An Evaluation of Definition Paradigms in Lexicography for Word Sense Alignment

Sina Ahmadi, John P. McCrae

Data Science Institute, National University of Ireland Galway
E-mail: firstname.lastname@insight-centre.org

Sense definitions are principal components of monolingual dictionaries describing various meanings of words in plain text. Since antiquity, there have been many theories and discussions on how to define a concept, i.e., *definiendum*, and the words and phrases which are used for this purpose, i.e., *definiens*. Durkin (2016) provides a description of such theories from historical, logical, and lexicographical points of view.

Dictionaries, as crucial resources for documenting languages, have been widely used in language technology and natural language processing. Given the increasing number of lexico-semantic resources thanks to community-driven initiatives such as Wiktionary[5] and Open Multilingual WordNet, the alignment of such resources is of importance to promote interoperability and facilitate the integration of various resources in a viable manner.

In the context of the word sense alignment task where word definitions are aligned automatically, we assume that retrieving the composing parts of sense definitions is useful to facilitate the alignment tasks. To this end, we carry out an evaluation of two analytical and relational paradigms on the MWSA English data (Ahmadi et al., 2020) containing annotated glosses of Webster's Dictionary 1913 and Princeton WordNet. The paradigms are defined as follows:

- Analytical definitions define a formal descriptive sentence consisting of four main components, namely species, verb, genus, and differentia.
- Relational definitions explain the meaning of a word in comparison to other entities, e.g., "extraneous (adjective)" defined as "not belonging to a thing" in Webster 1913.

As a preliminary study, we use a pattern-based approach as proposed by (Westerhout, 2010) where definitions are analyzed to retrieve genus and entity. This task is carried out using regular expressions with functional keywords, such as "opposite of", "belonging to" or "of or pertaining". Given definitions of a lemma with its part-of-speech in two dictionaries, in our case Webster 1913 and Princeton WordNet, two definitions are to be

[5] https://www.wiktionary.org/

aligned if they have an identical genus or entity after lemmatization.

Among the English aligned sense definitions, we select 100 definitions randomly among which 50 are alignable, i.e., specified with exact, and the rest are non-alignable, i.e., specified by none. Although non-alignable definitions are correctly classified in all cases, only 13 among the 50 other definitions are classified correctly. This indicates the poor performance of the pattern-based or symbolic approach for this task.

In addition to different phrase structures which lead to an unsimilar syntactic analysis, lexical choice determines the genus and entity of each definition. For instance, the definition of "angulation (noun)" as "the act of making angulate" and "making angular" use two semantically-related but different words "angular" and "angulate". On the other hand, descriptive phrases are missing in many definitions, as for "usurpation (noun)" defined as "wrongfully seizing" and "the act of usurping". It should also be noted that definitions are not of same granularity across resources.

Finally, we believe that such challenges which are faced in computational lexicography and natural language processing should be of interest to lexicographers and community-driven lexical content creators. This way, further computer-assisted techniques may be more efficiently integrated in the process of dictionary creation and compilation.

Keywords: electronic lexicography; natural language processing; lexical resource alignment

References

Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S., ... & Gabrovsek, D. (2020, May). *A multilingual evaluation dataset for monolingual word sense alignment*. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 3232-3242).

Durkin, P. (Ed.). (2016). *The Oxford handbook of lexicography*. Oxford University Press.

Grosse, J., & Saurí, R. (2020). *Principled Quality Estimation for Dictionary Sense Linking*. In Proceedings of the XIX EURALEX conference.

Westerhout, E. (2010). *Definition extraction for glossary creation: a study on extracting definitions for semi-automatic glossary creation in Dutch*. Netherlands Graduate School of Linguistics.