Lexical Sense Alignment using Weighted Bipartite b-Matching

Sina Ahmadi Mihael Arcan John McCrae
sina.ahmadi@insight-centre.org

Insight
Centre for Data Analytics

Introduction

Lexical resources are important components of natural language processing (NLP) applications providing linguistic information about the vocabulary of a language and the semantic relationships between words. While there is an increasing number of lexical resources, manual construction and maintenance of such resources is a cumbersome task. This can be efficiently addressed by NLP techniques.

Given various types of resources, aligned resources will **improve word, knowledge and domain coverage** and **increase multilingualism** by creating new lexical resources.

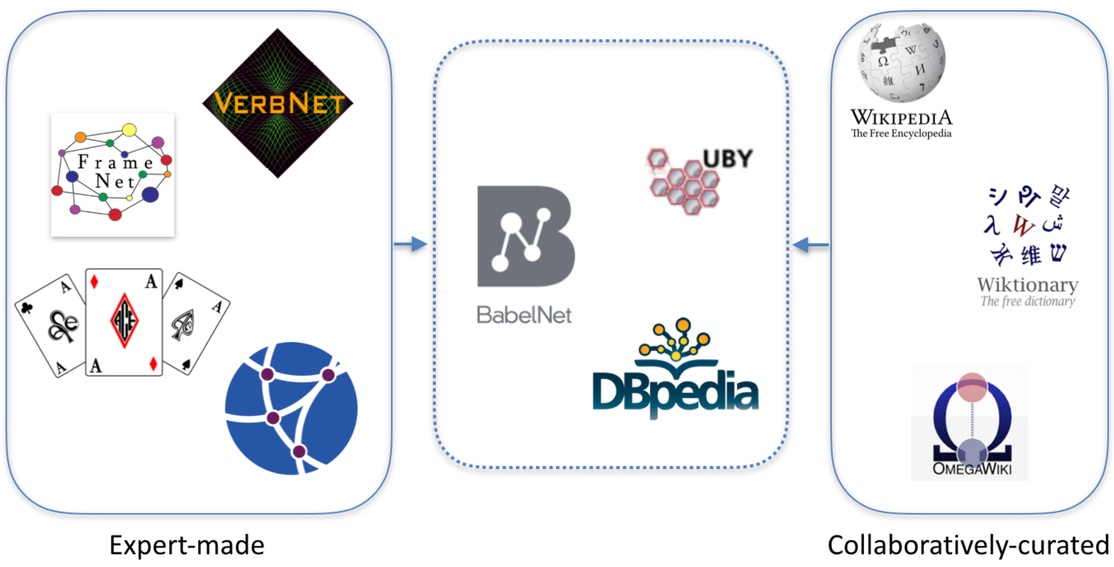

Figure: An example of various types of resources

Objectives

One of the current challenges in aligning lexical data across different resources is **word sense alignment (WSA)**. Different monolingual resources may use **different wordings** and **structures**, with dissimilar **level of granularity** for the same concepts and entries.

