Word Sense Disambiguation for Linking Domain-Specific Resources

Lucía Palacios

Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM), Madrid, Spain

Abstract

Word Sense Disambiguation (WSD) is a traditional task in Natural Language Processing (NLP), which involves determining the meaning of words based on their context and differentiating between multiple possible senses. Current approaches, reliant on general-purpose lexical resources like WordNet and Wikidata, often present limitations for domain-specific tasks. This research focuses on addressing the challenge of lexical ambiguity in specialized terminologies in Spanish, aiming to enhance the interconnection and interoperability of those resources. The proposed approach is based on accurately tagging terms with specific senses using automatically integrated sense inventories and hybrid WSD algorithms enabling the use of Entity Linking (EL) and Entity Matching (EM) techniques to facilitate the transformation of disambiguated terminologies into Linked Data formats. Ultimately, this work aims to contribute to the advancement of domain-specific applications, improving semantic analysis and knowledge extraction in specialized fields.

Keywords

Word Sense Disambiguation, Terminologies, Sense Inventories, Entity Linking, Entity Matching

1. Introduction and Motivation

High-quality language resources are essential for achieving optimal results in the development of any Natural Language Processing (NLP) system. General language resources are applicable across multiple contexts, while specialized language resources focus on terminologies pertinent to a specific domain.

Terminologies are collections of words that encompass the relevant vocabulary and concepts within a specific field or subfield. The meaning of each term depends on the domain of the terminology to which it belongs, i.e. it cannot be understood in isolation. In this context, we define a domain as a subdivision of general world knowledge, that can be further divided into more specific subdomains [1].

This research focuses on terminological resources published in Semantic Web formats, following the Linked Data principles and contributing to the population of the Linguistic Linked Open Data (LLOD)[2]. This paradigm not only establishes connections amongst terms in the resource but fosters their interoperability, reusability, and machine-readability. For terminologies to be effectively linked, it is crucial to accurately tag each term with the specific sense it denotes. This can be accomplished using Word Sense Disambiguation (WSD) techniques.

WSD is the process of mapping a word within a specific context to its most appropriate sense from a predefined lexical database, known as a sense inventory [3]. This task is a fundamental challenge in NLP due to its critical role in numerous applications, including Sentiment Analysis [4] and Machine Translation [5] among others.

Thanks to recent advances in Deep Learning (DL), several proposals have achieved remarkable success in determining the senses of words in open-domain contexts [6, 7], using general-purpose lexical resources such as WordNet[1] or Wikidata.[2] These computational lexicons are used as sense inventories, which are databases that collect all possible senses of words comprising one or more languages. They are designed for general language use but lack the specific data needed to develop disambiguation systems tailored to terminologies and nuanced domain-specific meanings [3].

Current research directions for domain-specific WSD systems are based on domain label inventories [8] and algorithms that use existing large lexical bases to apply them to sets of specialized literature [9]. A domain inventory can be defined as a collection of domain labels. The issue with these inventories is that they encompass only broad generic domains since they are used to label the contents of the general lexical databases. For example, in WordNet Domains,[3] which is a widely used resource for sense inventories, the domain "law" receives only the subdomain "state", excluding many others such as "labor law" or "commercial law".

Therefore, the investigation aims to develop a system capable of resolving the lexical ambiguity of domain-specific terminologies through the automatic generation of domain-specific sense inventories with adjustable granularity levels by integrating specialized sources. This system aims to assign the most suitable sense from these inventories to each target term. Moreover, this research will focus on developing a module for the system using methods like Entity Linking (EL) or Entity Matching (EM) to interconnect the disambiguated terminological resources in Linked Data formats.

2. State of the Art

In this section, related work for the WSD task addressing domain-specific lexical ambiguity is first overviewed, followed by a review of relevant work on EL and EM systems developed for connecting semantically associated resources. Both WSD and EL/EM tasks focus on resolving lexical ambiguity in language. However, they differ in their approach and purpose: EL connects textual mentions with their corresponding entities appearing in a knowledge base, which may or may not contain the exact mention (e.g., linking "Barack" with "Barack Obama") [10]. EM consists of aligning entries from different structured datasets that refer to the same entity, despite having different representations [11]. Finally, WSD assigns the correct meaning to a word within the context, which requires an exact match with predefined meanings (e.g., determining whether "bank" refers to a financial institution or a river bank) [3].

