elexis european lexicographic infrastructure

D3.5

MULTILINGUAL WORD SENSE DISAMBIGUATION AND ENTITY LINKING ALGORITHMS - FINAL REPORT

Authors: Marco Maru, Federico Martelli, Rexhina Blloshmi, Roberto Navigli, Paola Velardi

Date: 31 January 2022

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

Deliverable Number: D3.5

Dissemination Level: Public

Delivery Date: 31 January 2022

Version: Final

Authors: Marco Maru, Federico Martelli, Rexhina Blloshmi, Roberto Navigli, Paola Velardi

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
31 January 2022	-	-

Table of Contents

List of Tables

List of Figures

1 Introduction

This document accompanies and describes the set of tools and data that are released as deliverable D3.5 (Multilingual Word Sense Disambiguation and Entity Linking Algorithms - final report) with respect to task T3.1 (work package WP3: JRA Lexicographic Data for NLP).

The main objective identified in the context of WP3 Task 3.1 is that of **narrowing the gap that currently exists between English and other languages** when it comes to data needed to train and evaluate supervised systems for Word Sense Disambiguation (WSD) and Entity Linking (EL).

1.1 Word Sense Disambiguation and Entity Linking

Word Sense Disambiguation is a longstanding task in Natural Language Processing (NLP), an AI-complete problem dealing with the **automatic resolution of the key components of human language ambiguity, namely, polysemy and homonymy** (Bevilacqua et al., 2021). Traditionally, WSD is performed by means of assigning one or more senses to a target word in context, picking from a finite, predefined machine-readable dictionary (a sense inventory). Closely related to WSD is the task of Entity Linking, which requires systems to **label mentions with named entities found within existing knowledge bases** such as Wikipedia. Due to its inherent properties, a widely employed sense repository such as BabelNet (Navigli and Ponzetto, 2012), which includes both dictionary and encyclopaedic knowledge, enables systems to concurrently perform WSD and EL.

As of today, two main issues keep hampering the dissemination of WSD and EL applications in languages other than English: on the one hand, the **evaluation is usually carried out on stale frameworks** which, with a few outdated and heterogeneous exceptions, **do not account for low-resourced languages** (in fact, official Senseval and SemEval competitions featured test beds solely covering English, French, German, Spanish and Italian). On the other hand, **the paucity of high-quality training corpora** (so-called *knowledge acquisition bottleneck*), strictly related to the demanding time and money requirements to produce them, hinders state-of-the-art systems to unleash their full potential and be of use in a wider variety of languages and domains.

The remainder of this deliverable introduces four solutions to overcome the aforementioned issues. First, in Section 2, we provide details concerning XL-WSD: a new, cross-lingual evaluation benchmark

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

featuring datasets in 18 languages (Pasini et al., 2021). Secondly, in Section 3, we detail the creation of a novel, multilingual, manually-curated WSD dataset that has been initiated with the aim of covering 10 European languages (Martelli et al., 2021). As a third contribution, in Section 4, we describe MultiMirror (Procopio et al., 2021), a neural word alignment architecture for multilingual WSD which is able to perform seamless sense projection from a source to a target language in order to automatically produce high quality training data. Finally, in Section 5, we explore the usage of a state-of-the-art WSD system, namely ESCHER (Barba et al., 2021b), using dictionaries in the ELEXIS matrix as sense inventories and proving their aptness at being used as reference repositories for low-resourced languages in cutting-edge multilingual WSD. This latter goal is key to the success of the work package, in that the more the languages and the resources available and integrated into the ELEXIS matrix, the better the performance in disambiguation tasks.

2 XL-WSD

To address data scarcity, we first put forward **XL-WSD, the first large-scale multilingual evaluation framework for WSD** which (i) employs a unified sense inventory, and (ii) **covers 18 different languages** from 6 linguistic families, namely, Basque, Bulgarian, Catalan, Chinese, Croatian, Danish, Dutch, English, Estonian, French, Galician, German, Hungarian, Italian, Japanese, Korean, Slovenian, and Spanish.

2.1 Sense Inventory

Even though WordNet (Miller, 1995) represents the *de facto* sense inventory for the English language, no standard is equivalently established for other languages. **BabelNet**, on the other hand, being a unified multilingual repository of knowledge and lying at the very core of ELEXIS, inherently provides coverage for concepts in multiple languages (284, in its 4.0 version) thanks to the inclusion of language-specific lexicalizations for each distinct meaning.[1]

[1] At the time of developing XL-WSD, Task 2.3 (cross-lingual mapping through shared conceptualisation), whose aim is to link lexical resources across languages, was still undergoing.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

For the purposes of XL-WSD, we derived the subset of BabelNet 4.0 made up of the 117,659 synsets whose English lexicalizations are also featured in WordNet 3.0. Though constraining the number of word senses for other languages to the English WordNet may lead to unsound representations of a language semantics, this strategy allows for a fair evaluation of systems in a cross-lingual setting, and serves to establish a common and coherent testing ground.

2.2 Gold-standard Test Sets

Given that WordNet and WordNet-like repositories commonly employ usage examples to accompany a given word sense and its definition, we exploited this information to create **new evaluation benchmarks**. Particularly, we employed language-specific wordnets in Basque (Pociello et al., 2008), Bulgarian (Simov and Osenova, 2010), Catalan (Benítez et al., 1998), Chinese (Huang et al., 2010), Croatian (Raffaelli et al., 2008), Danish (Pedersen et al., 2009), Dutch (Postma et al., 2016), Estonian (Vider and Orav, 2002), Galician (Guinovart, 2011), Hungarian (Miháltz et al., 2008), Japanese (Isahara et al., 2008), Korean (Yoon et al., 2009), and Slovenian (Fišer, Novak, and Erjavec 2012).

