LexicoMatic: Automatic Creation of Multilingual Lexical-Semantic Dictionaries

Federico Martelli[1] **Luigi Procopio**[1,2*] **Edoardo Barba**[1] **Roberto Navigli**[1]
[1] Sapienza NLP Group, Sapienza University of Rome
[2] Sunglasses.ai
{martelli,barba,navigli}@diag.uniroma1.it
luigi.procopio@sunglasses.ai

Abstract

Lexical-semantic resources such as wordnets and multilingual dictionaries often suffer from significant coverage issues, especially in languages other than English. While improving their coverage manually is a prohibitively expensive undertaking, current approaches to the automatic creation of such resources fail to investigate the latest advances achieved in relevant fields, such as cross-lingual annotation projection. In this work, we address these shortcomings and propose LEXICOMATIC, a novel resource-independent approach to the automatic construction and expansion of multilingual semantic dictionaries, in which we formulate the task as an annotation projection problem. In addition, we tackle the lack of a comprehensive multilingual evaluation framework and put forward a new entirely manually-curated benchmark featuring 9 languages. We evaluate LEXICOMATIC with an extensive array of experiments and demonstrate the effectiveness of our approach, achieving a new state of the art across all languages under consideration. We release our novel evaluation benchmark at: https://github.com/SapienzaNLP/lexicomatic.

1 Introduction

Lexical-semantic resources, like wordnets and computational lexicons, play a key role in a wide range of Natural Language Understanding (NLU) tasks (Navigli, 2018) such as Word Sense Disambiguation (Bevilacqua et al., 2021, WSD), Semantic Role Labeling (Gildea and Jurafsky, 2000, SRL), Semantic Parsing (Martinez Lorenzo et al., 2022), when investigating semantic biases in Machine Translation (Campolungo et al., 2022, MT), and in a broad spectrum of NLU approaches. For instance, in WSD not only do these dictionary-like resources enable an explicit representation of words and their meanings, leveraged in many knowledge-based approaches (Agirre et al., 2014; Moro et al., 2014; Scozzafava et al., 2020), but they have also proven to be highly beneficial when integrated into neural systems (Bevilacqua and Navigli, 2020; El Sheikh et al., 2021; Conia and Navigli, 2021). For the purposes of this work, we refer to such resources as *multilingual semantic dictionaries*, and differentiate between these and the *bilingual dictionaries* typically used in MT (Klementiev et al., 2012; Irvine and Callison-Burch, 2014). In fact, bilingual dictionaries contain a list of possible translations in a target language for each word in a source language, with no distinction between the different meanings conveyed by such words.

Notably, a crucial limitation affecting multilingual semantic dictionaries is their insufficient coverage, especially when scaling to multiple languages or when considering mid- and low-resource ones. While, on the one hand, creating such resources manually is a very expensive endeavour, on the other hand, current automatic approaches have failed to investigate the benefits derived from recent breakthroughs achieved in relevant fields, such as cross-lingual label propagation (Procopio et al., 2021), WSD (Barba et al., 2021b) and word alignment.

In this paper we address these shortcomings and propose LEXICOMATIC, a novel approach to the automatic construction of multilingual semantic dictionaries. Starting from a monolingual semantic dictionary D in language L_0 and a set of target languages $\{L_1, \ldots, L_n\}$, we first build a synthetic L_0-centric parallel corpus and, then, leveraging WSD and word alignment, generate, for every sense in D, its corresponding lexicalizations in each target language. We use the wordnet creation task (Neale, 2018), i.e. the computational task of automatically constructing a wordnet, either from scratch, or by leveraging an already existing one, as our main test case. Since, to the best of our knowledge, no comprehensive multilingual evaluation suite is currently available, we propose a novel manually-

*Work carried out during the PhD programme at the Sapienza University of Rome.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 820–833
November 1–4, 2023.

curated framework comprising 9 languages, including mid- and low-resource ones. For each of these, we find that our approach achieves significantly better performances than its state-of-the-art alternatives in terms of both F_1 score, $F_{0.5}$ score and WordNet core coverage.[1] Furthermore, since LEXICOMATIC also generates silver WSD datasets for each language considered, we investigate their quality in our experiments.

