MOSAICo: a Multilingual Open-text Semantically Annotated Interlinked Corpus

Simone Conia[1], **Edoardo Barba**[1], **Abelardo Carlos Martínez Lorenzo**[1], **Pere-Lluís Huguet Cabot**[1], **Riccardo Orlando**[1], **Luigi Procopio**[2], **Roberto Navigli**[1]*
[1]Sapienza NLP Group, Sapienza University of Rome
[2]Litus AI, Italy
[1]{first.lastname(s)}@uniroma1.it
[2]luigi.procopio@litus.ai

Abstract

Several Natural Language Understanding (NLU) tasks focus on linking text to explicit knowledge, including Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing, and Relation Extraction. In addition to the importance of connecting raw text with explicit knowledge bases, the integration of such carefully curated knowledge into deep learning models has been shown to be beneficial across a diverse range of applications, including Language Modeling and Machine Translation. Nevertheless, the scarcity of semantically-annotated corpora across various tasks and languages limits the potential advantages significantly. To address this issue, we put forward MOSAICo, the first endeavor aimed at equipping the research community with the key ingredients to model explicit semantic knowledge at a large scale, providing hundreds of millions of silver yet high-quality annotations for four NLU tasks across five languages. We describe the creation process of MOSAICo, demonstrate its quality and variety, and analyze the interplay between different types of semantic information. MOSAICo, available at https://github.com/SapienzaNLP/mosaico, aims to drop the requirement of closed, licensed datasets and represents a step towards a level playing field across languages and tasks in NLU.

1 Introduction

There are two predominant schools of thought on how to integrate semantics into Natural Language Processing (NLP) systems. Most researchers lean towards the implicit integration of semantics through self-supervised language modeling (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020; Lewis et al., 2020), a framework that proved to be successful in numerous language understanding benchmarks (Rajpurkar et al., 2016; Wang et al., 2019b,a). However, this approach – while conceptually straightforward and effective by scaling up the number of parameters – has raised concerns regarding its cost-effectiveness and environmental impact (Strubell et al., 2019; Geiping and Goldstein, 2022). Furthermore, the black-box nature of this approach has also prompted calls for a critical inquiry into whether the impressive results on language benchmarks genuinely reflect a system's comprehension of language (Bender et al., 2021; Ray Choudhury et al., 2022; Maru et al., 2022; Tedeschi et al., 2023).

Conversely, other researchers have advocated for the integration of discrete symbols into AI systems, which has recently been resurfacing under the umbrella term of "neuro-symbolic AI" (d'Avila Garcez and Lamb, 2020; Raedt et al., 2020). In NLP, discrete symbols can be used to represent word meanings, and graphs of symbols can be interpreted as sentence-level semantics. The integration of such explicit semantics into NLP systems may have the potential to address the aforementioned concerns (Navigli, 2018). Indeed, explicit semantics has been shown to be effective in reducing the number of parameters of neural networks and in improving the interpretability of their outputs. However, to the best of our knowledge, we still lack a critical enabler for training such approaches, namely, a vast, high-quality dataset annotated with symbolic knowledge.

To address this issue, we introduce MOSAICo, a novel resource designed to foster explorations and modeling of explicit semantics on an extensive scale and across multiple languages. With its hundreds of millions of silver annotations on Wikipedia sentences across 5 languages for 4 important NLU tasks – namely, Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing, and Relation Extraction – MOSAICo is a large, diverse, and

*All authors contributed equally. The core of the work by Carlos and Pere-Lluís was carried out while working at Babelscape. Part of the work by Luigi was carried out while at the Sapienza University of Rome.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7990–8004

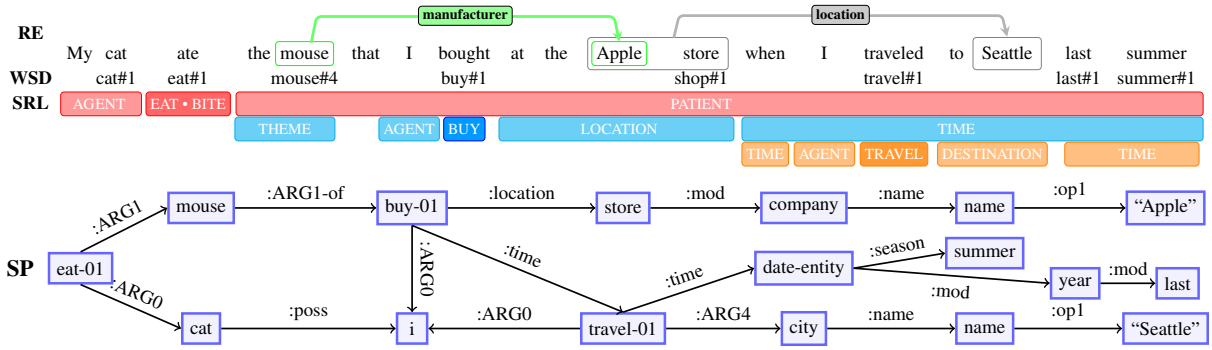

Figure 1: A visualization of the annotations (clickable) in MOSAICo for multilingual Word Sense Disambiguation (WSD), Semantic Role Labeling (SRL), Semantic Parsing (SP), and Relation Extraction (RE).

robust dataset that can bootstrap research on the integration of explicit semantics into NLP systems to train and evalute models. In addition, MOSAICo is also a tool for investigating the interaction between different types of symbols, structures, and semantics, as shown in Figure 1.

In this paper, we describe the construction of MOSAICo, which can save critical time and compute for future research, and show its effectiveness for training state-of-the-art systems on open data as opposed to the closed and copyrighted datasets that are traditionally used in the literature, which will aid in lowering the barrier for the development of next-generation systems. We also provide insights into the interaction between different NLU tasks, as diverse sources of semantic knowledge allow us to look at new opportunities enabled by the availability of a large, cross-lingual, semantically-tagged corpus. We hope MOSAICo will be a tool that facilitates deeper investigations of the intersection between natural language and explicit semantics, while also fostering analysis of the connection between different NLU tasks across languages.

2 Background and Motivation

This section briefly introduces the four semantic tasks we considered in MOSAICo, and further explains how our resource can help the research community. Due to space constraints and the extensive literature available for each task, we focus on the works that are within the scope of this paper.