After having retrieved our sense repositories, for each synset *s* in a given wordnet, and one of its usage examples *e* made up of the words $w_1 \ldots w_n$, we picked as target word the word whose PoS tag matches that of *s* and whose lemma is featured in the lexicalizations of *s*. Subsequently, exploiting the mapping between the language-specific wordnet to the English WordNet, and hence to BabelNet, we labeled the target word in *e* with its appropriate BabelNet synset.

In addition, we took into account the multilingual gold standards made available as part of past SemEval competitions i.e., the Italian and Chinese datasets in SemEval-10 Task 17 (Agirre et al., 2010), French, German, Italian and Spanish datasets in SemEval-13 Task 12 (Navigli et al., 2013), and Italian and Spanish datasets in SemEval-15 Task 13 (Moro and Navigli 2015).

With respect to the English language instead, we included in XL-WSD the datasets originally featured in the unified evaluation benchmark of Raganato et al. (2017), with the addition of data from SemEval-10 Task 17 and the coarse grained test set taken from SemEval-07 Task 7 (Navigli et al., 2007).

Finally, we gathered our datasets sharing the same language and performed a random split over their instances so as to obtain two subsets: one for testing (80% of the overall data), and one for

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

development (data not featured in the test sets). For English instead, we set aside SemEval-07 as a development set (English-Dev) and provided two evaluation grounds: **English-Fine**, made up of the test sets of Senseval-2, Senseval-3, SemEval-10, SemEval-13 and SemEval-15, and **English-Coarse**, featuring SemEval-07 Task 17.

2.3 Silver-standard Training Sets

Enabling WSD in a multilingual scenario entails the provision of adequate test beds, as well as the creation of datasets to train systems with. So far, the most widely employed training sets for WSD, already featured in the unified evaluation framework of Raganato et al. (2017), i.e., **SemCor** (SC) and the **Princeton WordNet Gloss Corpus** (WNG), only cover the English language. To deal with the paucity of training sets in other languages, **we created silver training data by exploiting Opus-MT**, the machine translation models of Tiedemann and Thottingal (2020). In particular, thanks to the models in Opus-MT, we were able to translate both SC and WNG in 15 out of 17 non-English languages featured in XL-WSD (exception made for Chinese and Korean) and hence to produce silver training corpora (T-SC+WNG) by means of a simple unsupervised sense projection technique which concurrently takes into account PoS tags and synset lexicalizations information to correctly identify the target word to label in the translated corpus.

Table 1 shows the general statistics for all corpora in XL-WSD, including the average ambiguity level for words in each corpora (Word-Type Polysemy), computed as the total number of candidate synsets for each word type divided by the total number of word types.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Language	Word Types			Polysemous Words			Word-Type Polysemy			Instances			Unique Synsets		
	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev
English-Fine	106906	2882	330	24658	2199	308	1.458	3.689	6.209	840471	8062	455	117653	3469	361
English-Coarse	-	980	-	-	750	-	-	4.255	-	-	1816	-	-	2190	-
Basque	12503	771	304	5294	525	253	2.331	3.224	4.543	197309	1580	395	16604	1423	388
Bulgarian	12413	2450	1413	2412	1325	839	1.304	1.670	1.938	148479	9968	2493	12600	2658	1517
Catalan	18603	1276	428	8378	1107	384	2.291	3.940	4.981	331757	1947	487	25624	1767	479
Chinese	-	1786	1173	-	1402	955	-	2.638	3.045	-	9568	2392	-	2687	1524
Croatian	6882	4389	1416	1161	1652	675	1.268	1.244	1.758	94575	6333	1584	6739	4543	1449
Danish	15822	2623	816	3324	1318	428	1.338	1.722	1.950	234681	3502	876	16707	2693	817
Dutch	28351	2935	985	9121	2122	766	1.711	2.356	3.067	305692	4400	1100	30490	2716	950
Estonian	10460	1615	460	1768	917	281	1.246	1.815	2.091	132240	1999	500	10462	1852	490
French	17850	549	203	5978	339	130	1.585	2.413	2.744	252756	1160	289	21510	584	213
Galician	8390	1244	486	3799	773	349	2.079	2.219	2.852	247379	2561	641	11821	1474	548
German	16213	421	154	2332	166	64	1.203	1.639	1.864	184952	862	214	16437	417	155
Hungarian	13234	3491	1022	2908	1931	625	1.367	1.842	2.346	161119	4428	1107	13297	4285	1103
Italian	23773	985	385	9540	758	316	2.021	3.790	4.569	385248	2278	561	29869	1212	475
Japanese	1008	4338	1538	581	2390	1001	2.516	1.871	2.460	23217	7602	1901	1141	5964	1755
Korean	-	1886	740	-	920	408	-	1.373	1.815	-	3796	950	-	1452	683
Slovenian	7577	104	87	1296	93	81	1.245	3.519	3.954	128395	2032	509	7705	243	172
Spanish	22020	847	329	11784	696	270	2.811	4.955	5.435	393539	1851	452	32151	1103	422

Table 1. Statistics of training, test and development sets in XL-WSD.

2.4 Experimental Setup

To establish our baseline model, we employ a **Transformer-based text encoder** (Vaswani et al. 2017) followed by a 2-layer feedforward network with swish activation function and batch-normalization, and stack on top of it an unbiased softmax linear layer for classification. Each sub-token is represented by summing the outputs of the last four layers of the encoder and each word by averaging its sub-token representations. Finally, a linear transformation is performed and the resulting vectors are fed to a linear layer for classification.

As text encoders, we use XLMR-Base, XLMR-Large (Conneau et al. 2020), BERT-Large, M-BERT (Devlin et al. 2019) and the language-specific versions of BERT (L-BERT) for each language in the Hugging Face library (Wolf et al. 2020).

With respect to training, we employ the language-specific T-SC+WNG translations of SemCor and the Princeton WordNet Gloss Corpus for all non-English languages.

Results are reported in terms of the traditional F1 score.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2.5 Results

In Table 2 we report the performances on XL-WSD by our reference models, as trained on (i) SC+WNG and (ii) T-SC+WNG, respectively.

