Creating a Multilingual Wide-Coverage Multi-Layered Semantically Annotated Corpus

Simone Conia **Edoardo Barba** **Abelardo Carlos Martinez Lorenzo***
Pere-Lluís Huguet Cabot* **Riccardo Orlando** **Luigi Procopio** **Roberto Navigli**
Sapienza University of Rome
first.lastname@uniroma1.it

Relevant UniDive working groups: WG3

1 Introduction

NLU tasks link text to explicit knowledge, such as Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing, and Relation Extraction. Integrating curated knowledge into deep learning models benefits Language Modeling and Machine Translation. However, to the best of our knowledge, we currently lack large high-quality datasets annotated with different types of word- and sentence-level semantics. As a result, this shortcoming also hinders the investigation of how different types of explicit knowledge can interact with each other, and the potential advantages that might result from such interactions.

To tackle these challenges, we present a novel resource aimed to offer a comprehensive repository of high-quality semantic annotations, and designed for the exploration and modeling of explicit semantics on a broad scale and across various languages. The key goal is to enhance and fortify research endeavors in the field, marking our resource as a valuable resource for advancing the understanding of explicit semantics.

Crucially, our resource drops the requirement of licensed datasets and represents a key step towards a level playing field across languages and tasks. Figure 1 provides an example of the annotations available in our resource for each of the tasks considered.

2 Tasks

Word Sense Disambiguation. Word Sense Disambiguation (WSD) involves assigning a word in context its most appropriate meaning from a predefined sense inventory, such as WordNet (Miller, 1992). Despite the challenges posed by WSD (Bevilacqua et al., 2021b), linking words to specific senses offers advantages such as improved Language Modeling (Levine et al., 2020; Barba et al., 2023), Machine Translation (Campolungo et al., 2022b), and benchmarking lexical bias in applications (Campolungo et al., 2022a). Traditional datasets for WordNet-based WSD are constrained in size, outdated, and predominantly focused on English, while generating high-quality training datasets for non-English languages remains challenging (Pasini, 2020). Our resource addresses two key limitations in current WSD approaches: limited sense coverage in English datasets, and a lack of high-quality annotations for non-English languages.

Semantic Role Labeling. Semantic Role Labeling (SRL) is the task of answering questions about actions in a sentence, such as "*who* did *what* to *whom*, *where*, *when*, and *how*?" (Màrquez et al., 2008). SRL involves four subtasks: predicate identification, predicate sense disambiguation, argument identification, and argument classification. Predicate senses and semantic roles are defined based on inventories like PropBank (Palmer et al., 2005), FrameNet, or VerbAtlas (Di Fabio et al., 2019). Our resource tackles three shortcomings in SRL: lack of open data, limited multilingual coverage, and absence of cross-inventory annotations. Our resource provides annotations using PropBank and VerbAtlas, creating a high-quality, large-scale silver corpus for SRL.

Semantic Parsing. Semantic Parsing (SP) involves encoding the meaning of a sentence into a machine-interpretable structure (Kate and Wong, 2010), focusing on formalisms like Abstract Meaning Representation (Banarescu et al., 2013, AMR). Our resource addresses SP challenges, benefiting from recent advances in accuracy. Despite progress, SP faces issues akin to WSD and SRL – i.e., shortage of open datasets in English and other languages and lack of domain diversity – compounded by manual effort in annotating text with complex graphs. Our resource aims to alleviate SP challenges by creating the first large-scale dataset with AMR annotations in five languages.

*Work carried out at Babelscape, Italy.

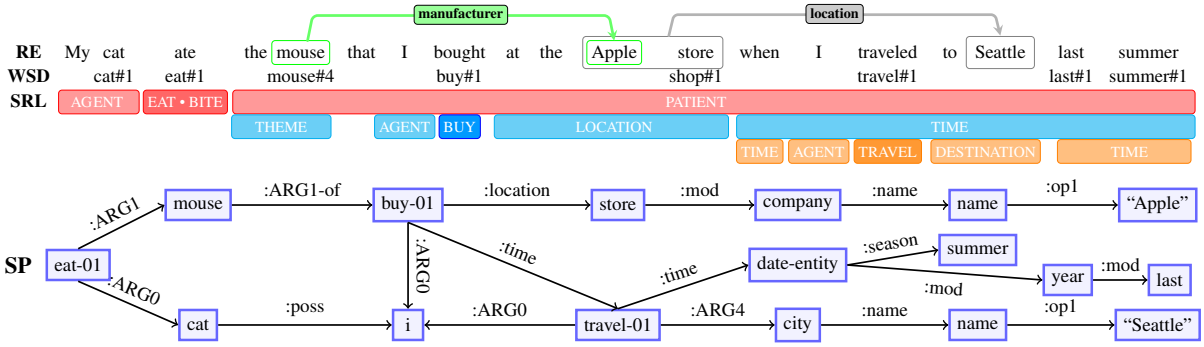

Figure 1: A visualization of the annotations (clickable) in our corpus for multilingual Word Sense Disambiguation (WSD), Semantic Role Labeling (SRL), Semantic Parsing (SP), and Relation Extraction (RE).