---

[1]https://wordnet.princeton.edu/
[2]https://www.wikidata.org/
[3]https://wndomains.fbk.eu/

2.1. Word Sense Disambiguation

WSD tasks are crucial in advancing NLP techniques due to the pervasive nature of lexical ambiguity in language. As a long-standing challenge, WSD has been approached from various perspectives. Three primary methodologies have been developed to create WSD systems. First, Knowledge-based WSD algorithms navigate through the structure of computational lexicons to leverage the encoded semantic information, making them independent of training corpora. Notable examples include SREF [12] and the so-called Semantic Specialization for WSD [13]. Another approach is the Neural Network-based WSD. These algorithms treat WSD as a classification task. The state-of-the-art neural systems use pre-trained language models that are fine-tuned with corpora annotated in a term-context-sense format. Typically, the training corpora are manually or automatically [14] annotated with senses defined in a computational lexicon. Lastly, the Hybrid WSD approach is based on the integration of language models with knowledge graphs and it is considered the most effective approach [3].

The above-mentioned approaches make use of large sense inventories. The most widely used are Wordnet, BabelNet[4], and Wikidata. As mentioned in the previous section, they present significant limitations for domain-specific WSD, especially granularity limitations. That is narrow coverage in specific contexts, inconsistent updating of content, redundancy, limited ability to customize resources for particular tasks or needs, and noise due to the integration of too diverse and too large databases [3].

Early efforts to tackle the problem of large sense inventories focused on grouping all senses associated with words sharing the same lemma [15]. In contrast, WordNet employs higher-level categories called SuperSenses, which organize synsets based on their grammatical and broad semantic types. However, many synsets are tagged as "ALL", indicating a domain-general classification. Some related work is the WordNet Domains[5] project, which is a semi-supervised annotated domain inventory that labels WordNet synsets with 165 hierarchically organized domains. However, a significant number of synsets are labeled as "FACTOTUM", expressing that they do not belong to a specific domain.

Building on this approach, BabelDomains[6] is a proposal for labeling BabelNet synsets using 42 domain types extracted from Wikipedia's predefined categories. The main limitation of BabelDomains is its exclusive focus on nouns and that it is not open source. The best results have been achieved with the Coarse Sense Inventory (CSI) [8], as shown by its outcome of almost 86 points F1 with a supervised WSD system. This inventory aims to reduce the granularity of WordNet. It is manually annotated with 45 domain labels and demonstrates very high inter-annotator agreement. However, the domain information to be tagged from WordNet is too general to deal with very specialized terminologies and it is exclusively available in English.

The methods presented address granularity issues to varying degrees, but none facilitate domain-specific disambiguation tasks. Therefore, this work aims to explore automatic resource generation techniques for domain-specific and Spanish WSD tasks. Domain information can be extracted from various sources that have not traditionally been used for WSD. For example,

---

[4]https://babelnet.org/
[5]https://wndomains.fbk.eu/
[6]http://lcl.uniroma1.it/babeldomains/

Bevilacqua et al. [3] proposed using Wiktionary. Other promising options include using IATE[7] domain labels as suggested by Sainz et al. [16], or dictionaries curated and frequently updated by language academies, such as the *Diccionario Panhispánico de Español Jurídico*[8]. By integrating sources relevant to the specific domain instead of extracting domain knowledge from a general database, a framework for domain-specific hybrid WSD can be developed with high precision and without granularity problems.

2.2. Entity Linking and Matching

After employing WSD algorithms to tag each term with its specific sense, the next objective is to establish connections between terms that are semantically related. The ultimate goal of this research work is to develop a system that enables terminological resources to be integrated into an intelligent and interconnected system (Semantic Web), where data is not only human-readable but also machine-interpretable. To connect the specialized terminologies, this work will leverage EL (if the resources are in an unstructured format) or EM (if the resources are in a structured format) techniques.

EL involves connecting entities from textual sources to their corresponding entries in a knowledge base or database. Early systems were based on rules [17] and machine-learning [18] approaches. State-of-the-art systems are based on embedding generation [19], feature extraction [20], or pre-trained language models [21]. As for EL systems for highly specialized terminologies, Zhang et al. [22] presents a notable example in the biomedical domain, among many other proposals for various domains.