With respect to the first setting (0-shot) it can be seen how XLMR-Large is able to achieve the best results across the board. Also, it can be noted that supervised systems systematically attain higher performances when trained with English data only (0-shot columns), as compared to language-specific training sets (either L-BERT or Language-Specific XLMR-Large columns), clearly testifying to the **aptness of large multilingual pre-trained language models at dealing with WSD**, as opposed to the hitherto employed knowledge-based approaches.

As regards the language-specific setting instead, pre-trained language-specific BERT models perform in the same ballpark as their multilingual counterparts, owing to model size.

Wrapping up, XL-WSD serves to evidence how there is **still room for improvement in both multilingual and 0-shot WSD**. This is particularly striking when considering **performance drops averaging at 10 points when moving from English to another language**, independent of that language being low- or high-resourced.

Data and code for XL-WSD are freely available for research purposes at https://sapienzanlp.github.io/xl-wsd/.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

	∅-Shot (SC+WNG)			Language-Specific (T-SC+WNG)	
Dataset	XLMR-Large	XLMR-Base	M-BERT	XLMR-Large	L-BERT
English-Fine	76.28	74.50	72.40	76.28	**76.77**
English-Coarse	91.30	91.02	89.70	91.30	**91.57**
Basque	**47.15**	43.80	42.41	41.96	43.04
Bulgarian	**72.00**	71.59	68.78	58.18	57.85
Catalan	**49.97**	47.77	47.35	36.00	36.98
Chinese	**51.62**	49.77	48.99	-	-
Croatian	**72.29**	72.13	70.65	63.15	62.89
Danish	**80.61**	79.18	76.04	78.67	76.41
Dutch	**59.20**	58.77	56.64	57.27	56.64
Estonian	**66.13**	64.82	64.33	50.78	51.23
French	**83.88**	82.33	81.64	71.38	71.12
Galician	66.28	64.79	**68.07**	56.18	56.95
German	83.18	82.13	80.63	73.78	73.78
Hungarian	67.64	**68.38**	65.24	52.60	52.17
Italian	77.66	76.73	76.16	**77.70**	75.68
Japanese	**61.87**	61.46	60.34	50.55	50.16
Korean	**64.20**	63.65	63.37	-	-
Slovenian	**68.36**	66.34	62.16	51.13	49.66
Spanish	75.85	76.55	74.66	**77.26**	74.88
Micro AVG	**65.66**	64.82	62.84	-	-

Table 2. Comparison of supervised and language-specific models.

3 The ELEXIS Parallel Sense-Annotated Dataset

With XL-WSD, we provided a much needed benchmark featuring gold test sets in multiple languages, as well as silver training sets translating the most commonly employed datasets for English. Yet, though extremely useful, silver data can inherently include noisy data, hence suboptimal data. For this reason, in order to also provide gold standard training sets in multiple languages, the **creation of a novel, multilingual, manually-curated dataset**, featuring five annotation layers, (i.e., tokenization, sub-tokenization, lemmatization, PoS tagging and Word Sense Disambiguation) has been initiated with the aim of covering 10 European languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish (Martelli et al., 2021). This work is particularly impactful in that, instead of seeking automatic alternatives to generate datasets, it exploits manual curation in order to devise gold standards in low-resourced languages.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3.1 Automatic Extraction and Manual Validation

The annotation of the ELEXIS parallel sense-annotated dataset has been initially carried out following an automatic process of extraction of a set of parallel sentences from WikiMatrix (Schwenk et al., 2021). In particular, language combinations having English as the first language and the other nine languages listed in Section 3 have first been extracted. Subsequently, the 2,500 sentences having the highest overlap across our language combinations have been selected.

After the completion of the automatic sentence extraction, annotators validated the sentences by means of, e.g., removing incorrect punctuation or notes in square brackets. Additionally, in order to ensure a more challenging environment, sentences not featuring at least 5 words (of which 2 polysemous) have been discarded. Given that some translations in some language pairs were missing, annotators manually provided the missing sentences. As a result, **the final dataset comprises an overall figure of 2,024 parallel sentences**.

The annotation of the dataset implies two distinct stages: the first one involves the morpho-syntactic sentence labeling, comprising tokenization, sub-tokenization, lemmatization, and PoS tagging (Section 3.2), whereas the second is related to the actual disambiguation of the instances in the dataset (Section 3.3).

3.2 Annotation: Morpho-syntactic Layers

With respect to the first step, annotators employed a user-friendly interface called Pipeline Annotation and made available by Babelscape[2] (see Figure 1). As a general guideline, annotators were instructed to follow the Universal Dependencies (UD)[3] to ensure data consistency. Given the overall scarcity of datasets to perform EL, named entities were labeled in addition to concepts, so as to create a resource to be used for both tasks that are part of this work package. Table 3 reports the number of tokens, unique lemmas and the open-class part-of-speech distribution for each of the target languages.