Our contributions are therefore as follows:

1. We propose a novel approach to the automatic construction of multilingual semantic dictionaries.

2. We put forward a new manually-curated multilingual evaluation suite for the wordnet creation task, covering 9 languages, ranging from mid- to low-resources ones.

3. We evaluate our approach extensively and carry out a performance analysis. Furthermore, we demonstrate the scalability of LEXICOMATIC when adopting a sense inventory different from Princeton WordNet[2] (Miller, 1995, PWN) (see Appendix A).

We release our novel evaluation benchmark at: https://github.com/SapienzaNLP/lexicomatic.

2 Related Work

Among lexical-semantic resources, wordnets are arguably the most popular and widely used. Their automatic creation has been addressed by several approaches put forward during the course of recent decades (Neale, 2018). Depending on the strategy adopted, such approaches have been divided into two main paradigms (Vossen, 1998):

1. **Expand** or extend approaches, in which translations of PWN synsets are used to create new wordnets in other languages;

2. **Merge** approaches, which, instead, create wordnets independently and then map them to PWN.

While the merge approaches are able to overcome several linguistic issues, the expand or extend approaches have become the *de facto* standard in literature, thanks to their speed gain and ease of connection with PWN.

Broadly speaking, expand approaches present a common structure consisting of two steps: i) *candidate retrieval*, where a list of candidate words in the target language to be assigned to a given PWN synset is produced; ii) *candidate selection*, where a scoring function is used to discard or assign candidate words to a given PWN synset. For example, Lee et al. (2000) propose the construction of a Korean wordnet by linking words in Korean derived from a bilingual machine-readable dictionary to PWN. Semantic ambiguities are then removed by combining 6 different heuristics with decision-tree learning. Instead, Montazery and Faili (2010) rely on different word similarity scores, such as the mutual information and other measures based on WordNet, to link words in Persian to PWN synsets. Along these lines, Lam et al. (2014) leverage publicly available wordnets, machine translation and bilingual dictionaries to translate synsets derived from existing wordnets into different languages, including endangered languages, such as Dimasa and Karbi. Finally, a scoring method is used to identify the best translations. Instead, Taghizadeh and Faili (2016) propose the automatic construction of a Persian WordNet, by leveraging a bilingual dictionary and a monolingual corpus as resources and unsupervised WSD to disambiguate candidate words.

With the advent of word embeddings, several approaches started using these new representations. For instance, Al Tarouti and Kalita (2016) use word2vec (Mikolov et al., 2013) to improve the candidate selection step. Instead, Khodak et al. (2017) compute word and synset representations to score the association between a candidate word and a given synset, discarding (word, synset) pairs below a certain threshold.

More closely related to our work, some approaches rely on WSD, parallel corpora and word alignment systems to retrieve, for each synset, a list of words in the target language. For instance, Sagot and Fišer (2008) use five different parallel corpora and disambiguate each aligned word by intersection of its possible senses in its corresponding inventories. Finally, closest to our work, Oliver (2014) leverages a WSD system on the English side of a parallel corpus and retrieves all the possible translations in the target language by using a word alignment system. More recent research works, however, did not investigate this promising direction further. To fill this gap, in this work we aim

[1] https://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt

[2] https://wordnet.princeton.edu/

to demonstrate that recent neural breakthroughs in WSD (Barba et al., 2021a) and cross-lingual label projection via word alignment (Procopio et al., 2021) allow us to achieve a new state of the art in the wordnet creation task.