EM allows finding which entries across two knowledge bases refer to the same entity. Traditional methods are based on similarity calculations concerning entity features, and later these techniques were combined with ontological rules [23]. With the advent of DL, the focus shifted towards Representation Learning (RL) methods that allow models to learn a low-dimensional vector representation of entities (i.e. Knowledge Graph Embeddings), like the TransE model [24]. Those that occupy the state of the art incorporate additional information such as knowledge graph structure information or external resources [25].

3. Open Research Problem, Hypothesis and Research Questions

The Open Research Problem (ORP) addressed by this paper can be formulated as follows: *There is a lack of adequate sense inventories to disambiguate the entities of terminological resources. As a result, it is very complex and costly to generate interconnected linguistic resources belonging to specific domains.*

Therefore, the subsequent Research Hypothesis (RH) is proposed: *The automatic generation of domain-specific sense inventories by integrating various sources would allow the development of a system that encompasses WSD and EL/EM techniques to transform specialized terms into the Linked Data format.* Based on the identified ORP and the RH, the following Research Questions (RQS) are proposed:

---

[7]https://iate.europa.eu/
[8]https://dpej.rae.es/

– **RQ1**: What techniques and methodologies are applicable for automating the creation of domain-specific sense inventories?

– **RQ2**: How can these inventories be evaluated considering granularity levels and senses not included in the inventory?

– **RQ3**: What strategies are effective in developing a system that employs sense inventories to disambiguate and interlink terminological resources in semantic web formats?

– **RQ4**: What methods can be employed to establish a robust evaluation framework that ensures specialized terminologies are accurately linked to semantically related resources?

4. Objective and Sub-Objectives

Based on the RQs presented before, this work has two objectives: (RO1) develop a system to resolve the lexical ambiguity of domain-specific terminologies and (RO2) create a module to interconnect disambiguated and semantically related terms to ensure their Linked Data format employing EL or EM techniques.

To achieve this objective, the following sub-objectives (SOs) are proposed:

- **SO1**. Identification and/or construction of corpora and sense inventories. Automatic and rigorous generation of domain-specific resources for each domain, addressing challenging aspects such as levels of granularity and uncommon or newly emerged senses.

- **SO2**. Development and evaluation of lexical disambiguation algorithms. Framework design and implementation, selecting the best-suited method for the task, such as embedding approaches, language models, etc.

- **SO3**. Implementation of EL and EM techniques for the integration and reuse of domain-specific terminologies in semantic applications.

- **SO4**. Evaluation and Validation. Design and execute comprehensive tests to assess the system's accuracy, efficiency, and scalability. Finally, validate the results with domain experts and make necessary adjustments to improve the system quality.

5. Research Methodology

To address the research objectives (ROs) presented, a methodology combining theoretical research with NLP systems development is proposed. This methodology is structured into three distinct phases (Ps), as illustrated in Figure 1. Each Phase encompasses two Sub-Phases, detailed as follows:

- **P1. Review and collection of existing work**. Comprehensive literature review on lexical disambiguation in domain-specific terminologies, EL and EM techniques, and evaluation methods. This review serves as a theoretical basis to justify the design and implementation of the proposed systems. After the review, this work will focus on the following sub-stages:

- P1.1: Identification and collection of existing resources. Selecting those that produce the best results for specialized terminologies.

- P1.2: Evaluation of existing frameworks. State-of-the-art WSD and EL/EM systems will be evaluated to identify those that produce the best results for specialized terminologies.

- **P2. Generation of new resources and frameworks**. Election of the most effective algorithm thanks to a series of experiments that will be conducted to determine the optimal methods for addressing the specific tasks chosen in this research. After selecting the most appropriate frameworks:

  - Automatic generation of domain-specific resources. Identification and processing of specialized and relevant sources for their integration into domain-specific sense inventories and corpora.

  - Development of the system framework. Design, training, and tuning the system that will address the tasks of this work.

- **P3. Evaluation and Validation**. Various tests will be conducted to verify the effectiveness of the entire system.

  - P3.1 Automatic evaluation. Initial tests in terms of accuracy, efficiency, and scalability using metrics such as F1 score, recall, and others.