[2] https://babelscape.org

[3] https://universaldependencies.org

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

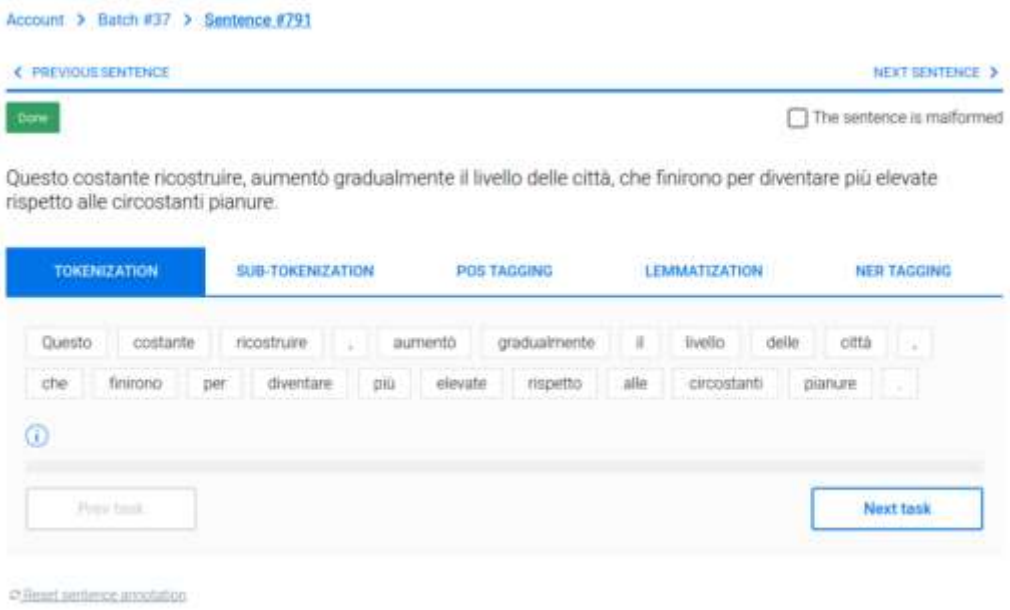

Figure 1. The interface used for morpho-syntactic annotation.

Language	Tokens	Unique Lemmas	Nouns	Verbs	Adjs	Advs
Bulgarian	33,994	6,683	7,892	3,970	3,313	1,157
Danish	32,524	6,832	7,322	3,099	2,626	1,677
Dutch	34,923	6,488	7,142	3,004	2,833	1,020
English	34,228	6,297	6,716	2,946	2,818	1,079
Estonian	37,693	6,112	8,189	3,327	2,310	1,487
Hungarian	29,657	7,457	6,930	2,485	3,561	1,173
Italian	39,067	6,371	7,864	3,022	2,961	1,368
Portuguese	38,723	6,260	7,372	3,181	2,757	1,302
Slovene	31,237	6,688	7,550	2,579	3,820	1,077
Spanish	37,693	6,112	8,189	2,806	3,141	1,140

Table 3. Statistics for the morpho-syntactic annotation phase.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3.3 Annotation: Semantic Layer

As of now, **the morpho-syntactic annotation has been completed**, and manual sense disambiguation is being carried out employing language-specific sense inventories (see Table 4) that will be pivoted on the BabelNet semantic network to allow for cross-lingual evaluation and usage.

Language	Resource
Bulgarian	Dictionary of Modern Bulgarian
Danish	DanNet (The Danish WordNet)
Dutch	Open Dutch WordNet
English	English WordNet
Estonian	EKI Combined Dictionary
Hungarian	The Explanatory Dictionary of the Hungarian Language
Italian	PAROLE-SIMPLE-CLIPS + ItalWordNet
Portuguese	Dictionary of the Lisbon Academy of Sciences
Slovene	sloWNet
Spanish	Spanish Wiktionary

Table 4. Sense inventories for the ELEXIS parallel dataset.

The semantic annotation is performed via the LexTag interface of Babelscape (see Figure 2), which has several features that, *inter alia*, make it perfect for the purposes of this task, namely:

1) it allows annotators to easily switch between BabelNet and the language specific sense inventory to perform the labeling task;
2) it offers sense editing, in order to ameliorate already existing inventories;
3) it allows multiword expressions tagging, to catch, e.g., idioms and other language-specific phenomena.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

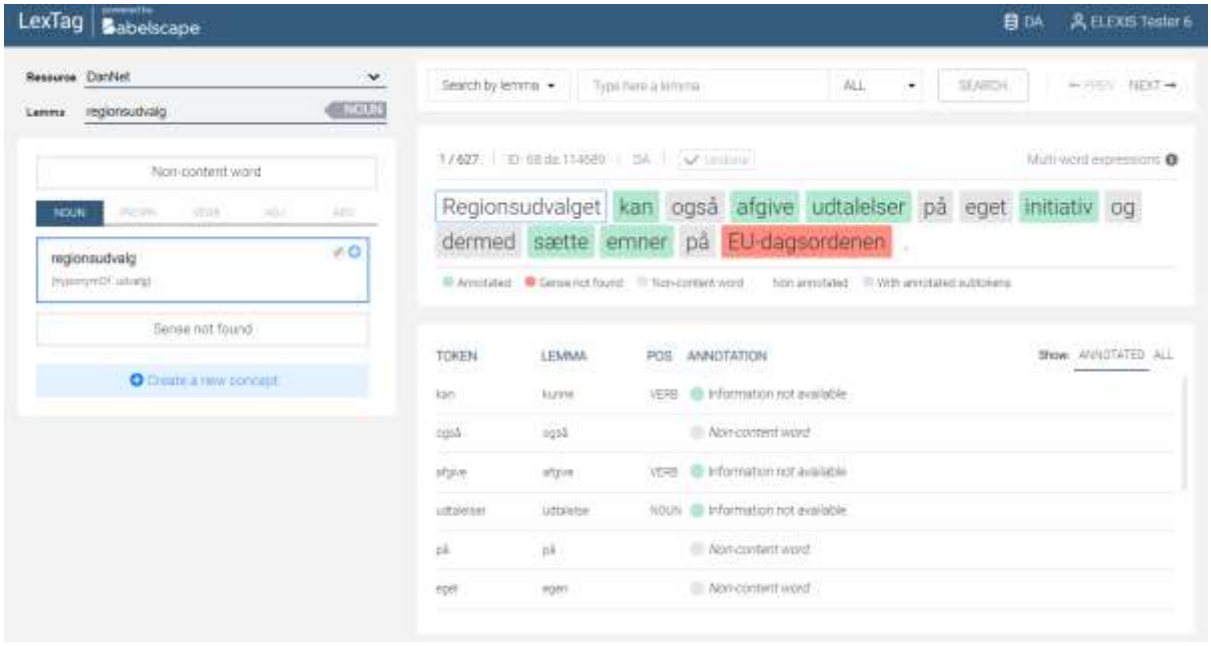

Figure 2. The LexTag annotation interface.