3 LexicoMatic

In this section, we introduce LEXICOMATIC, our approach to the automatic construction and expansion of multilingual semantic dictionaries. First, we outline our overall process (Section 3.1). Then, we focus on the core modules of our approach, describing the systems we adopt for WSD (Section 3.2) and word alignment (Section 3.3). Finally, we detail our data aggregation strategy (Section 3.4).

3.1 Formulation

Starting from a monolingual dictionary in language L, we frame the automatic expansion of this dictionary towards a set of target languages as a 3-stage process over L-centric parallel corpora.[3] Formally, let D be a dictionary in language L, comprising a list of lexemes $l_1, \ldots, l_m$, with each l_i $\forall i \in [1, m]$ associated with a collection of $c(l_i)$ textual definitions $\delta(l_i) = \{d_1^{l_i}, \ldots, d_{c(l_i)}^{l_i}\}$ expressing its possible meanings; both $l_1, \ldots, l_m$ and their definitions are in language L. Then, given a list of target languages $\tilde{L}_1, \ldots, \tilde{L}_k$, we formulate our objective as follows: for each target language $\tilde{L}$, we wish to yield the lexemes in $\tilde{L}$ corresponding to every pair $(l_i, d_j^{l_i})$ $\forall i \in [1, m]$ $\forall j \in [1, c(l_i)]$.

To achieve this objective, we put forward the following approach: denote by $s_1, \ldots, s_n$ a list of sentences in language L, and by $t_1, \ldots, t_n$ its parallel counterparts in a target language $\tilde{L}$. Then, for each $(l_i, d_j^{l_i})$, we generate the corresponding lexemes in $\tilde{L}$ as follows: we first *disambiguate* the words in $s_1, \ldots, s_n$ against D, that is, we pair each word with its most suitable meaning in D. Subsequently, we perform *word alignment* on each pair of parallel sentences (s_j, t_j) $\forall j \in [1, n]$. These two steps essentially yield a corpus where words in language L are paired with their meanings in D and linked to their counterparts in language $\tilde{L}$. With this data at our disposal, we conclude our process by employing an *aggregation strategy* over the processed parallel sentences, producing, for each $(l_i, d_j^{l_i})$, a list of corresponding lexicalizations in language $\tilde{L}$.

3.2 Word Sense Disambiguation

The first step in our approach aims at disambiguating the words in the sentences $s_1, \ldots, s_n$ against D. That is, $\forall j \in [1, n]$, we need to link each word form in s_j, whose lexeme we denote by $\tilde{l}$, to the definition in $\delta(\tilde{l})$ that best expresses its meaning. Note that these word forms might be single tokens (e.g., nouns) or multi-words (e.g., phrasal verbs); henceforth, we will use $\sigma_{j,1}, \ldots, \sigma_{j,|\sigma_j|}$ to denote the $|\sigma_j|$ word forms occurring in sentence s_j.

To perform this task, previous WSD approaches (Bevilacqua and Navigli, 2020; Conia and Navigli, 2021) adopt supervised classifiers that require training corpora sense-tagged against D. However, often such corpora labeled with D are not available across languages and, since manually producing them is prohibitively expensive, using these approaches would restrict the applicability of our process to a limited number of dictionaries. Thus, in order to drop this requirement, we focus here on the recent trend in literature of models tackling WSD via definition-selection formulations (Huang et al., 2019; Blevins and Zettlemoyer, 2020; Barba et al., 2021a): given a word in context, models dynamically receive a pool of textual definitions and are trained to select the most suitable one from among these. This framing allows us to perform disambiguation against D even in cases where D has either scarce or no sense-tagged data. Indeed, in these scenarios, we can train on already existing WSD corpora, annotated with some $D' \neq D$, and, then, *zero-shot* over D.

3.3 Word Alignment

As our second step, we perform word alignment between each pair of parallel sentences (s_j, t_j) $\forall j \in [1, n]$. To this end, we build on top of the Transformer-based discriminative model for word alignment introduced by Procopio et al. (2021) and employ a framework consisting of 2 steps inspired by the Expectation-Maximization algorithm (Dempster et al., 1977).