Word Sense Disambiguation (WSD) aims at associating a word in context with the most suitable meaning from a predefined sense inventory (Navigli, 2009; Bevilacqua et al., 2021b). WSD is a challenging task for machines and humans, especially when using fine-grained sense inventories like WordNet (Miller, 1992). Nevertheless, grounding words to discrete senses can bring several benefits, including better Language Modeling (Levine et al., 2020; Barba et al., 2023), Machine Translation (Campolungo et al., 2022b), and lexical bias benchmarking in applications (Campolungo et al., 2022a). However, traditional datasets for WordNet-based WSD are limited in size, outdated, and mostly English-specific. Even though sense inventories are available for multilingual WSD, such as BabelNet (Navigli et al., 2021), creating high-quality training datasets for non-English languages is difficult (Pasini, 2020). MOSAICo aims to overcome two limitations in today's WSD: i) sense coverage, i.e., current English datasets provide limited coverage of the long tail of the sense distribution, and ii) multilingual coverage, i.e., the lack of high-quality annotations for non-English languages.

Semantic Role Labeling (SRL) is informally described as the task of answering "*who* did *what* to *whom*, *where*, *when*, and *how*?" (Màrquez et al., 2008). Given a sentence, SRL usually requires: i) predicate identification, i.e., finding all the words that denote an action; ii) predicate sense disambiguation, i.e., selecting the most appropriate sense for each identified predicate; iii) argument identification, i.e., detecting the text spans that represent the participants in the action; and, iv) argument classification, i.e., assigning a *semantic role* to each predicate-argument pair. The predicate senses and semantic roles used in SRL are defined according to an inventory of predicate-argument structures, e.g., PropBank (Palmer et al., 2005; Pradhan et al., 2022), FrameNet (Baker et al., 1998), or VerbAtlas (Di Fabio et al., 2019), among others. MOSAICo provides annotations using PropBank – due to its popularity – and VerbAtlas – due to its manual linkage to WordNet and BabelNet and its

cross-frame relations – and addresses three shortcomings of SRL: i) open data availability, i.e., popular datasets are not open (Carreras and Màrquez, 2004; Hajič et al., 2009; Pradhan et al., 2012), ii) multilingual coverage, i.e., limited availability of large multilingual datasets, and iii) cross-inventory annotations, i.e., data annotated with more than one type of predicate-argument structure inventory.

Semantic Parsing (SP) is the task of encoding the meaning of a sentence in a machine-interpretable structure (Kate and Wong, 2010). Since the focus of MOSAICo is on semantics, we direct our attention to formalisms that encode the abstract meaning of a sentence, such as AMR (Banarescu et al., 2013) and BMR (Navigli et al., 2022; Martínez Lorenzo et al., 2022). Thanks to recent advances (Bevilacqua et al., 2021a), modern SP systems can now be used with high accuracy. However, SP still suffers from issues similar to those in WSD and SRL, perhaps amplified by the extent of the manual effort required to annotate text with complex graphs. More specifically, the role of MOSAICo is to provide the first large-scale dataset of annotations for AMR in five languages with the aim of mitigating two problems in SP: i) shortage of open datasets in English but also in other languages, and ii) coverage, i.e., the lack of semantic parses for a broad range of domains.

Relation Extraction (RE) is the task of extracting semantic relationships between entities from unstructured text. Since MOSAICo is based on Wikipedia, we focus on relation types from Wikidata. While previous work has mostly focused on named entities (Roth and Yih, 2004; Riedel et al., 2010), equally important relations are also present between concepts. MOSAICo allows us to explore interactions between entities identified through Wikipedia hyperlinks, and, unlike previous efforts, concepts identified through WSD, so as to disambiguate the subject and object of a semantic relation from Wikidata. This enables the creation of a vast network of concepts and entities interconnected through semantic relations tied to their multilingual textual context. To leverage them, we will explore how many of these relations are already present in Wikidata and whether MOSAICo can provide new and reliable relation annotations. Therefore, MOSAICo provides for RE: i) wide-coverage extraction of relational triplets[1] connected to their lexical-semantic context, which was unfeasible without the interconnection with WSD at hand, ii) a novel way of augmenting Wikidata, and iii) multilingual coverage, i.e., addressing the limited availability of high-quality, high-recall, multilingual datasets for RE.

[1]We define a triplet as a tuple of three elements: a subject, an object, and the relation type that connects them.

3 MOSAICo

Our main objective with MOSAICo is to create a large collection of silver annotations on multilingual data. However, we believe that the value of MOSAICo does not only lie in the creation of annotations whose quality is comparable to gold-standard datasets. Indeed, while several groups may have the resources and expertise to reproduce similar efforts, we introduce MOSAICo to also address the following issues:

- Large silver datasets require compute: by making our resource available to the wider community, researchers can save more than 10 thousand GPU hours;
- Processing from scratch: even if compute were not a limitation, recreating a large dataset from scratch is not efficient and green;
- Common reference: the open availability allows our resource to be less affected by hidden implementation details;
- Task interaction: having uniform annotations for four types of task enables the analysis of possible inter-task interactions.

We now describe the creation process of MOSAICo, our novel resource of silver-quality annotations for modeling semantics at a large scale. We start with an outline of the data collection process (Section 3.1) and then detail how we produce the annotations (Section 3.2).

3.1 Data collection and preprocessing

We create MOSAICo by collecting documents from Wikipedia in five languages: English, French, German, Italian, and Spanish.[2] There are several reasons for choosing Wikipedia as the core of MOSAICo. First, Wikipedia is released under a permissive license. Second, Wikipedia is often already included as part of the pretraining corpus of modern language models (Devlin et al., 2019; Liu et al., 2019). Providing annotations for textual data that

[2]We use the Wikipedia dump released on June 13, 2022.

		EN	DE	ES	FR	IT
MOSAICo	#documents	441K	441K	441K	441K	441K
	#sentences	19.6M	16.4M	7.1M	11.8M	8.6M
	#tokens	518M	335M	190M	319M	261M
	avg. sent. length	26	20	27	27	30
M-Core	#documents	17.2K	17.2K	17.2K	17.2K	17.2K
	#sentences	2.6M	2.2M	0.7M	1.3M	1.1M
	#tokens	69M	46M	18M	35M	36M
	avg. sent. length	27	21	28	28	31

Table 1: Number of documents, sentences, tokens, and average sentence length in tokens in MOSAICo and MOSAICo Core (M-Core).