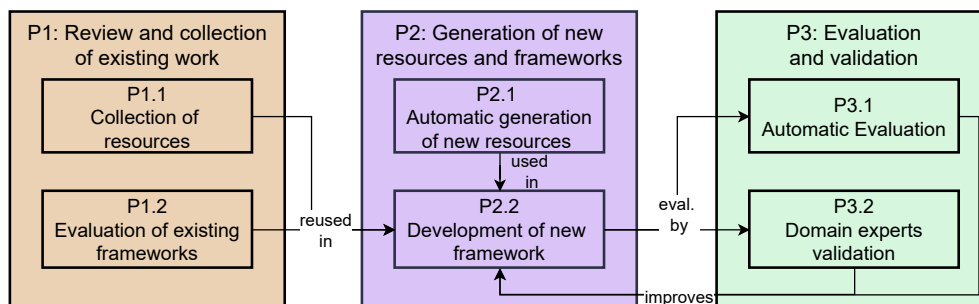 - P3.2 Domain experts validation. Second evaluation to ratify the results obtained with domain experts.

Figure 1: Modular approach for the research plan stages.

6. Conclusions and Specific Research Elements Proposed for Further Discussion

The goal of this work is to investigate how to develop a system that can solve WSD and EL/EM tasks for specialized terminologies to create interconnected, machine-readable data. The

research is currently in its initial stage, which involves an in-depth study of the related work. The purpose of this phase is to identify the challenges, methodologies, and technological gaps in current WSD and EL/EM systems for domain-specific terminologies. Through this ongoing literature review, preliminary future research directions have been established to evaluate the current systems and resources that are potentially suitable for the tasks at hand. Specifically, the key elements proposed for further discussion and research include:

- **Automatic enrichment and improvement of sense inventories**: Develop methods that allow continuously updating and expanding the domain-specific sense inventories, ensuring they remain manageable and comprehensive.

- **Cross-Domain System Development**: Design a versatile system that can adapt to multiple domains beyond the initially targeted ones. Methodologies that allow the system to handle diverse domain-specific terminologies effectively will be researched.

- **Integration of NLP models**: Investigation on how various NLP models can be combined with the enriched sense inventories and corpora.

It is expected that the successful development of this system would facilitate the integration and reuse of domain-specific knowledge in various applications, contributing to advancements in fields such as healthcare, legal, and scientific research.

Acknowledgments

References

[1] B. Hjørland, H. Albrechtsen, Toward a new horizon in information science, Journal of the Association for Information Science and Technology (1995). URL: https://api.semanticscholar.org/CorpusID:215898905.

[2] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Open Data Cloud, Springer International Publishing, Cham, 2020, pp. 29–41. URL: https://doi.org/10.1007/978-3-030-30225-2_3. doi:10.1007/978-3-030-30225-2_3.

[3] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: International Joint Conference on Artificial Intelligence, 2021. URL: https://api.semanticscholar.org/CorpusID:237100274.

[4] X. Zhang, R. Mao, K. He, E. Cambria, Neuro-symbolic sentiment analysis with dynamic word sense disambiguation, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. URL: https://openreview.net/forum?id=SQodZvCM5g.

[5] V. Iyer, E. Barba, A. Birch, J. Pan, R. Navigli, Code-switching with word senses for pre-training in neural machine translation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 12889–12901. URL: https://aclanthology.org/2023.findings-emnlp.859. doi:10.18653/v1/2023.findings-emnlp.859.

[6] E. Barba, T. Pasini, R. Navigli, ESC: Redesigning WSD with extractive sense comprehension, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 4661–4672. URL: https://aclanthology.org/2021.naacl-main.371. doi:10.18653/v1/2021.naacl-main.371.

[7] J. Zhang, R. He, F. Guo, C. Liu, Quantum interference model for semantic biases of glosses in word sense disambiguation, Proceedings of the AAAI Conference on Artificial Intelligence 38 (2024) 19551–19559. URL: https://ojs.aaai.org/index.php/AAAI/article/view/29927. doi:10.1609/aaai.v38i17.29927.

[8] C. Lacerra, M. Bevilacqua, T. Pasini, R. Navigli, Csi: A coarse sense inventory for 85% word sense disambiguation, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 8123–8130. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6324. doi:10.1609/aaai.v34i05.6324.

[9] D. Chopard, P. Corcoran, I. Spasic, Word sense disambiguation of acronyms in clinical narratives, Frontiers in Digital Health 6 (2024). URL: https://api.semanticscholar.org/CorpusID:268178900.