4 MultiMirror

The provision of a huge number of gold test sets for low-resourced languages seen in XL-WSD, as well as the creation of a fully manual resource described in Section 3, are milestones on their own. And yet, enabling disambiguation in the widest possible spectrum of languages also entails innovative strategies to build datasets with which systems can be effectively trained, without resorting to costly and time-consuming infrastructures. To this end, we worked on, and created MultiMirror, a **sense projection approach for multilingual WSD**. It is based on a neural discriminative model for word alignment which is trained with a very low number of instances and which, given a pair of parallel sentences as input, **can effectively align, with state-of-the-art quality, source and target tokens across different language combinations.**

4.1 Cross-lingual Word Alignment Model

We propose a discriminative word alignment model, shown in Figure 3, that takes two parallel sentences as input. To obtain continuous representations of each token, multilingual BERT is

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

employed (Devlin et al., 2019, mBERT), using its final layer to compute a representation for each token by averaging the vectors associated with the subwords that token was split into. Additionally, to enable token-level contextualization, a 6-layer Transformer Encoder resembling mBERT is used, and each possible alignment is classified separately.

Our model is then trained by minimizing the Binary Cross Entropy loss between the word alignment matrix and the reference matrix containing the gold annotations.

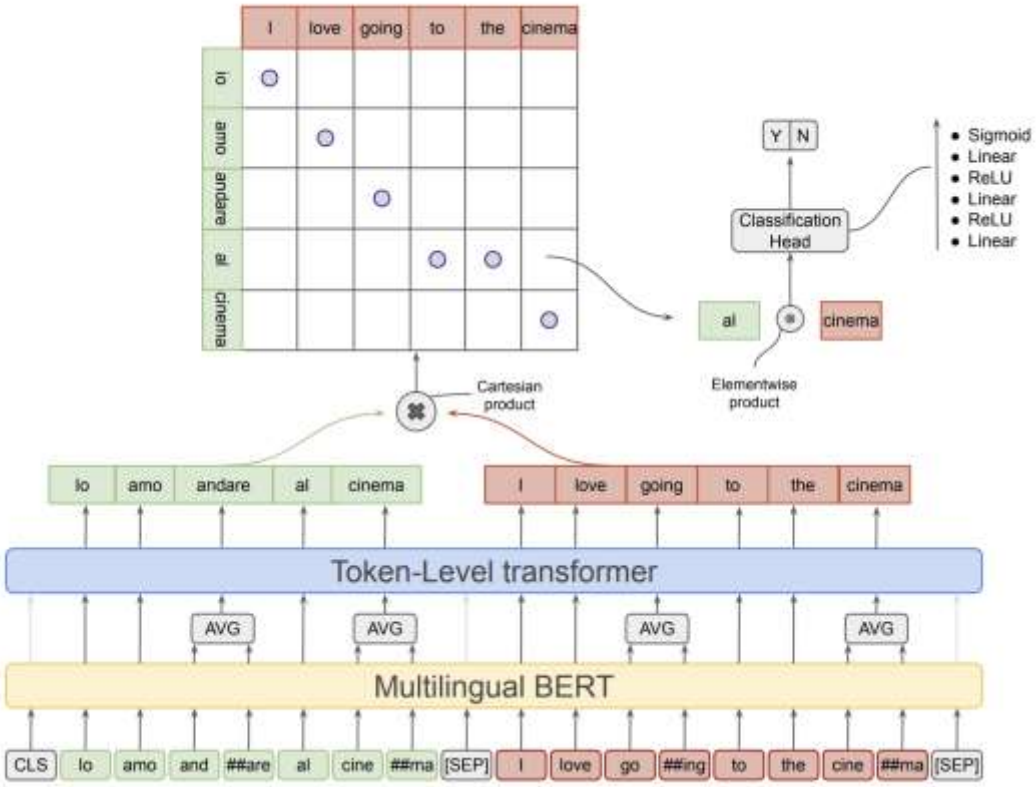

Figure 3. The MultiMirror word alignment model.

To show how MultiMirror fares in terms of alignment, we devised a simple experiment. Particularly, we selected the datasets described in Nagata et al. (2020), i.e., 4 manually curated datasets for word alignment in the following language combinations: English-French (En-Fr), German-English (De-En), Japanese-English (Ja-En) and Romanian-English (Ro-En), and compared our results in terms of F1 score against the hitherto best performing system of Nagata et al. (2020).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Results, shown in Table 5, evidence how the **MultiMirror word aligner performs better in all settings**, interestingly, **with a recall value that is systematically higher than that of its competitor**.

Languages	Method	P	R	F1
Ja-En	Nagata *et al.* [2020]	77.3	78.0	77.6
	MultiMirror Word Aligner	78.3	80.5	**79.4**
De-En	Nagata *et al.* [2020]	89.9	81.7	85.6
	MultiMirror Word Aligner	90.1	83.6	**86.7**
En-Fr	Nagata *et al.* [2020]	79.6	93.9	86.2
	MultiMirror Word Aligner	81.5	92.7	**86.8**
Ro-En	Nagata *et al.* [2020]	90.4	85.3	87.8
	MultiMirror Word Aligner	90.6	88.5	**89.1**

Table 5. Word alignment performance of MultiMirror in terms of precision, recall and F1 on different language pairs.

4.2 Sense Projection

Importantly, thanks to performing word alignment, **MultiMirror can be used to project word senses across different languages**. Given a corpus of sense-tagged sentences in a source language as input, and their translations in a target language, MultiMirror can label the translated sentences according to the word senses in the original corpus.

To this end, the alignment matrix obtained from our cross-lingual word alignment model is employed along with a technique to prevent target spans of the sentences to overlap with other targets, while, at the same time, preserving the PoS tags when moving from the source to the target language.

