Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation

Federico Scozzafava
Sapienza University of Rome

Marco Maru
Sapienza University of Rome

Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation

Federico Scozzafava[1,2], **Marco Maru**[1,2,3], **Fabrizio Brignone**[4],
Giovanni Torrisi[4], and **Roberto Navigli**[1,2]

[1]Sapienza NLP Group
[2]Department of Computer Science, Sapienza University of Rome
[3]Department of Literature and Modern Cultures, Sapienza University of Rome
[4]Babelscape, Italy
firstname.lastname@uniroma1.it, lastname@babelscape.com

Abstract

Exploiting syntagmatic information is an encouraging research focus to be pursued in an effort to close the gap between knowledge-based and supervised Word Sense Disambiguation (WSD) performance. We follow this direction in our next-generation knowledge-based WSD system, SyntagRank, which we make available via a Web interface and a RESTful API. SyntagRank leverages the disambiguated pairs of co-occurring words included in SyntagNet, a lexical-semantic combination resource, to perform state-of-the-art knowledge-based WSD in a multilingual setting. Our service provides both a user-friendly interface, available at http://syntagnet.org/, and a RESTful endpoint to query the system programmatically (accessible at http://api.syntagnet.org/).

1 Introduction

In Natural Language Processing, Word Sense Disambiguation (WSD) is an open problem concerning lexical ambiguity. It is aimed at determining which sense – among a finite inventory of many – is evoked by a given word in context (Navigli, 2009).

This challenge has been tackled by exploiting huge amounts of hand-annotated data in a supervised fashion (Raganato et al., 2017b; Bevilacqua and Navigli, 2019; Vial et al., 2019; Bevilacqua and Navigli, 2020) or, alternatively, by harnessing structured information (Agirre et al., 2014; Moro et al., 2014; Scarlini et al., 2020), such as that available within existing lexical knowledge bases (LKBs) like WordNet (Fellbaum, 1998). Despite achieving better overall results, supervised systems require tremendous efforts in order to produce data for several languages (Navigli, 2018; Pasini, 2020), whereas knowledge-based approaches can easily be applied in multilingual environments due to the wide array of languages covered by LKBs like BabelNet[1] (Navigli and Ponzetto, 2012), or the Open Multilingual WordNet (Bond and Foster, 2013). Moreover, it is widely acknowledged that the performance of a knowledge-based WSD system is strongly correlated with the structure of the LKB employed (Boyd-Graber et al., 2006; Lemnitzer et al., 2008; Navigli and Lapata, 2010; Ponzetto and Navigli, 2010). In fact, the knowledge available within LKBs reflects the fact that words can be linked via two types of semantic relations: paradigmatic relations – i.e. the most frequently encountered relations in LKBs – concern the substitution of lexical units, and determine to which level in a hierarchy a language unit belongs by semantic analogy with units similar to it; conversely, syntagmatic relations concern the positioning of such units, by linking elements belonging to the same hierarchical level (e.g., words), which appear in the same context (e.g., a sentence). As a case in point, a paradigmatic relation exists, independently of a given context, between the words *farm*$_n$ and *workplace*$_n$ (where a farm is a type of workplace), whereas a syntagmatic relation is entertained between the words *work*$_v$ and *farm*$_n$, e.g., in the sentence *'her husband works in a farm as a labourer.'*

In our most recent study (Maru et al., 2019, SyntagNet), we provided further evidence that the nature of LKBs impacts on system performance: the injection of syntagmatic relations – in the form of disambiguated pairs of co-occurring words – into an existing LKB biased towards paradigmatic knowledge enables knowledge-based systems to rival their supervised counterparts.

To make the above results accessible to the research community, in this paper we introduce a Web interface and a RESTful API for SyntagRank, our multilingual WSD system, which applies the

[1] https://babelnet.org/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 37–46
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

Personalized PageRank (PPR) algorithm (Haveliwala, 2002) to an LKB made up of WordNet, the Princeton WordNet Gloss Corpus (PWNG) and the lexical-semantic syntagmatic combinations available in the SyntagNet resource. SyntagRank is the first system to perform multilingual WSD by leveraging an underlying LKB connecting a sizeable amount of syntagmatically-related concepts.

2 Preliminaries

Our disambiguation algorithm relies on an LKB, i.e. a graph in which each node represents a concept, and each connection between nodes represents a semantic relation. In this Section we describe the LKBs whose resulting union we use as our reference graph, and then go on to provide details of the PPR algorithm.

