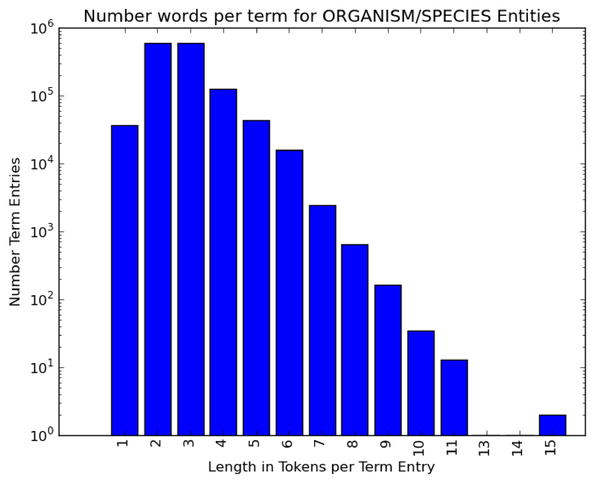
Motivation

- In the biomedical domain multi-word entity names are very common
- A deeper understanding of their characteristics is necessary
- Differences of occurrence and linguistic properties among entity types
- Statistics for frequent entity types: chemicals, protein/genes, diseases, ...
- Comparison of dictionary look-up and word embeddings

Differences between Entity Types



Number of Words per Term for GENE/PROTEIN Entities



Number of Words per Term for ORGANISM/SPECIES Entities

- Awareness of differences in distribution and usage between entity types

Methods

Statistics over a large Term Resource:

- Token counts and distributions per entity type over a large term resource
- Percentages of multi-word entities very different over entity types

Dictionary look-up:

- Matching terminology against a collection of 10,000 PubMed abstracts
- Statistics for different entity types

Word Embeddings:

- Word embedding model built with Word2Vec Phrases
- Effectiveness check using UMLS semantic network
- Illustration of characteristics per entity type