Given a list of parallel sentences (s_j, t_j) $\forall j \in [1, n]$, first we use the discriminative model to perform word alignment, pairing the tokens $s_{j,1}, \ldots, s_{j,|s_j|}$ in s_j to their counterparts $t_{j,1}, \ldots, t_{j,|t_j|}$ in t_j. Subsequently, since the disambiguation is performed over word forms that

[3] As will be seen, note that *L-centric* parallel corpora are not a strict requirement for LEXICOMATIC, which can also simply rely on (non *L*-centric) standard parallel corpora. We highlight this requirement at this point only for presentation simplicity and computational efficiency.

might span over multiple tokens, we bring together the alignments produced at span level. Note that, however, since the discriminative model may emit non-contiguous alignments (e.g., $s_{j,1}$ aligned to $t_{j,1}$ and $t_{j,4}$), $\forall k \in [1, |\sigma_j|]$, this operation involves choosing the most suitable option for $\sigma_{j,k}$ from among the multiple aligned spans $\tau_{j,1}, \ldots, \tau_{j,|\tau_j|}$ identified over t_j.

To address this issue, we adopt a strategy consisting of 2 passes. We first employ a simple heuristic, processing $\sigma_{j,1}, \ldots, \sigma_{j,|\sigma_j|}$ by increasing number of spans aligned over t_j and selecting each time the first target span that does not overlap with any of those previously chosen. Once this pass is completed, we leverage the aligned parallel sentences to compute a score for each proposed possible alignment:

$$f(\sigma, \tau) = \frac{\text{\# sentences with } (\sigma, \tau) \text{ aligned}}{\text{\# sentences with } (\sigma, \tau)}$$

where σ and τ are two possible word forms in languages L and $\tilde{L}$, respectively. We now perform a second pass over the parallel sentences, using the computed score in the selection process: $\forall k \in [1, |\sigma_j|]$, we select the target span aligned to $\sigma_{j,k}$ with the highest score, enforcing a minimum of α to reduce potential noise. We use $\alpha = 0.5$ in all our experiments since it expresses the condition where both the prior probability and the conditional probability given by the classifier[4] line up.

3.4 Aggregation Strategy

Once completed, the disambiguation and word alignment steps of our process produce a corpus where each source span $\sigma_{j,k}$ $\forall j \in [1, n]$ $\forall k \in [1, |\sigma_j|]$ is paired with its most suitable definition, and either associated with its counterpart span $\tau_{j,k'}$, with $k' \in [1, |\tau_j|]$, occurring in sentence t_j or marked as unaligned.[5]

Therefore, to conclude our process, we first group the collection of successfully aligned target spans by the lexeme-definition pair assigned to their source counterparts; this process results in a set of translation candidates $\mathrm{T}(l_i, d_j^{l_i}) = \{\tau_1, \ldots, \tau_p\}$ for each lexeme-definition pair $(l_i, d_j^{l_i})$. Then, we sort these candidates and produce $\hat{\mathrm{T}}(l_i, d_j^{l_i}) = [\hat{\tau}_1, \ldots, \hat{\tau}_p]$ s.t. $g(l_i, d_j^{l_i}, \hat{\tau}_z) > g(l_i, d_j^{l_i}, \hat{\tau}_{z+1})$ $\forall z \in [1, p]$, where $g(l_i, d_j^{l_i}, \hat{\tau}_z)$ represents the number of times $\hat{\tau}_z$ was aligned in the bitext to l_i with the meaning expressed by $d_j^{l_i}$. Finally, to reduce the amount of noise introduced by spurious alignments and wrong disambiguation, we apply a simple method to remove the translation candidates that are most likely to be wrong: for each lexeme-definition pair $(l_i, d_j^{l_i})$, we apply an L1-normalization on the vector $v = \left\langle g(l_i, d_j^{l_i}, \hat{\tau}_1), \ldots, g(l_i, d_j^{l_i}, \hat{\tau}_p) \right\rangle$ and select the first h candidates such that their normalized scores sum up to a hyperparameter β. Besides filtering out potential noise, this hyperparameter also allows us to bias LEXICOMATIC towards a more precision- or recall-oriented behavior.