		EN	DE	ES	FR	IT
MOSAICo	WSD	192.0M	83.1M	61.0M	102.2M	84.0M
	SRL	114.5M	44.2M	29.3M	42.8M	54.5M
	SP	17.6M	13.6M	5.5M	8.7M	6.3M
	RE	18.2M	6.9M	3.4M	7.0M	3.8M
M-Core	WSD	5.6M	2.9M	2.2M	3.4M	2.9M
	SRL	4.8M	2.2M	1.3M	1.8M	1.7M
	SP	2.5M	2.1M	0.6M	1.2M	1.0M
	RE	3.6M	1.2M	0.7M	1.4M	0.7M

Table 2: Overview of the number of annotations for each semantic task in MOSAICo and MOSAICo Core.

these language models have already used opens the door to easier future integration of explicit semantics into such models. Second, we aim to retain information about the provenance of each sentence, i.e., the Wikipedia article it belongs to and its position within the article. This design choice allows a user to connect and leverage annotations at the document level and for the same article across different languages.

Preprocessing. First, for each language, we retain only those Wikipedia articles that appear in all the five languages we consider. There are two main reasons for this selection strategy. First, we hypothesize that the larger the number of languages an article is written in, the greater must be the attention that the Wikipedia community has devoted to its creation. Second, forcing an article to be available in all five languages of interest makes MOSAICo "comparable" by design, i.e., fully parallel as regards the articles (concepts and entities) covered in each language, even if not parallel (i.e., only comparable) as regards the textual contents. Despite the above cross-lingual constraint, the filtering process still results in 440,000 articles per language. Finally, we preprocess each article using Stanza NLP[3] for sentence splitting, tokenization, lemmatization, and part-of-speech tagging. As shown in Table 1, MOSAICo contains millions of sentences and tokens across the five languages under consideration.

Since MOSAICo includes hundreds of thousands of articles for each language, we also identify a subset of these which includes those articles whose quality is higher than the average, i.e., the articles labeled as "good" or "featured" in at least one of the five languages by the Wikipedia community.[4] We name this subset MOSAICo Core.

[3] https://stanfordnlp.github.io/stanza/

[4] Wikipedia articles need to satisfy a set of strict quality requirements to be considered "good" or "featured".

3.2 Construction and annotation

We now describe how we annotated the collected texts across the four semantic tasks. One key aspect of the construction of MOSAICo is that producing a massive quantity of annotations, as can be seen in Table 2, is time-consuming. Therefore, an indirect contribution of MOSAICo is that its open availability can save thousands of hours of computation, enabling researchers to bootstrap their future work.[5]

Word Sense Disambiguation. We tag each document in MOSAICo using ESCHER, a state-of-the-art WSD model (Barba et al., 2021). ESCHER takes advantage of word sense definitions to better generalize on unseen patterns and inventories. This is a strong desideratum in our case, as Wikipedia features many words and senses that are not included in traditional training datasets (Miller et al., 1994; Pasini et al., 2021, SemCor, XL-WSD). We train ESCHER on SemCor and the WordNet Gloss Corpus (WGC) using `deberta-v3-base` for English and on the XL-WSD training datasets using `mdeberta-v3-base` for the other languages. We use BabelNet 5.1[6] (Navigli et al., 2021) as our default unified sense inventory to annotate in multiple languages.

Semantic Role Labeling. For SRL, we adopt Multi-SRL (Conia and Navigli, 2020), a state-of-the-art system for PropBank-style dependency- and span-based SRL. In this work, we focus on span-based SRL, as dependency-based formalisms are tied to syntax and, therefore, are more difficult to employ in downstream applications. Moreover, we leverage the mapping from PropBank to VerbAtlas frames created by Di Fabio et al. (2019) to train

[5] The annotation process of MOSAICo requires about 12 470 hours on Nvidia RTX 3090: 1050 for preprocessing, 1140 for WSD, 1600 for SRL, 5080 for SP, and 3600 for RE.

[6] https://babelnet.org

Multi-SRL to predict VerbAtlas-style labels as well, thus labeling each predicate using two inventories.

Semantic Parsing. For the English sentences, we leverage the state-of-the-art English AMR parser LeakDistill (Vasylenko et al., 2023),[7] which is an extension of the SPRING model (Bevilacqua et al., 2021a) – a popular system for AMR parsing (Bai et al., 2022; Martínez Lorenzo et al., 2023a; Lou and Tu, 2023; Gao et al., 2023). For the multilingual setting, we extend CLAP (Martínez Lorenzo and Navigli, 2024), an efficient implementation of SPRING. LeakDistill and CLAP are autoregressive models fine-tuned to "translate" natural sentences into linearized AMR graphs. To extend CLAP cross-lingually: i) we follow Blloshmi et al. (2020) and create a training corpus for French, German, Italian, and Spanish from the automatic translation[8] of the original English sentences in AMR 3.0 paired with the original (English) AMR graphs; then, ii) we fine-tune an mBART-based (Liu et al., 2020) instance of SPRING for each language.

Relation Extraction. For RE, we employ a two-step approach based on the mREBEL multilingual mBART-based RE system (Huguet Cabot et al., 2023) and the cRocoDiLe data extraction pipeline (Huguet Cabot and Navigli, 2021). In the first step, we perform inference on the corpus of MOSAICo using mREBEL to obtain the first portion of our annotated RE data. In the second step, we exploit the Wikipedia hyperlinks – along with their propagations[9] – and WSD annotations as sources of disambiguated entities and concepts. We extract the second portion of annotations by applying the cRocoDiLe pipeline which collects relational triplets from Wikipedia articles by leveraging the disambiguated entity and concept mentions and connecting them with the relations between them defined in Wikidata. Unlike previous work, we do not restrict this extraction to Wikipedia abstracts and extract the triplets from the entire pages. Finally, we remove false positives by utilizing an NLI-based Triplet Critic (Huguet Cabot et al., 2023).

4 Experiments and Results

To evaluate the quality of MOSAICo, we train a set of state-of-the-art models on our silver datasets for WSD, SRL, RE, and SP; then, we compare their results with those obtained by the same systems trained on gold datasets. Table 9 in the Appendix provides an overview of these benchmarks and resources in comparison to MOSAICo.

4.1 Word Sense Disambiguation

We evaluate the effectiveness of MOSAICo for WSD by testing ESCHER when trained under three different settings: i) using SemCor and WGC for English and the training datasets of XL-WSD for the other 4 languages (ESCHER$_{gold}$),[10] ii) using a randomly sampled subset of MOSAICo of the same size as the gold training sets (ESCHER$_{M\text{-}Ref}$), iii) using MOSAICo Core (ESCHER$_{M\text{-}Core}$), and iv) using a randomly sampled subset of MOSAICo Core of the same size as the gold training sets (ESCHER$_{M\text{-}Core\text{-}Ref}$).