2.1 Lexical Knowledge Bases

WordNet (Fellbaum, 1998) is a lexical-semantic database of English, in which concepts are expressed by means of sets of cognitive synonyms (synsets) that are interlinked to form a semantic network through relation edges.

Relations in WordNet are mainly of a hierarchical, and thus paradigmatic nature, with the most frequently encoded relation being the super-subordinate relation (instantiated in terms of hypernymy and hyponymy; see also Section 1). Other relations linking concepts in WordNet include part-whole relations (meronymy, e.g. between *wheel*$_n$ and *car*$_n$), antonymy relations and cross-part-of-speech relations holding among semantically similar words sharing a stem with the same meaning (e.g. between *speed*$_n$ and *speedy*$_a$). As of today, WordNet is the most widely used and *de facto* standard sense inventory for the WSD task (Raganato et al., 2017a).

Princeton WordNet Gloss Corpus (PWNG) is the semantically-annotated gloss corpus made available by WordNet since its 3.0 release.[2] Glosses are short definitions providing proper meanings for synsets, and in PWNG they have been tagged according to the senses in WordNet. Following Agirre et al. (2014), we induce new WordNet relations from PWNG by linking the synset to which the gloss refers to each of the synsets that have been tagged in the gloss itself. In this way, additional contextual relations are provided, inadvertently covering syntagmatic relations, too.

SyntagNet (Maru et al., 2019) is a database containing almost 90,000 pairs of manually-disambiguated lexical collocations and free word associations. Pairs in SyntagNet link nouns to other nouns or verbs tagged according to the WordNet 3.0 sense inventory and such pairs can therefore be exploited as new relation paths connecting nodes (synsets) in a WordNet-based semantic network. For our purposes, we are especially interested in the fact that SyntagNet is the only high-quality resource to systematically provide syntagmatic information in the form of lexical-semantic combinations. This kind of information becomes particularly valuable when used to enrich semantic networks otherwise biased towards paradigmatic knowledge, by creating direct routes between those concepts whose lexicalizations tend to appear together in the same context more often than by mere chance.

2.2 Personalized PageRank

The original PageRank (Brin and Page, 1998) is an algorithm which uses the connectivity of a graph to assess the probability that each of its nodes has to be reached and visited starting from a random position. As the probability mass (distribution) over the graph nodes is uniform, then, iteratively, the number of ingoing and outgoing connections serves as a means to increase or decrease the relative weight of each node. In order to apply this approach to WSD, following Agirre et al. (2014), SyntagRank uses a variant of the PageRank algorithm, the Personalized PageRank (PPR), in which the initial probability mass is distributed over a restricted set of specific nodes (i.e. the nodes representing the content words to be disambiguated in a given context[3]). Hence, given an initial set of nodes, the outcome of the PPR algorithm is a vector encoding all the information concerning the probability distributions of all the nodes in the graph.

3 Architecture of SyntagRank

SyntagRank is a knowledge-based disambiguation system which uses the PPR algorithm to determine

[2] http://wordnetcode.princeton.edu/glosstag.shtml

[3] In SyntagRank, a context is equivalent to a single whole sentence. Therefore, given an input paragraph made up of, say, three sentences, the system will perform the disambiguation task separately for each of these three sentences.

the most appropriate sense of a given word in context. This approach, already discussed by Agirre and Soroa (2009), is here presented in an optimized, rebuilt version, employing the LKBs described in Section 2.1 to achieve state-of-the-art knowledge-based performance across five languages: English, German, French, Spanish, and Italian. Our architecture (Figure 1) is composed of three main modules: (i) multilingual NLP pipeline, (ii) candidate retrieval, and (iii) disambiguator.

3.1 Multilingual NLP Pipeline

In order to allow the user to provide an unprocessed text as input for SyntagRank to disambiguate, our system employs a multilingual NLP pipeline which preliminarily performs the functions of tokenization, sentence splitting, lemmatization and Part of Speech (PoS) tagging. Depending on the input language, SyntagRank utilizes either the Stanford CoreNLP suite[4] (Manning et al., 2014), or the models provided by The Italian NLP Tool (Palmero Aprosio and Moretti, 2016, TINT).