4 Wordnet Construction

We now assess the effectiveness of LEXICOMATIC, using the wordnet creation task as our test case. We first present a novel evaluation suite comprising 9 languages that we propose for this task. Subsequently, we describe the experimental setup which we use in this setting and, finally, evaluate LEXICOMATIC with current state-of-the-art alternatives.

4.1 Evaluation Suite

To the best of our knowledge, no comprehensive multilingual evaluation suite is currently available for the wordnet creation task. Therefore, we address this limitation and propose a novel framework partially inspired by ML50 (Tang et al., 2020) and spanning over 9 languages, namely, Arabic, Chinese, French, German, Italian, Korean, Russian, Spanish and Swedish. Since parallel corpora are the main requirement on target languages for LEXICOMATIC,[6] we follow the same classification adopted by Tang et al. (2020) and divide these languages into three groups depending on the availability of such resources: low-resource (Swedish), mid-resource (Arabic, Italian and Korean) and high-resource (Chinese, French, German, Russian and Spanish).[7]

For each of these languages, we manually create a test set as follows. First, we lemmatize and label with the part of speech (POS) the corresponding

[4] For a given (σ, τ) pair to be considered, the classifier must have yielded a probability for their alignment that is > 0.5.

[5] This might occur if the selection process filters out all possible alignments, or no alignments have been provided.

[6] While the alignment model needs manually-aligned data, a few hundred sentences suffice (Procopio et al., 2021).

[7] Corresponding to 10k−100K, 100K−1M and 10M+ groups in Tang et al. (2020).

Wikipedia corpus[8] using Stanza[9] (Qi et al., 2020) and compute the absolute frequency ϕ of each lexeme, that is, each (lemma, POS) pair.[10] Then, we discard lexemes with $\phi \leq 1000$, to reduce potential noise, and divide those remaining into three frequency classes, depending on their ϕ value: specifically, denoting by ϕ_{25th} and ϕ_{50th} the 25-th and 50-th percentiles, the three classes are comprised of samples such that $\phi \geq \phi_{50th}$, $\phi_{25th} \leq \phi < \phi_{50th}$ and $\phi < \phi_{25th}$, respectively. For each class, we manually validate all extracted lexemes, discarding spurious ones,[11] and, then, randomly sample 200 elements. Finally, for each of these, we retrieve the corresponding synsets from BabelNet (Navigli and Ponzetto, 2012), a large multilingual semantic network built by combining a number of heterogeneous resources including PWN, and ask professional linguists to manually validate each (lexeme, synset) pair. These final pairs over the three classes constitute the test set for the language under consideration. Table 1 reports coverage statistics of our test sets on each language, both per POS class and aggregated. Further information regarding the annotation process and guidelines can be found in the next subsection.

4.2 Annotation Process and Guidelines

The manual creation of our comprehensive multilingual evaluation suite is carried out by six professional linguists or translators. We require each professional annotator to work in a language in which they have a C2 level of proficiency according to the Common European Framework of Reference for Languages, as well as proven experience in the creation and expansion of lexical-semantic resources. All annotators are paid at an agreed hourly rate which is higher than the legal minimum pay per hour in their country of residence, if available.

In order to ensure data consistency across languages, we devise and adopt specific annotation guidelines. In this way, shared linguistic criteria are adopted to perform the manual annotation and validation. For instance, we use hypernyms to determine whether a (lemma, POS) pair should be associated with a given synset, i.e. lemmas pertaining to a given synset should share the same hypernym according to reputable lexicographic resources such as WordNet for the English language. Importantly, during the annotation process, we encounter some language-specific peculiarities and exceptions, e.g. verb aspects in Russian or compounds (Komposita) in German. Such cases are discussed and subsequently addressed in joint annotation sessions.

