Lexical Sense Alignment using Weighted Bipartite b-Matching

Sina Ahmadi
Insight Centre for Data Analytics, Data Science Institute
National University of Ireland Galway
sina.ahmadi@insight-centre.org

Mihael Arcan
Insight Centre for Data Analytics, Data Science Institute
National University of Ireland Galway
mihael.arcan@insight-centre.org

John P. McCrae
Insight Centre for Data Analytics, Data Science Institute
National University of Ireland Galway
john.mccrae@insight-centre.org

1 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 the words. While there is an increasing number of lexical resources, particularly expert-made ones such as WordNet [8] or FrameNet [2], as well as collaboratively-curated ones such as Wikipedia[1] or Wiktionary[2], manual construction and maintenance of such resources is a cumbersome task. This can be efficiently addressed by NLP techniques. Aligned resources have shown to improve word, knowledge and domain coverage and increase multilingualism by creating new lexical resources such as Yago [13], BabelNet [9] and ConceptNet [12]. In addition, they can improve the performance of NLP tasks such as word sense disambiguation [10], semantic role tagging [15] and semantic relations extraction [14].

2 Objective

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 for the same concepts and entries. There are various approaches in aligning definitional texts based on semantic similarity and linking techniques. For instance, Meyer and Gurevych [7] use semantic similarity and Personalized PageRank (PPR) to estimate the semantic relatedness in linking Wiktionary and WordNet. Pilehvar and Navigli [11] go beyond the surface form semantic similarity by transforming resources into semantic networks. Differently, Matuschek and Gurevych [5] present Dijkstra-WSA algorithm which aligns word senses using Dijkstra's shortest path algorithm.

In this study, we present a similarity-based approach for WSA in English WordNet and Wiktionary with a focus on the polysemous items. Our approach relies on semantic textual similarity using features such as string distance metrics and word embeddings, and a graph matching algorithm. Transforming the alignment problem into a bipartite graph matching enables us to apply graph matching algorithms, in particular, weighted bipartite b-matching (WBbM).

[1] `https://www.wikipedia.org/`
[2] `https://www.wiktionary.org/`

LDK 2019 - Posters Track.
Editors: Thierry Declerck and John P. McCrae

3 Method

WBbM is one of the widely studied classical problems in combinatorial optimization for modeling data management applications, e-commerce and resource allocation systems [1, 3, 4]. WBbM is a variation of the weighted bipartite matching, also known as assignment problem. In the assignment problem, the optimal matching only contains one-to-one matching with the highest weight sum. This bijective mapping restriction is not realistic in the case of lexical resources where an entry may be linked to more than one entries. Therefore, WBbM aims at providing a more diversified matching where a node may be connected to a certain number of nodes. Formally, given $G = ((U, V), E)$ with weights W and vertex-labelling functions $L : U \cup V \rightarrow \mathbb{N}$ and $B : U \cup V \rightarrow \mathbb{N}$, WB$b$M finds a subgraph $H = ((U, V), E')$ which maximizes $\sum_{e \in E'} W(e)$ having $u \in [L(u), B(u)]$ and $v \in [L(v), B(v)]$. In other words, the number of the edges that can be connected to a node is determined by the lower and upper bound functions L and B, respectively.

Algorithm 1: Greedy WBbM

Input: $G = ((U, V), E, W)$, bounds L and B

Output: $H = ((U, V)), E', W)$ satisfying bound constraints with a greedily-maximized score $\sum_{e \in E'} W(e)$

$E' = \emptyset$

Sort E by descending W(e)

for e **to** E **do**

    if $H = ((U, V)), E' \cup \{e\}, W)$ *does not violate L and B* **then**

        $E' = E' \cup \{e\}$

return $H = ((U, V)), E', W)$

Algorithm 1 presents the WBbM algorithm with a greedy approach where an edge is selected under the condition that adding such an edge does not violate the lower and the upper bounds, i.e. L and B.