Table 3 provides an overview of the results on ALL (Raganato et al., 2017), 42D (Maru et al., 2022), and XL-WSD test sets, which are the standard evaluation framework for WordNet-based WSD, a new multi-domain challenge set for rare senses, and the largest benchmark for multilingual WSD, respectively. As expected, ESCHER$_{M\text{-}Ref}$ and ESCHER$_{M\text{-}Core\text{-}Ref}$ do not achieve the same results as ESCHER$_{gold}$ (74.9 and 76.0 vs. 76.4 in F1 score averaged across all benchmarks), but we can see that the higher quality of MOSAICo Core leads to higher overall performance. When we move to the larger scale of the whole MOSAICo Core dataset, ESCHER$_{M\text{-}Core}$ improves over ESCHER$_{gold}$ in every test set and language (77.3 vs. 76.4 on average), confirming the quality of our WSD annotations.

Since WSD is not as computationally expensive as the other tasks, here we also analyze the difference in performance when training on MOSAICo Core compared to a random 10% of MOSAICo (ESCHER$_{M\text{-}10\%}$) and the entire MOSAICo (ESCHER$_{M\text{-}100\%}$). The last two rows of Table 3 show that ESCHER$_{M\text{-}Core}$ compares favorably against both ESCHER$_{M\text{-}10\%}$ and ESCHER$_{M\text{-}100\%}$, suggesting that MOSAICo Core features a good balance between size and quality. Finally, it is worth noting that, for the rare senses in 42D, coverage is the most important aspect, as ESCHER$_{M\text{-}100\%}$ outperforms both ESCHER$_{M\text{-}Core}$ and ESCHER$_{gold}$ (58.4 vs. 56.2 vs. 54.4, respectively).

[7] This model does not scale cross-lingually.

[8] We employ DeepL to obtain high-quality translations.

[9] We propagate links to the other mentions of the same entity as in Tedeschi et al. (2021).

[10] We note that the XL-WSD training datasets are silver-quality datasets and that we refer to them as *gold* just to be consistent with the naming convention of the other tasks.

	ALL EN	42D EN	XL-WSD DE	XL-WSD ES	XL-WSD FR	XL-WSD IT	Avg.
ESCHER$_{gold}$	81.0	54.4	83.2	77.5	84.3	78.2	76.4
ESCHER$_{M\text{-}Ref}$	79.0	51.5	81.2	76.8	83.8	77.3	74.9
ESCHER$_{M\text{-}Core}$	**82.0**	56.2	**84.1**	**77.8**	**84.5**	**79.1**	**77.3**
ESCHER$_{M\text{-}Core\text{-}Ref}$	80.5	53.7	82.7	77.5	83.9	77.7	76.0
ESCHER$_{M\text{-}10\%}$	81.7	55.9	83.8	77.6	84.4	78.8	77.0
ESCHER$_{M\text{-}100\%}$	81.5	**58.4**	83.8	77.4	84.1	78.8	**77.3**

Table 3: WSD results in terms of F1 score. Best in **bold**.

		ON5 EN	PBE EN	X-SRL EN	X-SRL DE	X-SRL ES	X-SRL FR	Avg.
Senses	M-SRL$_{gold}$	**95.5**	80.7	**97.2**	67.6	75.1	71.1	81.2
	M-SRL$_{M\text{-}Ref}$	94.8	80.1	96.5	67.0	74.9	71.2	80.8
	M-SRL$_{M\text{-}Core}$	95.3	**82.6**	97.0	**68.4**	**75.5**	**72.3**	**81.9**
Roles	M-SRL$_{gold}$	**87.3**	74.9	**92.0**	69.5	75.1	72.1	78.4
	M-SRL$_{M\text{-}Ref}$	85.5	74.7	91.4	69.3	75.0	71.8	77.9
	M-SRL$_{M\text{-}Core}$	87.0	**75.7**	**92.0**	**70.6**	**75.5**	**72.2**	**78.8**

Table 4: SRL results in terms of F1 score on the test sets of OntoNotes (ON5), PropBank Examples (PBE), and X-SRL. Top half: predicate sense disambiguation. Bottom half: argument labeling. Best results in **bold**.

4.2 Semantic Role Labeling

We assess the quality of the SRL annotations in MOSAICo by measuring their impact on Multi-SRL. Table 4 shows the results of Multi-SRL when trained on three sources of SRL annotations: i) OntoNotes 5.0, i.e., the most recent gold standard for PropBank 3 (Pradhan et al., 2022) and derived from CoNLL-2012, ii) a random subsample of the MOSAICo annotations of the same size as the gold training dataset (M-SRL$_{M\text{-}Ref}$), and iii) MOSAICo Core (Multi-SRL$_{M\text{-}Core}$).

We evaluate Multi-SRL trained using the above datasets on three benchmarks: on OntoNotes, which is a standard benchmark in the area, on PB-Examples (PBE), which is a comprehensive benchmark recently introduced by Orlando et al. (2023), and on X-SRL, a multilingual benchmark by Daza and Frank (2020). For each benchmark, we report the F1 score typically used in the literature, which combines the accuracy in predicate sense disambiguation with the F1 score on argument labeling.

On the OntoNotes test set, Multi-SRL$_{M\text{-}Core}$ achieves F1 scores in the same ballpark as Multi-SRL$_{gold}$ on predicate sense disambiguation (95.3 vs. 95.5 points) and argument labeling (87.0 vs. 87.3 points), even though we note that OntoNotes features different genres compared to MOSAICo Core.[11] This result acquires greater value if we consider that current datasets for SRL, such as OntoNotes, are closed-source. Instead, the open availability of MOSAICo can lift barriers for researchers looking for high-performance SRL systems, fostering future work on multilingual SRL.

PB-Examples is an out-of-domain benchmark constructed from the annotated examples in PropBank[12] and it enables us to also test the generalizability of an SRL system on verbal, nominal, adjectival, and adverbial predicates, especially the latter two types, which are not present in the OntoNotes training set. Thanks to the wide variety of text included in MOSAICo Core, Multi-SRL$_{M\text{-}Core}$ outperforms Multi-SRL$_{gold}$ in predicate sense disambiguation (82.6 vs. 80.7 in F1 score) and argument labeling (75.7 vs. 74.9 in F1 score) on PBE. To better understand how MOSAICo Core helps Multi-SRL, we conduct two fine-grained analyses on PBE. First, we split PBE by predicate type, i.e., verbal predicates, nominal predicates, and adjectival predicates. Here, we find that Multi-SRL$_{M\text{-}Core}$ generalizes better than Multi-SRL$_{gold}$ on the adjectival predicates, which are particularly challenging as they do not occur in the OntoNotes training set, achieving an absolute improvement of 11.0 (74.5 vs. 63.5) and 3.2 (59.5 vs. 56.3) points in F1 score on predicate sense disambiguation and argument labeling, respectively. Second, we analyze the results of Multi-SRL on those predicates whose ground-truth sense labels are "unseen" with respect to the OntoNotes training set, finding that Multi-SRL$_{M\text{-}Core}$ outperforms Multi-SRL$_{gold}$ by 7.0 (73.5 vs. 66.5) and 1.5 (69.4 vs. 67.8) points in F1 score on predicate sense disambiguation and argument labeling, respectively.