	Low	Mid			High				
POS	SV	AR	IT	KO	DE	ES	FR	RU	ZH
NOUN	1543	1768	1672	801	1352	2021	1703	1178	1076
ADJ	94	213	646	16	324	524	506	-	-
VERB	151	581	503	40	641	910	596	516	445
ADV	21	5	112	53	21	94	84	-	30
TOTAL	1809	2567	2933	910	2338	3549	2889	1694	1551

Table 1: Number of distinct synsets, divided by POS, in LEXICOMATIC test sets.

4.3 LexicoMatic Setup

Word Sense Disambiguation As our disambiguation system, we use ESCHER (Barba et al., 2021a), a Transformer-based architecture that frames WSD as a text extraction problem. Specifically, since the dictionary under consideration is PWN, we employ the model released by the authors[12] that is trained on SemCor (Miller et al., 1993), a large manually-annotated English dataset featuring 33 362 sentences and 226 036 tagged instances. Our choice of this system is motivated by the strong performance which ESCHER attains both when evaluated on the same sense inventory used at training time and in zero-shot scenarios.

Word Alignment To train the word alignment model, we use the manually-annotated datasets made available by Procopio et al. (2021) covering English and one or other of the following languages: French, German, Spanish and Italian. Instead, as far as the remaining languages are concerned, we leverage proprietary in-house datasets, which are created with the same method as that adopted for the aforementioned datasets: approximately 300 sentences are collected from WikiMatrix (Schwenk et al., 2019) and professional linguists are asked to manually annotate them.

Parallel Corpora LEXICOMATIC relies on parallel corpora to transfer sense annotations and

[8] https://en.wikipedia.org/. We use the dump of December 2021.

[9] https://stanfordnlp.github.io/stanza/index.html

[10] In order to include both single tokens and multiword expressions, we create a vocabulary of multiword expressions from titles of Wikipedia pages and a manually-selected set of common multiword expressions.

[11] Lexemes are considered to be incorrect and thus discarded if one or more of the following issues can be identified: i) lexicalization issues; ii) wrong language, i.e., a lemma is in a language other than the one under consideration.

[12] https://github.com/SapienzaNLP/esc

gather different lexicalizations for each synset. However, the quality and the coverage of the resources created is directly proportional to the heterogeneity and number of source sentences considered. This is especially troublesome for low- and mid-resource languages, where the amount of gold parallel data is relatively low (<1M sentences). Therefore, to cope with this issue, we here adopt synthetic parallel sentences, that is, sentences translated via MT systems. Specifically, we use the neural machine translation system presented by Tang et al. (2020, mBART50). This strategy allows us to generate an arbitrarily large amount of parallel sentences for each language in our evaluation suite. To ensure a good coverage, we use a sample of 1M randomly selected sentences from English Wikipedia as the source corpus.

4.4 Comparison Systems

As our baseline, we consider a simple approach (PWN + MT) where we use mBART50 to translate the lexicalizations of each synset in PWN. For each language, in order to give more context to the MT systems, we postpend the synset definition to the comma-separated list of its lexicalizations. For example, to get the lexicalizations for the synset {fire, flame, flaming} with definition *the process of combustion of inflammable materials producing heat and light*, we input to the model the sequence *fire, flame, flaming: the process of combustion of inflammable materials producing heat and light*, and extract, from the translated sentence, the lexicalizations in the target language. As comparison systems, we consider Universal Wordnet (De Melo and Weikum, 2009, UWN), an expand approach built upon bilingual dictionaries and an ensemble of statistical heuristics, and Extended Open Multilingual Wordnet (Bond and Foster, 2013, EOWN) which merged Open Multilingual Wordnet (Bond and Paik, 2012) with data collected automatically from Wiktionary;[13] both these works cover all languages in our evaluation suite. Furthermore, we consider Sagot and Fišer (2008, WOLF) for French and the system recently proposed by Khodak et al. (2017, AWCWE) for French and Russian.