3.2 Candidate Retrieval

English Candidate Retrieval With each token in the input text already pre-processed, and considering that each node in our graph corresponds to a unique WordNet synset (see Section 2), in this phase we can retrieve, for each content word (target word) in a single sentence, all those candidate concepts (synsets) for which a coincident lexicalization exists. In doing so, in line with the word-to-word heuristics described in (Agirre et al., 2014), we exclude the target word when retrieving the candidate concepts so as to avoid the probability mass being distributed across the most frequent sense of the target word. The resulting set of collected concepts C, which will now include all the possible senses for the non-target words in the input sentence, thus establishes the starting nodes for the PPR algorithm.

In view of the fact that, according to the Linearity Theorem (Jeh and Widom, 2003), the PPR vector computed starting from a set of nodes C is equivalent to the weighted average of the PPR vectors calculated using each of the nodes in C as single starting points, all the PPR vectors in SyntagRank have been preliminarily determined for each node in the graph, with the purpose of minimizing execution times[5]. Thus, the PPR vector for a precise context (i.e. an input sentence) is calculated simply by determining the weighted average of the pre-computed PPR vectors for each of its nodes[6]. The weight factor $p(w, s)$, for each candidate s associated with a content word w, is computed as follows:

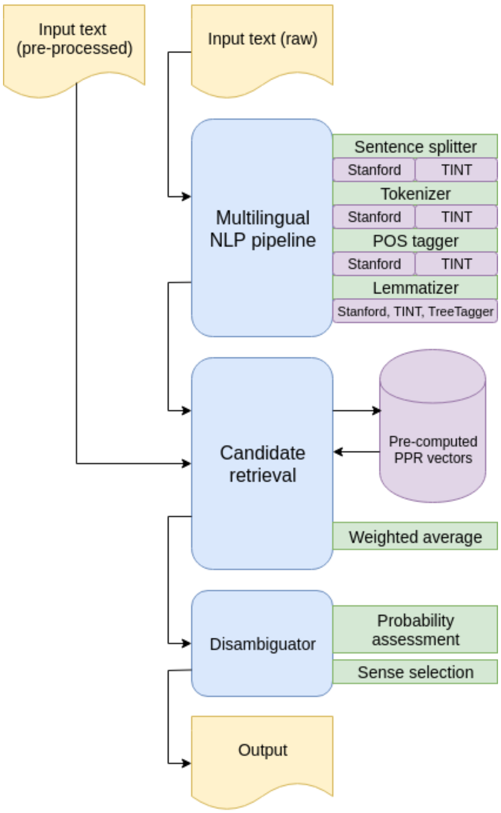


Figure 1: Architecture diagram of SyntagRank.

$$p(w, s) = \frac{1}{N * |senses_w|} freq_{ws} \tag{1}$$

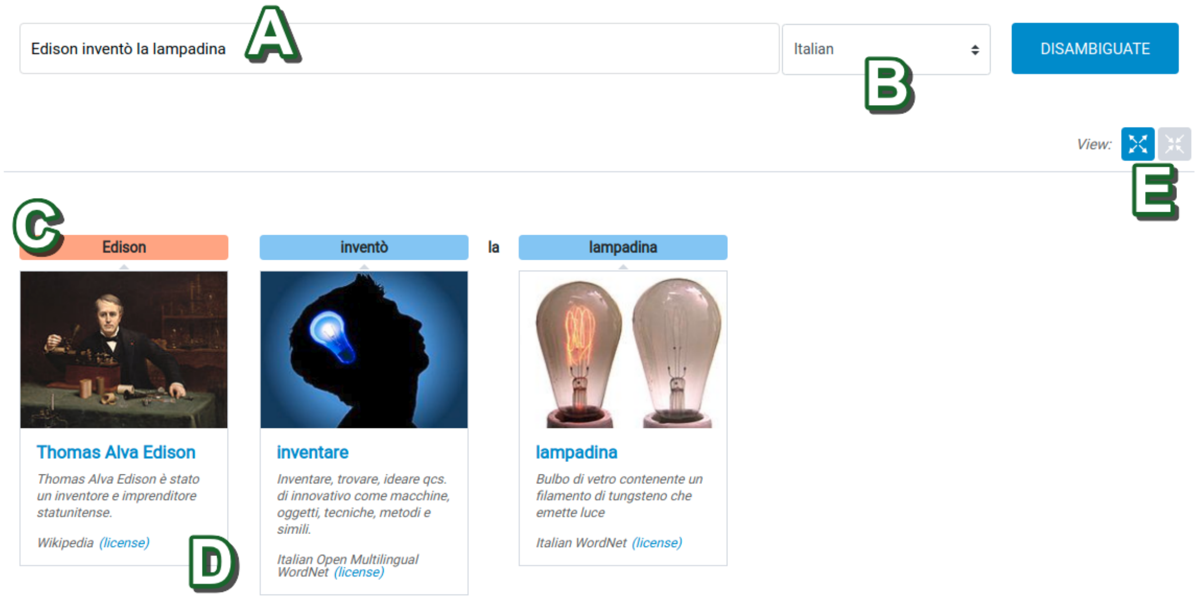
where N is the number of content words in the input sentence and $senses_w$ is the set of sense candidates associated with w. Moreover, since the graph connectivity gets denser around most frequent senses (MFS) – according to their distribution in SemCor[7] (Miller et al., 1993) –, and in view