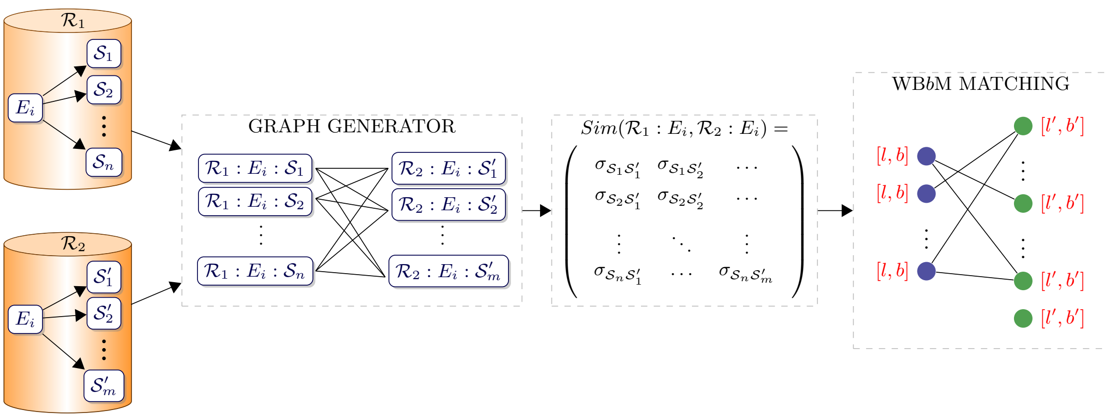

Figure 1 Sense alignment system

We evaluate the performance of our approach on aligning sense definitions in WordNet and Wiktionary using an aligned resource presented by Meyer and Gurevych [7]. Given an identical entry in English WordNet and Wiktionary, we first convert the senses to a bipartite graph where each side of the graph represents the senses belonging to one resource. Then, we extract the similarity scores between those senses using a similarity function. The similarity function is a trained model based on similarity features such as word length ratio, longest common subsequence, Jaccard measure, word embeddings and forward precision, which is performed by NASIC [6]. And finally, the senses in the the weighted bipartite graph are matched by the WBbM algorithm. This process is illustrated in Figure 1 where senses of entry E_i in resource $\mathcal{R}_1$, $\{S_1, S_2, ..., S_n\}$, are aligned with the senses of the same entry in $\mathcal{R}_2$, $\{S_1^{'}, S_2^{'}, ..., S_n^{'}\}$. The lower and upper bounds of the right side and left side of the graph, respectively $[l, b]$ and $[l', b']$, are the parameters to be tuned.

4 Evaluation

In order to evaluate the performance of our alignment approach, we calculated macro precision P_{macro}, macro recall R_{macro}, average F-measure F_{avg} and average accuracy A_{avg} as follows:

$$P = \frac{TP}{TP + FP} \qquad P_{macro} = \frac{1}{|E|}\sum_{i=1}^{|E|}\frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$R = \frac{TP}{TP + FN} \qquad R_{macro} = \frac{1}{|E|}\sum_{i=1}^{|E|}\frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$F = 2 \times \frac{P \times R}{P + R} \qquad F_{avg} = \frac{1}{|E|}\sum_{i=1}^{|E|} F_i \tag{3}$$

$$A_{avg} = \frac{1}{|E|}\sum_{i=1}^{|E|}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{4}$$

where E refers to the set of entries, TP, TN, FN and FP respectively refer to true positive, true negative, false negative and false positive.

Table 1 provides the evaluation results using the WBbM algorithm with different combinations of the matching bounds over the left side (WordNet senses) and the right side (Wiktionary senses) of the alignment graph. We observe that a higher upper bound increases the recall. On the other hand, setting the lower bound to 1 provides a higher precision, while parameters with a lower bound of 0, e.g. [0, 3], lack precision. Note that [0, 1] parameter performs similarly as a bijective mapping algorithms such as the assignment problem where a node can be only matched to one node. Our approach delivers superior results in comparison to the baseline results provided by McCrae and Buitelaar [6].

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 1 WBbM algorithm performance on alignment of WordNet and Wiktionary

5 Conclusion

We revisited WordNet-Wiktionary alignment task and proposed an approach based on textual and semantic similarity and WBbM algorithm. We demonstrated that this approach is efficient for aligning resources in comparison to the baseline results thanks to the flexibility of the matching algorithm. However, tuning the parameters of the matching algorithm needs further investigations of the resource and is not following a rule. In future work, we plan to extend our experiments to more resources.

Acknowledgements

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

References

1 Faez Ahmed, John P Dickerson, and Mark Fuge. Diverse weighted bipartite b-matching. *arXiv preprint arXiv:1702.07134*, 2017.

2 Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

3 Cheng Chen, Sean Chester, Venkatesh Srinivasan, Kui Wu, and Alex Thomo. Group-aware weighted bipartite b-matching. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 459–468. ACM, 2016.

4 Cheng Chen, Lan Zheng, Venkatesh Srinivasan, Alex Thomo, Kui Wu, and Anthony Sukow. Conflict-aware weighted bipartite b-matching and its application to e-commerce. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1475–1488, 2016.

5 Michael Matuschek and Iryna Gurevych. Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164, 2013.

6 John P McCrae and Paul Buitelaar. Linking datasets using semantic textual similarity. *Cybernetics and Information Technologies*, 18(1):109–123, 2018.

7 Christian M Meyer and Iryna Gurevych. What psycholinguists know about chemistry: Aligning Wiktionary and Wordnet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892, 2011.

8 George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

9 Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
10 Roberto Navigli and Simone Paolo Ponzetto. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1399–1410. Association for Computational Linguistics, 2012.
11 Mohammad Taher Pilehvar and Roberto Navigli. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 468–478, 2014.
12 Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017.
13 Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.
14 Robert S Swier and Suzanne Stevenson. Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 883–890. Association for Computational Linguistics, 2005.
15 Nianwen Xue and Martha Palmer. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.