Finally, we find that Multi-SRL$_{M\text{-}Core}$ performs better than Multi-SRL$_{gold}$ in X-SRL, a benchmark that includes German, Spanish, and French: it achieves an average improvement of 0.8% (72.1 vs. 71.3) and 0.5% (72.8 vs 72.3) respectively in predicate disambiguation and argument labeling F1 across the 3 languages.

4.3 Semantic Parsing

For SP, we measure the efficacy of MOSAICo in text-to-AMR parsing by training our SP model in three different settings: i) the gold-standard training split of AMR 3.0 for English and on its translations for the other 4 languages (SPRING$_{gold}$), as

[11] For example, OntoNotes contains dialogues with informal writing, which are rare in the Wikipedia articles.

[12] Most of the predicate senses in PropBank come with a usage example. For example, PropBank provides the following example for *give.01*: [The executives]$_{ARG0}$ gave [the chefs]$_{ARG2}$ [a standing ovation]$_{ARG1}$.

	AMR3	TLP	Bio	AMR-4T			Avg.
	EN - ID	EN - OOD		DE	ES	IT	
SPRING$_{gold}$	**83.0**	81.0	60.5	**69.2**	**72.9**	71.6	73.0
SPRING$_{M\text{-}Ref}$	80.4	80.1	56.8	67.3	70.9	70.1	70.9
SPRING$_{M\text{-}Core}$	82.7	**81.4**	**62.0**	**69.2**	72.3	**72.9**	**73.4**

Table 5: Text-to-AMR results in terms of SMATCH score on the test set of AMR 3.0 (AMR3), The Little Prince (TLP), biomedical (Bio), and AMR 2.0 – Four Translations (AMR-4T). Best results in **bold**.

Model	CONLL04	NYT	ADE	Avg.
BART$_{gold}$	71.2	91.8	81.7	81.6
BART$_{REBEL}$	**75.4**	**92.0**	**82.2**	**83.2**
BART$_{M\text{-}cRoco}$	72.7	91.6	81.5	81.9
BART$_{M\text{-}Core}$	**75.4**	91.9	81.9	83.1

Table 6: RE results in terms of F1 score. Best in **Bold**.

explained in Section 3.2, ii) a random subsample of the MOSAICo annotations of the same size as AMR 3.0 (SPRING$_{M\text{-}Ref}$), and iii) the silver annotations from MOSAICo Core (SPRING$_{M\text{-}Core}$). We compare the results on the test sets of AMR 3.0 (LDC2020T02), the out-of-domain benchmarks of The Little Prince (TLP) and Bio AMR, and the multilingual test set of "AMR 2.0 – Four Translations" (LDC2020T07) in German, Italian, and Spanish. We use the well-established SMATCH metric (Cai and Knight, 2013) – which calculates the maximum overlap between two graphs – to evaluate the performance of the systems.

Table 5 presents the results in SP. Comparing SPRING$_{M\text{-}Ref}$ to SPRING$_{gold}$, we can see that, as in previous tasks, the model trained on the gold training dataset outperforms our baseline. However, similar to WSD and SRL, although the genres in MOSAICo Core and the test set of AMR 3.0 do not match, SPRING$_{M\text{-}Core}$ achieves results that are comparable to SPRING$_{gold}$ (82.7 vs. 83.0 in SMATCH score, respectively). Moreover, in out-of-domain benchmarks, due to its wide range of texts and domains, SPRING$_{M\text{-}Core}$ provides improvements over SPRING$_{gold}$ of 0.4 points on TLP (81.4 vs. 81.0), and 1.5 on Bio (62.0 vs. 60.5).

Finally, in the multilingual setting, SPRING$_{M\text{-}Core}$ and SPRING$_{gold}$ achieve results in the same ballpark for German (69.2 vs. 69.2), Spanish (72.3 vs. 72.9), and Italian (72.9 vs. 71.6). Interestingly, once again, since "AMR 2.0 — Four Translations" sentences are human translations of a portion of the AMR 3.0 test sentences, SPRING$_{M\text{-}Core}$ is evaluated in an out-of-genre setting, while SPRING$_{gold}$ in an in-genre one. These findings gain even more relevance considering that current training datasets for SP are closed-source.

4.4 Relation Extraction

The experimental setup for RE is different from the other tasks. While sense inventories can be shared across datasets, relation taxonomies differ among RE datasets, MOSAICo annotations, and more in general silver-standard resources for RE. Therefore our annotations cannot be used as direct training material to evaluate on gold corpora. For this reason the research community has relied on silver corpora to perform a first round of pretraining before fine-tuning models on a specific dataset and relation set (Yamada et al., 2020). In order to provide an extrinsic evaluation of the quality of the annotations in MOSAICo for the English language, we compare them with those used in REBEL (Huguet Cabot and Navigli, 2021) and design three different settings by pretraining BART with REBEL's same decoding scheme under i) the silver data provided by REBEL (BART$_{REBEL}$), equivalent to REBEL$_{pre-training}$ in the original paper, ii) MOSAICo Core's instances produced just with the cRocoDiLe pipeline on the Wikipedia hyperlinks as in REBEL annotation scheme (BART$_{M\text{-}cRoco}$), and iii) all the silver annotations in MOSAICo Core (BART$_{M\text{-}Core}$). As a reference, we include the performances of BART when trained only on each of the gold corpora (BART$_{gold}$). More specifically, we then evaluate the difference in performance on CoNLL04 (Roth and Yih, 2004), NYT (Riedel et al., 2010), ADE (Gurulingappa et al., 2012) datasets when further training any of the aforementioned pre-trained BART on these datasets. All results are reported as micro-F1 scores based on the labeled triplets, where a triplet is considered correct only if subject, object and relation type are correctly identified.