[4] Except for the English language, for which the Stanford CoreNLP pipeline has full coverage, in order to perform the lemmatization for German, French, and Spanish, we use instead TreeTagger (Schmid, 1995).

[5] All the pre-computed PPR vectors are stored in binary format, and are accessed via a memory-mapped file supported by a Least Recently Used (LRU) cache.

[6] With regard to our PPR implementation details, we opted for a damping factor of 0.85. In addition, the algorithm performs a variable number of iterations (random walks) over the graph until reaching convergence, i.e. when the difference between the scores of any node computed at two successive iterations falls below a threshold of 10^{-4}.

[7] SemCor is the largest, manually sense-annotated corpus of English, and is currently the *de facto* standard reference dataset for several WSD applications.



Edison inventò la lampadina

Italian

Thomas Alva Edison

Thomas Alva Edison è stato un inventore e imprenditore statunitense.

Wikipedia (license)

inventare

Inventare, trovare, ideare qcs. di innovativo come macchine, oggetti, tecniche, metodi e simili.

Italian Open Multilingual WordNet (license)

lampadina

Bulbo di vetro contenente un filamento di tungsteno che emette luce

Italian WordNet (license)

Figure 2: User interface of SyntagRank when the Italian language is selected and the sentence *'Edison inventò la lampadina'* (Edison invented the light bulb) is typed as input query. Disambiguation results are displayed in extended view by default. Overlaying letters over the image are detailed in Section 4.

of the fact that unsupervised systems tend to have a strong bias towards the MFS (Calvo and Gelbukh, 2015; Postma et al., 2016; Pasini et al., 2020), we accounted for potential skew towards MFS by including the parameter $freq_{ws}$, i.e. the normalized value resulting from the number of occurrences for a given word sense in SemCor, divided by the total number of occurrences for all the senses of the same word.