[10] A. Moro, A. Raganato, R. Navigli, Entity Linking meets Word Sense Disambiguation: a Unified Approach, Transactions of the Association for Computational Linguistics 2 (2014) 231–244. URL: https://doi.org/10.1162/tacl_a_00179. doi:10.1162/tacl_a_00179. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00179/1566870/tacl_a_00179.pdf.

[11] Y. Li, J. Li, Y. Suhara, J. Wang, W. Hirota, W.-C. Tan, Deep entity matching: Challenges and opportunities, J. Data and Information Quality 13 (2021). URL: https://doi.org/10.1145/3431816. doi:10.1145/3431816.

[12] M. Wang, Y. Wang, A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6229–6240. URL: https://aclanthology.org/2020.emnlp-main.504. doi:10.18653/v1/2020.emnlp-main.504.

[13] S. Mizuki, N. Okazaki, Semantic specialization for knowledge-based word sense disambiguation, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3457–3470. URL: https://aclanthology.org/2023.eacl-main.251. doi:10.18653/v1/2023.eacl-main.251.

[14] B. Scarlini, T. Pasini, R. Navigli, Sense-annotated corpora for word sense disambiguation in multiple languages and domains, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp.

5905–5911. URL: https://aclanthology.org/2020.lrec-1.723.

[15] M. Palmer, H. T. Dang, C. Fellbaum, Making fine-grained and coarse-grained sense distinctions, both manually and automatically, Natural Language Engineering 13 (2006) 137 – 163. URL: https://api.semanticscholar.org/CorpusID:18376438.

[16] O. Sainz, O. L. de Lacalle, E. Agirre, G. Rigau, What do language models know about word senses? zero-shot WSD with language models and domain inventories, in: G. Rigau, F. Bond, A. Rademaker (Eds.), Proceedings of the 12th Global Wordnet Conference, 2023, pp. 331–342. URL: https://aclanthology.org/2023.gwc-1.40.

[17] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program, Proceedings. AMIA Symposium (2001) 17–21. URL: https://api.semanticscholar.org/CorpusID:14187105.

[18] W. Zhang, C. L. Tan, Y. C. Sim, J. Su, Nus-i2r: Learning a combined system for entity linking, Theory and Applications of Categories (2010). URL: https://api.semanticscholar.org/CorpusID:18583658.

[19] J. G. Moreno, R. Besançon, R. Beaumont, E. D'hondt, A.-L. Ligozat, S. Rosset, X. Tannier, B. Grau, Combining word and entity embeddings for entity linking, in: The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 – June 1, 2017, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2017, p. 337–352. URL: https://doi.org/10.1007/978-3-319-58068-5_21. doi:10.1007/978-3-319-58068-5_21.

[20] O. Adjali, R. Besançon, O. Ferret, H. L. Borgne, B. Grau, Multimodal entity linking for tweets, Advances in Information Retrieval 12035 (2020) 463 – 478. URL: https://api.semanticscholar.org/CorpusID:215746363.

[21] N. Heist, H. Paulheim, Nastylinker: Nil-aware scalable transformer-based entity linker, in: The Semantic Web: 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2023, p. 174–191. URL: https://doi.org/10.1007/978-3-031-33455-9_11. doi:10.1007/978-3-031-33455-9_11.

[22] S. Zhang, H. Cheng, S. Vashishth, C. Wong, J. Xiao, X. Liu, T. Naumann, J. Gao, H. Poon, Knowledge-rich self-supervision for biomedical entity linking, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, 2022, pp. 868–880. URL: https://aclanthology.org/2022.findings-emnlp.61. doi:10.18653/v1/2022.findings-emnlp.61.

[23] W. Hu, J. Chen, Y. Qu, A self-training approach for resolving object coreference on the semantic web, in: Proceedings of the 20th International Conference on World Wide Web, WWW '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 87–96. URL: https://doi.org/10.1145/1963405.1963421. doi:10.1145/1963405.1963421.

[24] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Advances in Neural Information Processing Systems, volume 26, 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

[25] B. Zhu, T. Bao, R. Han, H. Cui, J. Han, L. Liu, T. Peng, An effective knowledge graph entity alignment model based on multiple information, Neural Netw. 162 (2023) 83–98. URL: https://doi.org/10.1016/j.neunet.2023.02.029. doi:10.1016/j.neunet.2023.02.029.