As we can see from the results in Table 6, BART$_{M\text{-}Core}$ and BART$_{REBEL}$ perform in the same ballpark, with the latter slightly outperforming the former, and both surpassing BART$_{gold}$. However, we highlight that BART$_{M\text{-}Core}$ is 66% smaller than BART$_{REBEL}$ (259,982 vs. 784,202 instances). We hypothesize that BART$_{M\text{-}Core}$ does not outperform BART$_{REBEL}$ due mainly to two reasons: i) the REBEL dataset differs from MOSAICo in its composition, as it is built only from Wikipedia lead sections and it features a different distribution of

	EN	IT	FR	DE	ES
ESCHER$_{gold}$	88.4	85.9	82.8	86.3	83.1

Table 7: F1 scores of ESCHER$_{gold}$ on Wiki-WSD.

relation types, which can be beneficial when transferring knowledge to smaller datasets used for evaluation, and ii) MOSAICo provides a higher density of triplets per training passage than REBEL (6.7 vs. 1.2), but the datasets used for evaluation are more similar to REBEL than MOSAICo (1.6 for ADE, 1.5 for CoNLL-2004, 1.7 for NYT). Moreover, when we compare the results of BART$_{M\text{-}cRoco}$ to those of BART$_{M\text{-}Core}$, we can see that, when the additional sources MOSAICo provides are dropped, performance decreases, reaffirming the importance of our work in providing an augmented annotation scheme that includes concepts and not only entities. Despite the slightly lower performance, a system trained on MOSAICo is able to extract more triplets per instance, and thanks to a larger annotation coverage, we believe it will enable richer RE systems to be built not just in English, but also in French, Italian, German and Spanish.

5 Insights and Opportunities

Here, we highlight the opportunities MOSAICo brings to interconnect the different types of knowledge captured by each semantic task.

MOSAICo is a test for multilingual WSD. Indeed, one feature often overlooked in Wikipedia is that, in each article, the first mention of a term has been manually hyperlinked to the corresponding Wikipedia article if it exists. Therefore, we can use the 1:1 mapping (Navigli et al., 2021) between Wikipedia and BabelNet to convert Wikipedia interlinks to BabelNet synsets and create Wiki-WSD, a large novel benchmark for multilingual WSD. For reference, we employ Wiki-WSD to evaluate ESCHER$_{gold}$, as shown in Table 7. Wiki-WSD is the largest available WSD benchmark for English, Italian, French, German, and Spanish, including almost 6.7 million crowdsourced sense-tagged instances compared to 14,166 instances in XL-WSD. While we stress that Wiki-WSD is restricted to nominal senses, we argue that, thanks to its large coverage, this benchmark can be employed for evaluating WSD systems, providing new insights into multilingual disambiguation, and for studying the interplay between different sense inventories.

Interconnecting WSD and SRL. A key step in SRL is that of predicate sense disambiguation, i.e., selecting the most appropriate "sense" for a predicate in context so as to make it possible to also assign appropriate semantic roles to each predicate-argument pair. We note that, in many respects, predicate sense disambiguation is equivalent to WSD. However, there has been little interest in interconnecting the two tasks and their linguistic inventories, perhaps due to a lack of parallel annotations for WSD and SRL. Instead, thanks to MOSAICo, we show that such an interconnection can open the door to interesting work. More specifically, we leverage the fact that VerbAtlas "senses" (more correctly, frames) are clusters of BabelNet synsets. Therefore, we convert each verbal synset predicted by ESCHER to a VerbAtlas frame and compare it to the frame predicted by Multi-SRL. Interestingly, although both systems report state-of-the-art results, the agreement is only 74%. A manual inspection reveals several classes of disagreement. The two simplest cases are when one of the two systems is wrong, i.e., there is either a *WSD error* or an *SRL error*. However, *part-of-speech errors* – which make the WSD and SRL predictions inevitably wrong – appear a non-negligible number of times, even though part-of-speech tagging is often considered a "solved" task, especially in English. Perhaps, most importantly, we observe *inventory errors*, i.e., predictions that are wrong by default because the inventory (e.g., BabelNet, VerbAtlas, or PropBank) does not include the correct answer among its possible options. It is clear that future work should aim at bridging not only the approaches, but also the inventories used in WSD and SRL.

Comparing structured data: SRL and SP. Since our SRL and SP annotations employ the same inventory – PropBank – we can leverage MOSAICo Core to investigate the relation between two different types of structured prediction, namely, the predicate-argument structures in SRL and the AMR subgraphs in SP. However, because AMR abstracts away from the sentence level, there is no direct relation with the words in a sentence, unlike SRL, which works at the span level. Therefore, in order to check the SRL-AMR overlap, we use the cross-lingual AMR sentence-graph aligner of Martínez Lorenzo et al. (2023b) to connect each AMR node to the respective sentence span. In the AMR graphs of MOSAICo Core, there are around 10M nodes tagged with a PropBank sense label. If

we compare the predicate lemmas found in these AMR graphs with those found in the SRL annotations of the corresponding sentences, we can observe a large overlap of around 5.6M predicate lemmas. It is interesting to note that SPRING and Multi-SRL agree on the sense labels of these predicates around 95.5% of the time. Among the rest of the predicates, we can attribute 3.2% of the disagreements to an annotation error by one of the two systems (or both). Instead, only 1.3% of the disagreements on the predicate sense can be attributed to differences in how SRL and AMR parses are constructed. Finally, we analyze the overlap between the semantic roles that appear in the AMR graphs and the SRL annotations, finding that 92.7% of these triplets are annotated with the same semantic roles by both SPRING and Multi-SRL. Given the strong ties between the SRL and SP annotations in MOSAICo, future work may explore leveraging their connection to produce cleaner annotations.

RE meets WSD. The combination of RE and WSD presents a unique opportunity for enriching relations. As outlined in Section 3.2, we produce a large portion of our RE annotations with the cRocoDile pipeline using the links to Wikidata in Wikipedia articles. However, by leveraging the mapping between BabelNet and Wikidata, we also enrich the RE corpora by combining the manual annotations provided by the Wikipedia community with the WSD annotations in MOSAICo. Compared to using only the Wikipedia links, using the WSD annotations results in the addition of more than 5 million triplets to MOSAICo Core, as highlighted in Table 8. Moreover, MOSAICo Core contains more commonsense annotated relational facts, i.e., triplets in which the subject and object are concepts, than those usually provided by a RE system, providing new opportunities for future research. Finally, the fact that nearly half of the relation triplets in MOSAICo Core originate from WSD validates our hunch on the presence and value of semantic relations, not only between entities but also between concepts (Martinelli et al., 2024).