4.3 Experimental Setup

To assess MultiMirror in the WSD setting, we employed the alignment datasets already described in Section 4.1, with the addition of **datasets manually crafted for Italian and Spanish** by aligning for each of them 300 sentence pairs (for around 4,000 tokens) taken from WikiMatrix, of which 50 are reserved for development.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

With respect to our reference source corpus, we used the concatenation of SemCor and the Princeton WordNet Gloss Corpus and, since official translations for this data is not available, we used the multilingual translation model of Tang et al. (2020) to generate target corpora in five distinct languages, namely French, German, Italian, Japanese, and Spanish. Owing to the same rationale as in XL-WSD, MultiMirror too makes use of BabelNet as its sense inventory, due to its capability of representing abstract concepts in a multilingual context.

In Table 6 we report statistics for the datasets generated by means of MultiMirror, as compared to its strongest competitor, i.e., MuLaN (Barba et al., 2021a), noting how **MultiMirror is able to systematically transfer a larger number of instances, senses and synsets from the source to the target corpus**.

		IT	ES	FR	DE	JA
MULTIMIRROR	# instances	519k	552k	387k	318k	301k
	# senses	77k	92k	62k	68k	98k
	# synsets	37k	50k	29k	22k	25k
	# multiwords	28k	38k	19k	19k	40k
MuLaN	# instances	415k	452k	310k	245k	310k
	# senses	44k	57k	29k	22k	27k
	# synsets	33k	43k	25k	19k	21k
	# multiwords	18k	22k	20k	6k	552

Table 6. Statistics of training sets from MultiMirror.

We tested our sense-tagged corpora against the SemEval-13 and SemEval-15 test sets, and included the Japanese test bed of XL-WSD to estimate the capability of MultiMirror to deal with distant languages.

We chose to employ a simple linear classifier on top of mBERT as our reference model, exploring two different strategies at training time: (i) fine-tuned (FT), where the whole model is fine-tuned, with mBERT being updated along with the linear classifier, and (ii) feature-based (FB), where, instead, we freeze weights and update only the linear classifier. As comparison systems, we report: (i) the Most

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Common Sense baseline, (ii) UKB+SyntagNet (Maru et al., 2019), a knowledge-based approach exploiting the Personalized PageRank algorithm and a WordNet graph enhanced via syntagmatic relations, (iii) ARES (Scarlini et al., 2020), a method for producing contextualized sense embeddings across languages, and (iv) the annotation projection technique of MuLaN (Barba et al, 2021a).

4.4 Results

Results in the multilingual WSD setting are shown in Table 7. There, we also report scores on the 4L setting, where the concatenation of the 4 datasets automatically generated for the European languages is used as training data.

As immediately evident, **MultiMirror^{FB} is able to achieve the state of the art**, outperforming its direct competitor, i.e., MuLaN, in all settings. The same holds for MultiMirror^{FT}, with even more impressive results. Interestingly, the widest performance gap is attained in Japanese, proving that **MultiMirror is also able to efficiently scale to distant languages**.

Finally, in the 4L setting, improvements are achieved across the board, even surpassing ARES on the Spanish section of the SemEval-15 test set.

			SemEval-13				SemEval-15		XL-WSD
	Model	Alignment Data	IT	ES	FR	DE	IT	ES	JA
	MCS	-	44.20	37.10	53.20	70.20	44.60	39.60	48.71
	UKB+SyntagNet	-	72.14	74.12	70.32	76.39	68.95	63.37	-
	ARES	-	77.00	75.30	81.20	79.60	71.40	70.10	-
1L	MuLaN	-	77.45	77.70	80.12	82.09	70.31	68.73	57.59
1L	MultiMirror^{FB}	Nagata *et al.* [2020]	-	-	81.09	82.16	-	-	58.34
1L	MultiMirror^{FT}	Nagata *et al.* [2020]	-	-	81.78	**83.18**	-	-	**62.60**
1L	MultiMirror^{FB}	Ours	78.59	79.68	80.81	81.13	**73.49**	69.03	-
1L	MultiMirror^{FT}	Ours	**79.53**	**81.83**	**83.44**	82.81	72.89	69.42	-
4L	MuLaN$_{4L}$	-	77.85	81.11	81.64	82.34	71.80	69.42	-
4L	MultiMirror$^{FB}_{4L}$	Best	78.59	81.67	81.64	82.43	73.39	69.42	-
4L	MultiMirror$^{FT}_{4L}$	Best	**79.60**	**82.17**	**83.64**	**83.71**	**73.69**	**70.42**	-

Table 7. WSD performances of MultiMirror and its competitors.

All datasets produced with MultiMirror are available at https://github.com/SapienzaNLP/multimirror.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

5 Extractive Sense Comprehension and the ELEXIS Matrix

We successfully reached the objective of effectively narrowing the gap between English and low-resourced languages in WSD, but the success of this work package lies in the exploitation of the wealth of knowledge encoded in the ELEXIS dictionary matrix. In fact, the use of such an interconnected and high quality set of manually-curated repositories in disambiguation is a crucial objective of the ELEXIS project to enable and enhance the tasks of WSD and EL in a wide array of languages. To this end, given the existence of the ESCHER state-of-the-art WSD system of Sapienza (Barba et al., 2021b), which allows for the use of arbitrary sense repositories to perform WSD, we retrieved a set of dictionaries from the ELEXIS matrix, and assessed their aptness at being successfully employed as machine-readable repositories for the tasks under investigation. In the remainder of this Section, we will first introduce the ESCHER disambiguation system and, subsequently, our experimental setup and results.

5.1 ESCHER

The standard formulation of WSD, i.e., that of a multi-label classification task, can hinder the capability of a model to effectively represent word meanings, with no system able to attend all of the possible definitions of a given word at once. To overcome this, Barba et al. (2021b) recently reframed WSD as a text extraction task deemed Extractive Sense Comprehension (ESC), in which a system is asked to extract the text span associated with the correct definition for a target word, given a pseudo-sentence made up of the concatenation of all of its possible glosses.[4]

On top of ESC, the ESCHER model, i.e., a transformer-based architecture implementing the task, is also introduced.