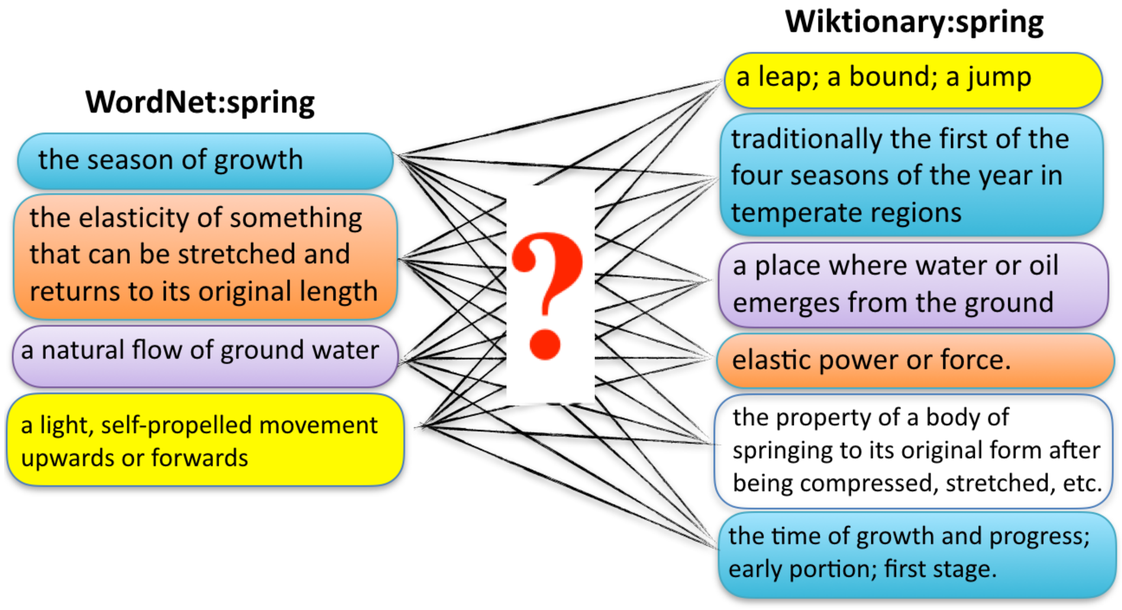

Figure: A few senses of the word *spring* (n) in WordNet and Wiktionary

Our objective in the current study is to **enhance WSA with respect to polysemous items.**

Method

We present a similarity-based approach which relies on **semantic** and **textual similarity** and **a graph matching algorithm**. Transforming the alignment problem into a bipartite graph matching, where **nodes and edges respectively represent senses and links** between them, enables us to apply graph matching algorithms, in particular, **weighted bipartite b-matching (WBbM)**.

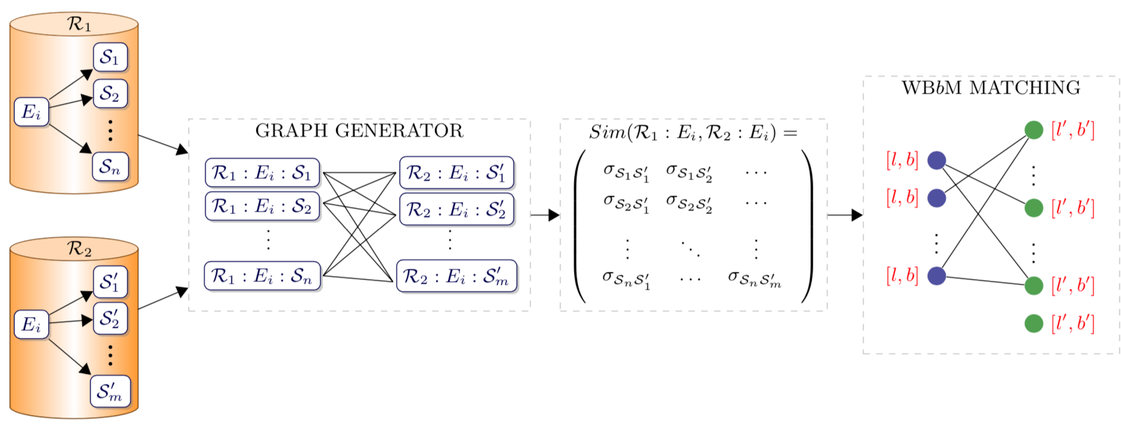

Figure: Our sense alignment system

WBbM aims at providing a **more diversified matching** where a node may be connected to a **certain number of nodes** (not only one), determined by lower and upper bound functions L and B. Given such a configuration, WBbM finds the matching which **maximizes the overall weight**.

Evaluation

We evaluate the performance of our approach on aligning sense definitions in **WordNet and Wiktionary** [1]. Our approach delivers superior results in comparison to the baseline results [2].

Left bound, right bound	P_{macro}	R_{macro}	F_{avg}	A_{avg}
[0, 1], [0, 1]	**81.86**	61.83	68.51	69.48
[0, 2], [0, 1]	78.13	70.74	73.28	76.57
[0, 3], [0, 1]	77.88	71.38	73.59	77.13
[1, 2], [1, 2]	81.21	74.17	76.59	79.49
[1, 3], [1, 3]	81.26	75.02	77.12	80.14
[1, 5], [0, 1]	81.25	**75.25**	**77.28**	**80.33**
[1, 5], [1, 2]	81.25	75.23	77.26	80.32

Table: WBbM algorithm performance on alignment of WordNet and Wiktionary

- High precision, high recall
- Efficient in linking polysemous items
- Still difficult to tune the parameters

WBbM is available at `/sinaahmadi/Bipartite_b_matching`.

References

1. Meyer, Christian M., and Iryna Gurevych. "What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage." *Proceedings of 5th International Joint Conference on Natural Language Processing*. 2011.
2. McCrae, John P., and Paul Buitelaar. "Linking Datasets Using Semantic Textual Similarity." *Cybernetics and Information Technologies* 18.1 (2018): 109-123.

A World Leading SFI Research Centre

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

DCU · NUI Galway OÉ Gaillimh · UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh · UCD Dublin · Science Foundation Ireland sfi For what's next