New Wikidata annotations. MOSAICo is a source of new relational data and textual evidence. From the predictions portion of the annotation (see Table 8), there are triplets not yet present in Wikidata, which is known to be incomplete (Conia et al., 2023). For instance, there are 141,128 unique relational facts in the English version of MOSAICo Core that were not present in Wikidata at the time of the RE annotations (September 9, 2022). However, compared to a more recent Wikidata dump (December 30, 2022), that number is reduced to 140,568. This means 560 newly added relational facts were already present in MOSAICo. For example, the fact that Leonhard Euler (Q7604) had worked in the field (P101) of Graph Theory (Q131476) was added to Wikidata on November 7, 2022, and is annotated five times in MOSAICo Core despite not being present at annotation time.

		Predictions	cRocoDiLe	WSD	Total
EN	**Triplets**	533,341	878,958	1,563,412	3,631,011
	Filtered	60,382	220,845	597,731	878,958
FR	**Triplets**	449,403	372,355	570,145	1,391,903
	Filtered	39,160	105,128	175,765	320,053
ES	**Triplets**	224,037	187,454	290,794	702,285
	Filtered	28,758	98,017	158,707	285,482
IT	**Triplets**	227,720	190,166	283,916	701,802
	Filtered	30,225	115,185	119,711	265,121
DE	**Triplets**	403,184	306,169	537,814	1,247,167
	Filtered	44,313	138,624	246,506	429,443

Table 8: Number of triplets annotated in MOSAICo-Core. Triplets come from three sources: predicting new triplets using a RE system (predictions), interlinking Wikipedia text and Wikidata (cRocoDiLe), and leveraging WSD annotations (WSD). We also report the number of triplets after Triplet Critic filtering (Filtered).

6 Conclusion and Future Work

In this paper, we presented MOSAICo, the first effort in the field of NLU aimed at equipping researchers with the key ingredients for modeling symbolic semantic knowledge at a large scale. MOSAICo provides hundreds of millions of annotations in four NLU tasks across five languages, filling a crucial gap in the availability of semantically-tagged corpora. After describing the creation process of MOSAICo, we analyzed the interplay between different sources of semantic information and provided insights into how they can be combined to bring different types of explicit knowledge together. Finally, we demonstrated the quality of the annotations in MOSAICo, which can improve the performance of state-of-the-art systems across standard benchmarks, showing its value for the research community. We release MOSAICo to the community, allowing future work to drop the requirement of licensed datasets and making research in this area accessible to a wider community.

Limitations

Here we discuss some of the limitations of our work, hoping not only to address them but also to encourage future work on top of ours.

Manual evaluation. Although some proxy results do confirm the quality of MOSAICo – for example, the performance of the systems trained on MOSAICo and tested on standard benchmarks (Section 4) – we did not evaluate the annotations directly. Performing such a manual effort would have been too expensive, especially when we consider that MOSAICo's annotations span across 5 languages and 4 different tasks. However, after conducting a qualitative analysis of the annotations, we found them to exhibit promising quality. This is evident, for example, when comparing them to the user annotations in Wikipedia, which confirms their effectiveness in WSD, as shown in Table 7. Overall, we believe that a human evaluation would be a valuable contribution and bring deeper insights into problems and challenges of each task. However, such a wide-range annotation process would require annotators trained on several tasks and languages, and a large budget. We would also need to include redundant annotations in order to mitigate issues like inter-annotator agreement, which is notoriously hard for these tasks. For instance, most SemEval tasks in WSD concern the manual annotation and validation of only 1000-2000 items, still involving considerable work in one or a few languages for a single task. We think that creating such a gold standard would merit a paper in itself.

Missing languages and tasks. While we believe that the multilingual nature of MOSAICo already provides a valuable foundation for cross-lingual studies, we recognize the desire to expand the language coverage of MOSAICo to include a broader range of languages. Furthermore, we strongly believe that the more semantic tasks we add to MOSAICo, the greater will be the opportunity to study possible intersections between them. For this reason, we designed MOSAICo as an ever-growing resource and will continue to update it by adding new tasks and languages.

Wikipedia. Despite the numerous advantages of utilizing Wikipedia as the underlying corpus for our resource, it is important to acknowledge certain limitations inherent to this choice. Two notable drawbacks are the topics discussed and the descriptive/encyclopedic writing style found in Wikipedia articles, which may introduce inherent biases and thus impact the generalizability of our resource to other domains or genres (Navigli et al., 2023). However, we would also like to note that the benchmarks that we use to evaluate the systems include out-of-domain sets as well. The results on these test sets suggest that training a system on MOSAICo's annotations actually brings improvements in the generalization capabilities of the system.

Another limitation is the relatively low coverage of low-resource languages within Wikipedia. Although Wikipedia contains vast amounts of information in multiple languages, there is an inherent bias towards languages with larger speaker populations, or those with more active contributors. Consequently, the availability and depth of annotations for low-resource languages may be limited.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487, the PNRR MUR project PE0000013-FAIR, the CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content), and the Marie Skłodowska-Curie project *Knowledge Graphs at Scale* (KnowGraphs) No. 860801 under the European Union's Horizon 2020 research and innovation programme.

Simone Conia and Edoardo Barba are fully funded by the PNRR MUR project PE0000013-FAIR. While working at Babelscape, Abelardo Carlos Martínez Lorenzo and Pere-Lluís Huguet Cabot were funded by KnowGraphs. Babelscape also funded the dataset translation work. CREATIVE is funded by the MUR Progetti di Ricerca di Rilevante Interesse Nazionale programme (PRIN 2020). We acknowledge the CINECA awards IsCa5_WSP and IsCa5_WRE under the ISCRA initiative for the availability of high-performance computing resources and support.

References

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

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. DMLM: Descriptive Masked Language Modeling. 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, and Angelina McMillan-Major et al. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

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.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022a. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

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.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proc. of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*.

Simone Conia, Min Li, Daniel Lee, Umar Minhas, Ihab Ilyas, and Yunyao Li. 2023. Increasing coverage and precision of textual information in multilingual knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1612–1634, Singapore. Association for 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*, pages 1396–1410. International Committee on Computational Linguistics.

Artur d'Avila Garcez and Luis C. Lamb. 2020. Neurosymbolic AI: The 3rd wave. *ArXiv preprint*, abs/2012.05876.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *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 627–637, Hong Kong, China. Association for Computational Linguistics.

Bofei Gao, Liang Chen, Peiyi Wang, Zhifang Sui, and Baobao Chang. 2023. Guiding AMR parsing with reverse graph linearization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13–26, Singapore. Association for Computational Linguistics.

Jonas Geiping and Tom Goldstein. 2022. Cramming: Training a language model on a single GPU in one day. In *ArXiv preprint*, volume abs/2212.14034.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5).

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

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, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

Chao Lou and Kewei Tu. 2023. AMR parsing with causal hierarchical attention and pointers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8942–8955, Singapore. Association for Computational Linguistics.