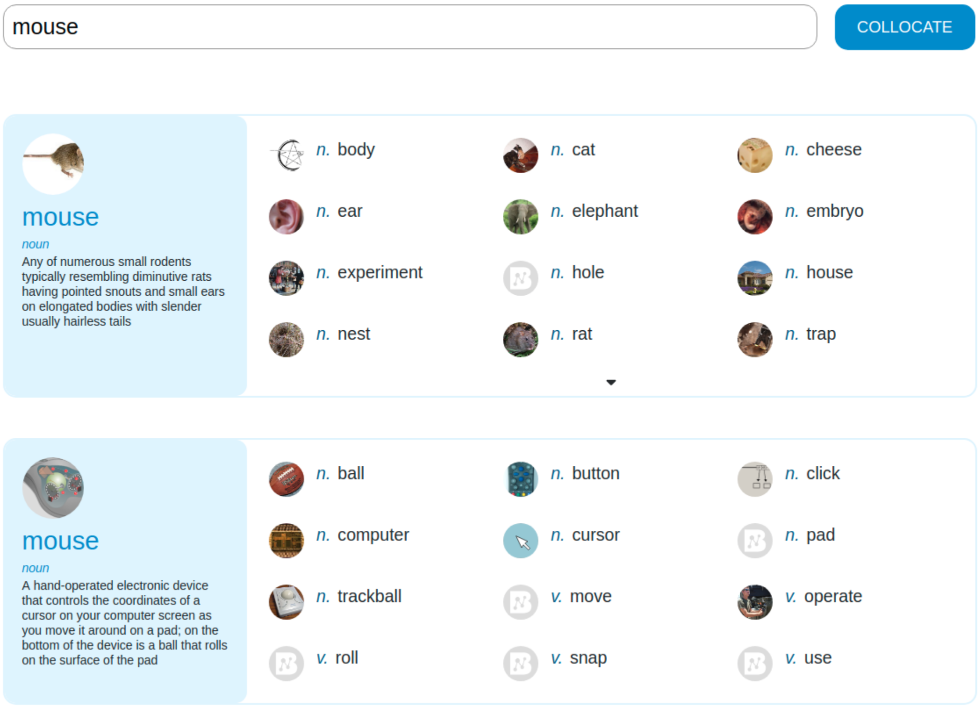
Multilingual Candidate Retrieval Concepts represented in a semantic network are language independent by definition. Still, in order to retrieve sense candidates for words in specific languages, we need the nodes in the graph to be mapped with lexicalizations in those languages. As mentioned in Section 2.1, WordNet provides this information for the English language only, therefore, in order to retrieve the lexicalizations in languages other than English we exploited the BabelNet semantic network, which inherently aligns lexicalizations in 284 distinct languages to the original WordNet 3.0 synsets. Nevertheless, two main flaws lie in this approach: (i) the lexicalizations in BabelNet are induced from automatically-linked resources, hence, their quality might be sub-optimal, and (ii) no SemCor equivalent exists for other languages, which means we do not have any accessible MFS information to exploit when computing the weighted average between vectors. In order to address both these flaws concurrently, we devised a strategy to mimic the MFS ranking function by associating a confidence score with each of the lexical resources from which BabelNet derives its lexicalizations (e.g. Wikidata, OmegaWiki or Wikipedia, among others). To this end, after conducting an empirical study to assess the quality of random translation samples provided by each individual resource mapped to BabelNet, we assigned a normalized confidence score to them. Consequently, for each unique lexicalization, we have been able to compute its "MFS" score as the average confidence among all the resources providing that lexicalization for a specific concept.

3.3 Disambiguator

After retrieving the PPR vectors for each candidate sense and computing their weighted average (as described in Section 3.2), the last module of SyntagRank serves as a means to finally: (i) extract the probability values for the senses of the target word from the averaged PPR vector, and (ii) select the sense with the highest probability value as the result of the disambiguation for the target word.

4 Web Interface

Figure 2 shows the Web interface of SyntagRank. Its components are explained below.



mouse

mouse

noun

Any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails

mouse

noun

A hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad

Figure 3: User interface of the SyntagNet Explorer when the English word *mouse* is typed as input query.

A. Query The system takes as input the text to be processed[8]. Users can enter either single words, multiword expressions (MWEs), or full sentences as input queries. In the event that the input text is a sentence, this will be processed by the disambiguator and the system will return a disambiguated sentence (see Paragraph C). Otherwise, if the query matches an entry in the SyntagNet database, the interface will switch to the SyntagNet Explorer (see Section 4.1) to display all the lexical-semantic combinations available for all the senses of the word/MWE provided as input query.

B. Language Selection The drop down menu allows the user to select the language in which the input text is provided. Currently, SyntagRank offers disambiguation in five different languages: English, German, French, Spanish and Italian.

C. Disambiguated Sentence If an input text has been provided, the interface will display the results of the disambiguation here, with tokens highlighted in different colors for *Concepts* (blue) and *Named Entities* (orange).

D. Disambiguated Token Each disambiguated token is accompanied by a tooltip which shows the image, word sense and definition, as retrieved from the corresponding entry in BabelNet 4.0.

E. View Selection The Web interface allows the user to display the disambiguated sentence in extended or compact form. In the extended view, the focus is placed on the tokens: the disambiguated sentence is shown as a horizontal slider, navigable by means of arrows located on the left and right ends of the container, and the user is thereby given a means to quickly leaf through all the disambiguation results at the same time. Instead, when selecting the compact view, the focus is shifted to the sentence. In this mode, the information associated with the disambiguated tokens will be shown only if the user hovers the mouse cursor over a highlighted token.