Particularly, ESCHER takes as input a context where the target word to be disambiguated is explicitly marked by means of special characters, and that is followed by the set of the available definitions for

[4] The ESC paradigm is being **currently reframed to account for the handling of Named Entities**, employing information taken from open data sources such as Wikipedia to specifically perform state-of-the-art Entity Linking (a publication describing this work is planned for submission at a top-tier venue in Q1 2022).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

that word according to a given sense inventory, and predicts a pair of indices representing the span of text in which the correct definition for the word is to be found.

As shown in Table 8, framing WSD as a text extraction task, along with the use of BART (Lewis et al., 2019) as its transformer architecture, allows ESCHER to reach unprecedented performances in WSD.[5] Particularly, ESCHER was originally tested on the five English all-words evaluation test beds included in the unified framework of Raganato et al. (2017), namely, Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015).

	Dev Set	**Test Sets**				**Concatenation of all Datasets**				
Model	SE07	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	ALL
EWISE	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
GLU	68.1	75.5	73.6	71.1	76.2	—	—	—	—	74.1
LMMS	68.1	76.3	75.6	75.1	77.0	—	—	—	—	75.4
SVC	—	—	—	—	—	—	—	—	—	75.6
GlossBERT	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
ARES	71.0	78.0	77.1	78.7	75.0	80.6	68.3	80.5	83.5	77.9
EWISER	71.0	78.9	**78.4**	78.9	79.3	81.7	66.3	81.2	85.8	78.3
BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	83.0	**87.9**	79.0
ESCHER	**76.3**	**81.7**	77.8	**82.2**	**83.2**	**83.9**	**69.3**	**83.8**	86.7	**80.7**

Table 8. WSD performances of ESCHER and its competitors.

Even more interesting is the fact that, owing to its particular formulation, ESCHER can dispose of the need for a fixed sense inventory, and can easily make use of multiple repositories at the same time maintaining top-notch performances. In Table 9, we report results for ESCHER as compared with BEM (which employs a bi-encoder to represent the target word and its sense definitions within the same space), when trained on:

1) SemCor (ESCHER$_{S,}$ BEM$_{S}$);

[5] Reported competitors are: GLU (Hadiwinoto et al., 2019), SVC (Vial et al., 2019), EWISE (Kumar et al., 2019), GlossBERT (Huang et al., 2019); BEM (Blevins and Zettlemoyer, 2020), EWISER (Bevilacqua and Navigli, 2020), LMMS (Loureiro and Jorge, 2019) and ARES (Scarlini et al., 2020).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2) on the online dataset of examples and definitions from oxforddictionaries.com provided by (Chang et al., 2018 - ESCHER_{OT}, BEM_{OT});
3) on their concatenation (ESCHER_{S+OT}, BEM_{S+OT}).

All models were then tested on ALL (Raganato et al., 2017), as well as on the OX_{test}, the dataset from Chang et al., (2018).

Model	ALL	OX_{test}
BEM_S	79.0	61.7
ESCHER_S	80.7	67.9
BEM_{OT}	67.2	84.3
ESCHER_{OT}	70.3	86.3
BEM_{S+OT}	78.8	85.2
ESCHER_{S+OT}	**81.5**	**87.7**

Table 9. ESCHER and BEM scores on different English inventories.

5.2 Experimental Setup and Results

To prove the effectiveness of ELEXIS dictionaries, we retrieved those featuring usage examples in addition to glosses and part-of-speech information to define lemma entries and subsequently proceeded to create quadruples of *<lemma, PoS, definition, usage example>*.[6] We then employed the Stanza Python natural language analysis package (https://stanfordnlp.github.io/stanza/) to perform tokenization, lemmatization and PoS tagging over each usage example so as to identify the position of the target word therein (i.e., the lemma entry to which the usage examples refers to).

As a result, we derived a distinct sense inventory based on each dictionary in the matrix, mapping lemma-PoS pairs to definitions and usage examples with explicit marking of the target token, hence ready to be fed as input to the ESCHER model with minimum intervention (e.g., concatenating all of the definitions for the same lemma).

[6] We discarded entries for which no usage examples could be retrieved.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

For the purposes of this experiment, we selected the three dictionaries featuring the largest figures of unique lemmas, namely, (i) the Slovene Lexical Database (SLD) provided by JSI for the Slovene language, (ii) the Algemeen Nederlands Woordenboek (ANW) provided by INT for the Dutch language, and the combined dictionary of Estonian language (CE) provided by EKI.

The instances in each of the three dictionaries have been randomized and divided into training/development/test subsets following a traditional split of 80%, 10% and 10% of the data, respectively. Table 10 shows the number of instances used for each split, whereas Table 11 reports the results achieved by ESCHER on the test sets derived from the three chosen ELEXIS dictionaries. As is clearly visible, results **testify to the high quality of the data featured in the dictionaries**, with performances ranging from 85.4 to 91.6, hence being in line and above the state-of-the-art results currently attained on the English language using traditional sense repositories and test benchmarks. This demonstrates how manually-curated dictionaries from the ELEXIS matrix can be perfectly integrated in a cutting-edge neural architecture for WSD such as ESCHER, hence (i) **curbing the need for ad hoc machine-readable dictionaries** to be used in order to enable the task, and (ii) **proving to be key tools to produce high-quality disambiguation on raw text in low-resourced scenarios**.

Language	Training	Validation	Testing
ET	37018	4267	4267
NL	137779	10000	10000
SL	113957	10000	10000

Table 10. ESCHER and ELEXIS dictionaries: number of instances.

Language	Validation	Testing
ET	85.7	85.4
NL	92.0	91.6
SL	87.1	86.8

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Table 11. ESCHER and ELEXIS dictionaries: results (F1 score).

The pretrained model, along with code and data for ESCHER, is available at https://github.com/SapienzaNLP/esc.