Relation Extraction. Relation Extraction (RE) involves extracting semantic relationships between entities from unstructured text. While prior efforts have concentrated on named entities (Roth and Yih, 2004; Riedel et al., 2010), our resource extends this focus to include equally significant relations between concepts. This enables wide-coverage extraction of relation triplets with their lexical-semantic context, which was previously challenging without the integration of WSD. Our resource also contributes to augmenting Wikidata by exploring the presence of these relations in the database and potentially providing new, reliable relation annotations. It also offers coverage, addressing the scarcity of high-quality, high-recall, multilingual datasets.

3 Methodology

3.1 Data collection and preprocessing

We create our resource by collecting articles from Wikipedia in five languages: English, French, German, Italian, and Spanish. We choose Wikipedia due to its permissive license, and because it is often already included as part of the pretraining corpus of language models (Devlin et al., 2019; Liu et al., 2019). We then employ a selection strategy in which only Wikipedia articles present in all five considered languages are retained. This is motivated by the assumption that articles appearing in multiple languages receive greater attention from the Wikipedia community. Moreover, this enables cross-lingual comparability. Despite this, the filtering process results in 440,000 articles per language. We then preprocess each article using Stanza NLP for sentence splitting, tokenization, lemmatization, and part-of-speech tagging. This results in 7M to 20M sentences, and from 190M to 518M tokens, depending on the language.

3.2 Construction and annotation

WSD. We tag each document using ESCHER, a state-of-the-art WSD model (Barba et al., 2021). We use BabelNet 5 (Navigli et al., 2021) as our default unified sense inventory to annotate in multiple languages.

SRL. For SRL, we adopt Multi-SRL (Conia and Navigli, 2020), a state-of-the-art system for PropBank-style dependency- and span-based SRL. We leverage the mapping from PropBank to VerbAtlas frames created by Di Fabio et al. (2019) to train Multi-SRL also to predict VerbAtlas-style labels, thus labeling each predicate using two inventories.

SP. We follow Blloshmi et al. (2020) and extend SPRING (Bevilacqua et al., 2021a) – a popular system for English AMR parsing (Bai et al., 2022; Yu and Gildea, 2022; Cheng et al., 2022) – to the multilingual setting. SPRING is an auto-regressive model fine-tuned to "translate" natural sentences into linearized AMR graphs.

RE. For RE, we employ a two-step approach based on the cRocoDiLe data extraction pipeline (Huguet Cabot and Navigli, 2021), which collects relation triplets from Wikipedia articles by detecting entity mentions and connecting them with the relations defined in Wikidata. We perform inference on our corpus using mREBEL (Huguet Cabot et al., 2023), a RE model based on mBART, to obtain our annotated RE data. We then exploit the Wikipedia hyperlinks and WSD annotations as sources of disambiguated entities and concepts with which we extract the second portion of annotations by applying the cRocoDiLe pipeline.

Acknowledgements

Simone Conia, Edoardo Barba, Pere-Lluís Huguet Cabot and Roberto Navigli gratefully acknowledge the PNRR MUR project PE0000013-FAIR / PE01-FAIR-SPOKE-5-DIAG CUP B53C22003980006.

References

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proc. of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*.

Laura Banarescu, Claire Bonial, and Shu et al. Cai. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Edoardo Barba, Niccolò Campolungo, and Roberto Navigli. 2023. Reducing disambiguation biases in NMT by leveraging explicit word sense information. In *Findings of the 2023 Conference of the Association for Computational Linguistics: Human Language Technologies*, Toronto, Canada. Association for Computational Linguistics.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proc. of the 2021 Conference of the NAACL: Human Language Technologies*.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021a. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. *Proc. of the AAAI Conference on Artificial Intelligence*, 35(14).

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021b. Recent trends in Word Sense Disambiguation: A survey. In *Proc. of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proc. of the 2020 Conference on EMNLP*, pages 2487–2500.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022a. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proc. of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*.

Niccolò Campolungo, Tommaso Pasini, Denis Emelin, and Roberto Navigli. 2022b. Reducing disambiguation biases in NMT by leveraging explicit word sense information. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4824–4838, Seattle, United States. Association for Computational Linguistics.

Ziming Cheng, Zuchao Li, and Hai Zhao. 2022. BiBL: AMR parsing and generation with bidirectional Bayesian learning. In *Proc. of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proc. of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proc. of the 2019 Conference on EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the EMNLP 2021*.

Pere-Lluis Huguet Cabot, Simone Tedeschi, Axel-Cyrile Ngonga Ngomo, and Roberto Navigli. 2023. REDFM: a filtered and multilingual relation extraction dataset. In *Proc. of the 61st Annual Meeting of the Association for Computational*.

Rohit J. Kate and Yuk Wah Wong. 2010. Semantic parsing: The task, the state of the art and the future. In *Proc. of the 48th Annual Meeting of the ACL: Tutorial Abstracts*.

Yoav Levine, Barak Lenz, and Or et al. Dagan. 2020. SenseBERT: Driving some sense into BERT. In *Proc. of the 58th Annual Meeting of the ACL*.

Yinhan Liu, Myle Ott, and Naman Goyal et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic Role Labeling: An introduction to the special issue. *Computational Linguistics*, 34(2).

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli, Michele Bevilacqua, and Simone et al. Conia. 2021. Ten years of BabelNet: A survey. In *Proc. of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proc. of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*.

Chen Yu and Daniel Gildea. 2022. Sequence-to-sequence AMR parsing with ancestor information. In *Proc. of the 60th Annual Meeting of the ACL (Volume 2: Short Papers)*.