4.1 SyntagNet Explorer

In addition to the SyntagRank disambiguation system, our Web interface also provides users with full access to the SyntagNet database. By typing into the query bar a word or MWE which is present in SyntagNet[9] (an autocomplete function will provide the user with search suggestions), the interface will switch to the SyntagNet Explorer (Figure 3). The SyntagNet Explorer displays a list of boxes, each containing a sense of the input word/MWE. Senses in the list are ordered according to (i) PoS tag and

[8] The Web interface only allows raw text as input.

[9] At the time of writing, the SyntagNet Explorer is available for the English language only.

	English						Multilingual						
							SemEval-13				SemEval-15		
System	Sens2	Sens3	Sem07	Sem13	Sem15	All	IT	ES	DE	FR	IT	ES	All
Babelfy	67.0	63.5	51.6	66.4	70.3	65.5	66.6	69.5	69.4	56.9	-	-	-
UKB	68.8	66.1	53.0	68.8	70.3	67.3	-	-	-	-	-	-	-
SyntagRank	**71.6**	**72.0**	**59.3**	**72.2**	**75.8**	**71.7**	**72.1**	**74.1**	**76.4**	**70.3**	69.0	63.4	71.2

Table 1: F1 scores (%) for English all-words fine-grained WSD (left) and for multilingual all-words fine-grained WSD (right). Statistically-significant differences against our results are underlined according to a χ^2 test, $p < 0.01$. Results under "All" refer to the concatenation of the English (left) and multilingual (right) datasets.

(ii) sense frequency (in line with BabelNet 4.0). On the left side (blue background), the boxes show information for word senses, along with PoS tags, sense definitions and illustrations. By clicking on a sense name, the corresponding BabelNet entry will open in a separate tab. On the right side (white background), all the lexical-semantic items (collocates) linked with the corresponding word senses via SyntagNet are listed. Further information about collocates is provided by hovering the mouse over each item. Finally, clicking on a collocate will start a new query with the selected word.

4.2 Usage of the RESTful API

The RESTful API we provide can be used effectively to query the SyntagRank system programmatically. Unlike the Web interface, our API allows the user to input a pre-processed text in addition to performing standard queries with raw text. For the full documentation of the RESTful API, along with the required parameters description, please refer to Appendix A: API Documentation.

5 Evaluation

In order to assess its performance, we tested SyntagRank on the five English all-words WSD evaluation datasets standardized according to WordNet 3.0 in the framework of Raganato et al. (2017a), namely: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015). As regards the appraisal of SyntagRank in a multilingual setting, we used the German, Spanish, French and Italian annotations available in the amended version of the SemEval-2013 and SemEval-2015 evaluation datasets[10], which is accordant with the BabelNet API 4.0.1 graph and enables testing on a larger number of instances than hitherto.

In Table 1, we report F1 scores for SyntagRank in the English (left), and multilingual (right) settings, along with comparisons to the best configurations of two distinct graph-based disambiguation systems: Babelfy (Moro et al., 2014) and UKB (Agirre et al., 2014). As can be seen, SyntagRank outperforms its direct competitors by a considerable margin[11], on both the English and multilingual settings. These results substantiate the idea that applying the PPR algorithm to a graph injected with high-quality syntagmatic knowledge is crucial to enhancing disambiguation performances.

6 Conclusion

In this paper we presented and described the architecture of SyntagRank, our state-of-the-art knowledge-based system for multilingual Word Sense Disambiguation using syntagmatic information. We also provided details concerning the use of SyntagRank's Web interface and RESTful API, accessible at http://syntagnet.org/ and http://api.syntagnet.org, respectively.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under the grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

[10] Made available at https://github.com/SapienzaNLP/mwsd-datasets.

[11] For the purpose of these experiments, we set a threshold $T = 0.4$ for the PPR values of any given sense; for values failing to reach the threshold, the MFS was chosen instead as the result of the disambiguation.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proc. of EACL*, pages 33–41, Athens, Greece.

Michele Bevilacqua and Roberto Navigli. 2019. Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation. In *Proc. of RANLP*, pages 122–131, Varna, Bulgaria.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA, USA.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding Dense, Weighted Connections to WordNet. In *Proceedings of the third international WordNet conference*, pages 29–36, South Jeju Island, Korea.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Hiram Calvo and Alexander Gelbukh. 2015. Is the Most Frequent Sense of a Word Better Connected in a Semantic Network? In *Proc. of ICIC*, pages 491–499, Fuzhou, China.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA, USA.