6 Conclusion

In this deliverable, **we successfully tackled impactful issues undermining the replication of Word Sense Disambiguation in languages other than English**. In particular, we provided tools and strategies to significantly narrow the gap existing between English and other languages in terms of enabling training and evaluation of monolingual and multilingual systems. To sum up our goals:

1) **We produced and released a novel set of 18 gold test beds and 15 silver training sets for as many different languages** owing to the new benchmark of XL-WSD;
2) We are in the process of finalizing a **gold standard dataset of parallel sentences for WSD and EL in 10 European languages, i.e., the ELEXIS parallel sense-annotated dataset;**
3) We introduced MultiMirror, **an effective strategy to propagate gold sense annotations in order to create high-quality training sets** from scratch in, virtually, any language;
4) With MultiMirror, we also **released four new additional datasets for word alignment**, each featuring 300 sentences in English and one of the following languages: French, German, Italian and Spanish;
5) **We demonstrated the aptness of ELEXIS dictionaries to be employed as high-quality sense repositories in the context of cutting-edge multilingual disambiguation**, thanks to their seamless integration in the state-of-the-art neural architecture of ESCHER.

References

Agirre, E., De Lacalle, O. L., Fellbaum, C., Hsieh, S. K., Tesconi, M., Monachini, M., ... & Segers, R. (2010, July). Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In Proceedings of the 5th international workshop on semantic evaluation (pp. 75-80).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Barba, E., Procopio, L., Campolungo, N., Pasini, T., & Navigli, R. (2021, January). Mulan: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (pp. 3837-3844).

Barba, E., Pasini, T., & Navigli, R. (2021, June). 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* (pp. 4661-4672).

Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., & Taulé, M. (1998). Methods and tools for building the Catalan WordNet. arXiv preprint cmp-lg/9806009.

Bevilacqua, M., & Navigli, R. (2020, July). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2854-2864).

Bevilacqua, M., Pasini, T., Raganato, A., & Navigli, R. (2021, August). Recent trends in word sense disambiguation: A survey. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conference on Artificial Intelligence, Inc.

Blevins, T., & Zettlemoyer, L. (2020, July). Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1006-1017).

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edmonds, P., & Cotton, S. (2001, July). Senseval-2: overview. In Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems (pp. 1-5).

Fišer, D., Novak, J., & Erjavec, T. (2012). sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)* (pp. 113-117).

Hadiwinoto, C., Ng, H. T., & Gan, W. C. (2019, November). Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. 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) (pp. 5297-5306).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Huang, C. R., Hsieh, S. K., Hong, J. F., Chen, Y. Z., Su, I. L., Chen, Y. X., & Huang, S. W. (2010). Chinese Wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, *24*(2), 14-23.

Guinovart, X. G. (2011). Galnet: WordNet 3.0 do galego. *Linguamática*, *3*(1), 61-67.

Huang, L., Sun, C., Qiu, X., & Huang, X. J. (2019, November). GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. 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) (pp. 3509-3514).

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008, May). Development of the Japanese WordNet. In *LREC*.

Kumar, S., Jat, S., Saxena, K., & Talukdar, P. (2019, July). Zero-shot word sense disambiguation using sense definition embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5670-5681).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Loureiro, D., & Jorge, A. (2019, July). Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5682-5691).

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J., Lipp, V., Váradi, T., Győrffy, A., László, S., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., & Munda, T. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In Proceedings of eLex 2021.

Maru, M., Scozzafava, F., Martelli, F., & Navigli, R. (2019, November). SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3534-3540).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., & Váradi, T. (2007). Methods and results of the Hungarian WordNet project. In *Proceedings of GWC* (Vol. 2008, p. 4th).

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Moro, A., & Navigli, R. (2015, June). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 288-297).

Nagata, M., Chousa, K., & Nishino, M. (2020, November). A Supervised Word Alignment Method Based on Cross-Language Span Prediction Using Multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 555-565).

Navigli, R., Litkowski, K. C., & Hargraves, O. (2007, June). Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 30-35).

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, 193, 217-250.

Navigli, R., Jurgens, D., & Vannella, D. (2013, June). Semeval-2013 task 12: Multilingual word sense disambiguation. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (pp. 222-231).

Pasini, T., Raganato, A., & Navigli, R. (2021, May). XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language resources and evaluation*, *43*(3), 269-299.

Pociello, E., Gurrutxaga, A., Agirre, E., Aldezabal, I., & Rigau, G. (2008, May). WNTERM: Enriching the MCR with a Terminological Dictionary. In *LREC*.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A., & Vossen, P. (2016). Open dutch wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)* (pp. 302-310).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007, June). Semeval-2007 task-17: English lexical sample, srl and all words. In Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007) (pp. 87-92).

Procopio, L., Barba, E., Martelli, F., & Navigli, R. (2021). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 (pp. 3915-3921).

Raffaelli, I., Tadic, M., Bekavac, B., & Agic, Ž. (2008). Building croatian wordnet. In *Proceedings of GWC* (pp. 349-360).

Raganato, A., Camacho-Collados, J., & Navigli, R. (2017, April). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 99-110).

Scarlini, B., Pasini, T., & Navigli, R. (2020, November). With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3528-3539).

Schwenk, H., Chaudhary, V., Sun, S., Hopkins, J., Gong, H., & Guzmán, F. (2021) WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia.

Simov, K. I., & Osenova, P. (2010, May). Constructing of an Ontology-based Lexicon for Bulgarian. In *LREC*.

Snyder, B., & Palmer, M. (2004, July). The English all-words task. In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (pp. 41-43).

Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., ... & Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Tiedemann, J., & Thottingal, S. (2020, November). OPUS-MT–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 479-480).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Vial, L., Lecouteux, B., & Schwab, D. (2019, July). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In Proceedings of the 10th Global Wordnet Conference (pp. 108-117).

Vider, K., & Orav, H. (2011, December). Estonian Wordnet and lexicography. In *Symposium on Lexicography XI* (pp. 549-558). Max Niemeyer Verlag.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45).

Yoon, A. S., Hwang, S. H., Lee, E. R., & Kwon, H. C. (2009). Construction of Korean WordNet. *Journal of KIISE: Software and Applications*, *36*(1), 92-108.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.