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

Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024. CNER: Concept and Named Entity Recognition. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City.

Abelardo Carlos Martínez Lorenzo, Pere Lluís Huguet Cabot, and Roberto Navigli. 2023a. AMRs assemble! learning to ensemble with autoregressive models for AMR parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1595–1605, Toronto, Canada. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Pere Lluís Huguet Cabot, and Roberto Navigli. 2023b. Cross-lingual AMR aligner: Paying attention to cross-attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1726–1742, Toronto, Canada. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In *Proc. of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*.

Abelardo Carlos Martínez Lorenzo and Roberto Navigli. 2024. Efficient amr parsing with CLAP: Compact linearization with an adaptable parser. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

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*.

George A. Miller, Martin Chodorow, and Shari et al. Landes. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

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 *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5697–5702. International Joint Conferences on Artificial Intelligence Organization.

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.

Roberto Navigli, Rexhina Blloshmi, and Abelardo Carlos Martinez Lorenzo. 2022. BabelNet Meaning Representation: A Fully Semantic Formalism to Overcome Language Barriers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory and discussion. *J. Data and Information Quality*. Just Accepted.

Riccardo Orlando, Simone Conia, and Roberto Navigli. 2023. Exploring non-verbal predicates in semantic role labeling: Challenges and opportunities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12378–12388, Toronto, Canada. Association for Computational Linguistics.

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.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for Word Sense Disambiguation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*. AAAI Press.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wrightbettner, and Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Luc De Raedt, Sebastijan Dumancic, Robin Manhaeve, and Giuseppe Marra. 2020. From statistical relational to neuro-symbolic artificial intelligence. In *Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org.

Colin Raffel, Noam Shazeer, and Adam Roberts et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proc. of the 15th Conference of the EACL: Volume 1, Long Papers*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proc. of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

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*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajic, Daniel Hershcovich, Eduard H Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. Incorporating graph information in transformer-based AMR parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, and Nikita Nangia et al. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*.

Alex Wang, Amanpreet Singh, and Julian Michael et al. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

A MOSAICo: Construction and Annotation

Here we provide more details on the systems that we used to create MOSAICo. More specifically, for each task, we have selected a system among the ones that are considered to be state-of-the-art, trained it from scratch, and used it to annotate Wikipedia. In the following, we provide the training details of the systems used for WSD, SRL, SP, and RE.

Word Sense Disambiguation. We used ESCHER (Barba et al., 2021) to tag MOSAICo with word senses since it is a state-of-the-art system that takes advantage of word sense definitions to better generalize on unseen patterns and inventories. We train every model using AdamW with 5000 warmup steps and for a total of 100 000 steps with a learning rate of 10^{-5} and a batch size of 4096 tokens. For English we follow the training splits described in Raganato et al. (2017), for all the other languages we follow the splits described in the XL-WSD framework.

Semantic Role Labeling. We used Multi-SRL (Conia and Navigli, 2020) to tag MOSAICo with predicate-structure inventories. Multi-SRL is a state-of-the-art system for SRL with strong performance in multilingual settings. We train every model using AdamW with a peak learning rate of 10^{-5}, 10 000 warmup steps, and a linear decay to 10^{-6} for the learning rate for 20 epochs. In every configuration, we train a model on English data from OntoNotes 5 and rely on the cross-lingual transfer capabilities of its underlying language model, i.e., XLM-RoBERTa-base to tag SRL annotations in non-English languages.

Semantic Parsing. We used LeakDistill (Vasylenko et al., 2023) to tag MOSAICo with predicate-structure inventories in English. LeakDistill is a state-of-the-art system for English AMR parsing. For the multilingual settings, we extend the efficient implementation of SPRING, CLAP (Martínez Lorenzo and Navigli, 2024), following Blloshmi et al. (2020). We train every model using Adafactor with a peak learning rate of 10^{-4}, 500 warmup steps, and a batch size of 2048 tokens.

Relation Extraction. We used mREBEL (Huguet Cabot et al., 2023) to tag MOSAICo with triplets as part of the RE annotation process along the Crocodile pipeline. mREBEL is an mBART-based version of REBEL (Huguet Cabot and Navigli, 2021), which reframes the RE task as a seq2seq problem by decoding a linearized version of the triplets. Similarly, to enable comparisons with REBEL on English benchmarks, we train our RE models on top of BART-large, using the same settings as the original paper for both pretraining, (1000 warmup steps) and finetuning on target datasets (10% steps warmup) with 10^{-5} peak learning rate and 32 batch size.

Task	Benchmark	Languages	#Sentences	#Annotations	Training	Dev	Test
WSD	SemCor	EN	33K	226K	33K	–	–
	WGC	EN	116K	497K	116K	–	–
	ALL	EN	1.1K	7.2K	–	–	1.1K
	42D	EN	–	–	–	–	0.4K
	XL-WSD	M	696K	1.5M	1.5M	2K	14.2K
	MOSAICo	M	63.5M	522M	63.5M	–	–
	MOSAICo-Core	M	7.9M	17M	7.9M	–	–
SRL	OntoNotes 5	EN	137K	384K	309K	43K	32K
	PB-Examples	EN	14K	14K	–	–	14K
	X-SRL	M	118K	219K	200K	7K	12K
	MOSAICo	M	63.5M	285.3M	285.3M	–	–
	MOSAICo-Core	M	7.9M	11.8M	11.8M	–	–
SP	AMR 3.0	EN	59,255	59,255	55,635	1,722	1,898
	TLP	EN	1,562	1,562	–	–	1,562
	Bio	EN	6,952	6,952	–	–	6,952
	AMR 4T	M	1,371	1,371	–	–	1,371
	MOSAICo	M	17.6M	17.6M	17.6M	–	–
	MOSAICo-Core	M	2.5M	2.5M	2.5M	–	–
RE	CoNLL04	EN	1.4K	2.1K	1,290	343	422
	NYT	EN	66.2K	111.3K	94,222	8,489	8,616
	ADE	EN	4.3K	6.8K	6,821	–	–
	MOSAICo	M	63.5M	39.3M	39.3M	–	–
	MOSAICo-Core	M	7.9M	7.6M	7.6M	–	–

Table 9: Data Distributions of Benchmarks for each Task. Rows: WSD, SRL, SP, and RE. Columns: Task, Benchmark Name, Languages, Total Number of Sentences, Total Number of Annotations, and Distribution of Training, Dev, and Test Sets annotations. The Language column can be En (English) or M (Multilingual), where, in the multilingual case, the number of annotations is per language.