Taher H. Haveliwala. 2002. Topic-Sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526, Honolulu, HI, USA.

Glen Jeh and Jennifer Widom. 2003. Scaling Personalized Web Search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, Budapest, Hungary.

Lothar Lemnitzer, Holger Wunsch, and Piklu Gupta. 2008. Enriching GermaNet with verb-noun relations - a case study of lexical acquisition. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008, May 28-30, 2018*, pages 156–160, Marrakech, Morocco.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MD, USA.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3532–3538, Hong Kong, China.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Proc. of HLT*, pages 303–308, Plainsboro, NJ, USA.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, CO, USA.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (Present and Future) Use. In *Proc. of IJCAI*, pages 5697–5702, Stockholm, Sweden.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, GA, USA.

Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Roberto Navigli and Simone P. Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence Journal*, 193:217–250.

Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*.

Tommaso Pasini. 2020. The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-20*, Yokohama, Japan.

Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. 2020. CluBERT: a Cluster-Based Approach for Learning Sense Distributions in Multiple Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA, USA.

Simone P. Ponzetto and Roberto Navigli. 2010. Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden.

Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016. Addressing the MFS Bias in WSD systems. In *Proc. of LREC*, pages 1695–1700, Portorož, Slovenia.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proc. of SemEval-2007*, pages 87–92, Stroudsburg, PA, USA.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017a. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL*, pages 99–110, Valencia, Spain.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017b. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proc. of EMNLP*, pages 1167–1178, Copenhagen, Denmark.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, USA.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proc. of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

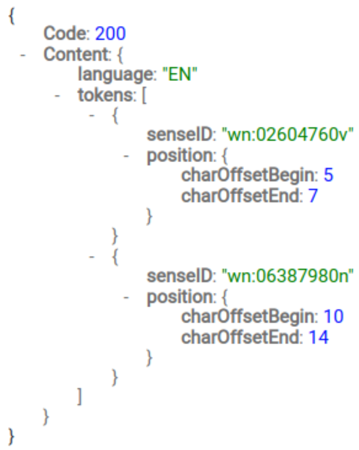
Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43, Barcelona, Spain.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of Global Wordnet Conference*, Wroclaw, Poland.

A API Documentation

In what follows we describe the typical usage of our RESTful API and its parameters. The SyntagRank API allows the user to perform two distinct requests: (i) `Disambiguate Text` and (ii) `Disambiguate Tokens`.

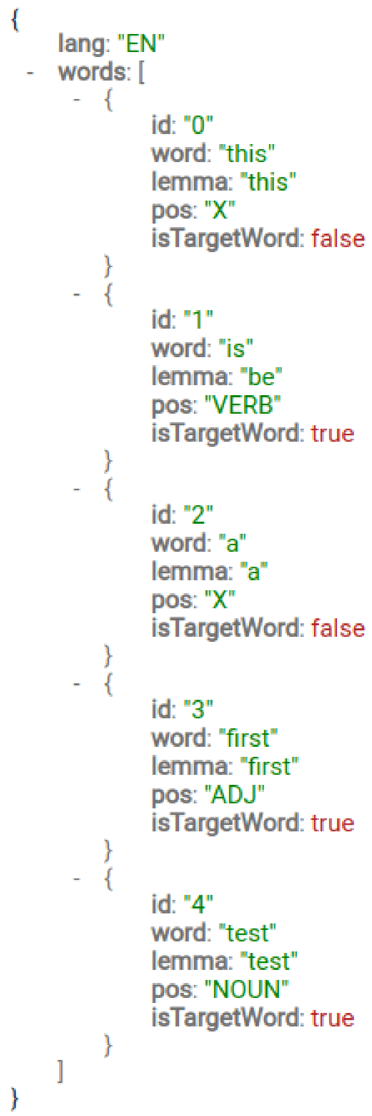
Disambiguate Text With `Disambiguate Text`, SyntagRank will process a raw text provided as input, given a target language among the five currently supported: EN (English), DE (German), FR (French), ES (Spanish), and IT (Italian).
Method type, URL, parameters and response description are specified in detail in Table 2. Figure 4 shows an example of a *success response* for the `Disambiguate Text` query.



```
{
    Code: 200
  - Content: {
        language: "EN"
      - tokens: [
          - {
                senseID: "wn:02604760v"
              - position: {
                    charOffsetBegin: 5
                    charOffsetEnd: 7
                }
            }
          - {
                senseID: "wn:06387980n"
              - position: {
                    charOffsetBegin: 10
                    charOffsetEnd: 14
                }
            }
        ]
    }
}
```

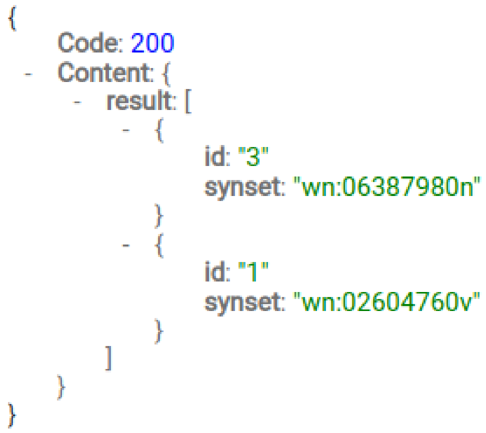
Figure 4: Example of a *success response* for `Disambiguate Text` when the language chosen is English and the input text is "this is a text".

Disambiguate Tokens With `Disambiguate Tokens`, SyntagRank will accept a pre-processed text as input to be disambiguated.
As for `Disambiguate Text`, language specification is required. Each token must show information concerning index (`id`), word form (`word`), lemma form (`lemma`), POS tag (`pos`), and a boolean indicating whether the token is a content word to be disambiguated (`isTargetWord`). In Table 3, we provide exhaustive details concerning method type, URL parameters, token parameters and response description for `Disambiguate Tokens`. Additionally, Figures 5 and 6 show, respectively, an example of a typical request, and its *success response*.



```
{
    lang: "EN"
  - words: [
      - {
            id: "0"
            word: "this"
            lemma: "this"
            pos: "X"
            isTargetWord: false
        }
      - {
            id: "1"
            word: "is"
            lemma: "be"
            pos: "VERB"
            isTargetWord: true
        }
      - {
            id: "2"
            word: "a"
            lemma: "a"
            pos: "X"
            isTargetWord: false
        }
      - {
            id: "3"
            word: "first"
            lemma: "first"
            pos: "ADJ"
            isTargetWord: true
        }
      - {
            id: "4"
            word: "test"
            lemma: "test"
            pos: "NOUN"
            isTargetWord: true
        }
    ]
}
```

Figure 5: A request example in English for `Disambiguate Tokens`.



```
{
    Code: 200
  - Content: {
      - result: [
          - {
                id: "3"
                synset: "wn:06387980n"
            }
          - {
                id: "1"
                synset: "wn:02604760v"
            }
        ]
    }
}
```

Figure 6: *Success response* with `Disambiguate Tokens` for the input shown in Figure 5.

Disambiguate Text	
Method	GET/POST
URL	http://api.syntagnet.org/disambiguate?lang=language&text=text
URL Parameters	
text (String)	The text to be disambiguated (max length: 1,500 characters). E.g.: text=this is a text.
lang (String)	The language of the input text, among the currently supported: EN, DE, FR, ES and IT.
Response description	
language	The language of the disambiguated tokens.
tokens	Contains a list of disambiguated tokens.
senseID	Identifies the WordNet 3.0 offset for the concept assigned to the token.
position	Contains information concerning the token positioning.
charOffsetBegin	Highlights the position where a given term instance starts. Expressed as char offset.
charOffsetEnd	Highlights the position where a given term instance ends. Expressed as char offset.

Table 2: Details for the Disambiguate Text request.

Disambiguate Tokens	
Method	POST
URL	http://api.syntagnet.org/disambiguate_tokens
URL Parameters	
lang (String)	The language of the input text, among the currently supported: EN, DE, FR, ES and IT.
words (List<Token>)	Contains a list of words, each representing a single token of the input text.
Token Parameters	
id (String)	Identifies the position of the token in the input text.
word (String)	Identifies the token, as it appears in the input text.
lemma (String)	The lemmatized form of the token.
pos (String)	The Part of Speech (PoS) of the token.
isTargetWord (boolean)	If true, identifies a token (for a content word) to be disambiguated.
Response description	
result	Contains a list of disambiguated tokens.
id	Identifies the position of the disambiguated token according to the input text.
synset	Identifies the WordNet 3.0 offset for the concept assigned to the token.

Table 3: Details for the Disambiguate Tokens request.