4.5 Results

Table 2 shows the performances achieved by the systems under consideration over our evaluation suite. In particular, besides precision, recall and F_1 score, we further report as in Khodak et al. (2017) the $F_{.5}$ score, a variant of F_1 score that is more biased towards precision, and the coverage statistic, that is, the percentage of synsets in core WordNet[14] that are present in the resource under evaluation.

As a first result, we note the significant performances that the MT baseline attains, particularly in terms of recall and coverage. This is especially interesting since the only discerning signals the models receive as regards the desired meaning of a given term are the definition and the other English lexicalizations of the corresponding synset.

Moving to our actual system, we consider here how LEXICOMATIC fares for different β values, namely $[0.7, 0.9, 1.0]$. Indeed, differently from the other systems that are skewed towards precision by design since it is more useful in practical scenarios (Khodak et al., 2017), our approach enables us to select the desired trade-off, which might depend upon the use case under consideration, by adjusting β: lower values result in more conservative selection strategies that favour precision over recall, whereas higher ones produce the opposite behavior. We can see this trend in Table 2, where moving from $\beta = 0.7$ to $\beta = 1.0$ causes precision to decrease and recall to increase.

Finally, compared to its competitors, LEXICOMATIC surpasses all its alternatives considered here across the board in terms of F_1 score and $F_{.5}$ score. Interestingly, even with $\beta = 0.7$, LEXICOMATIC is still more oriented towards recall than the majority of its alternatives. These findings suggest that LEXICOMATIC is indeed an effective option for the automatic creation of wordnets. As a matter of interest, we investigate the scalability of LEXICOMATIC when adopting a sense inventory different from PWN and report the results in Appendix A.

5 Multilingual WSD

As a by-product of our first two steps, our process results in the automatic creation of sense-tagged corpora in languages $\tilde{L}_1, \ldots, \tilde{L}_k$ that can be used to train WSD systems. This training operation, besides naturally yielding models for each language we cover, also acts as an interesting evaluation proxy for the created wordnets. Indeed, the results attained on multilingual WSD benchmarks provide hints as to the quality of the automatically-

[13] https://www.wiktionary.org/

[14] In order to enable multilinguality, we convert senses to synsets.

	Model	Precision	Recall	F_1	$F_{.5}$	Synsets	Senses	Coverage
Arabic	PWN + MT	41.5	6.8	11.7	20.6	117653	230851	93.4%
	Universal Wordnet	5.8	5.4	5.6	5.7	5991	7791	27.2%
	EOMWN	**49.3**	10.4	17.2	28.1	6891	7791	37.6%
	LEXICOMATIC$_{\beta=0.7}$	28.0	33.8	**30.6**	**29.0**	17038	21974	75.3%
	LEXICOMATIC$_{\beta=0.9}$	25.9	36.5	30.3	27.5	17038	28335	75.3%
	LEXICOMATIC$_{\beta=1.0}$	23.4	**38.7**	29.2	25.4	17038	38507	75.3%
Chinese	PWN + MT	22.7	25.8	24.2	23.3	117653	156223	99.7%
	Universal Wordnet	41.0	10.2	16.4	25.6	22665	104760	53.2 %
	EOMWN	**44.7**	14.3	21.7	31.4	12128	19079	49.4 %
	LEXICOMATIC$_{\beta=0.7}$	31.3	36.3	**33.6**	**32.2**	20251	23124	83.7 %
	LEXICOMATIC$_{\beta=0.9}$	27.8	41.7	33.4	29.8	20251	27843	83.7 %
	LEXICOMATIC$_{\beta=1.0}$	22.7	**45.0**	30.2	25.2	20251	37622	83.7 %
French	PWN + MT	36.8	35.6	36.1	36.5	117653	155751	99.6%
	WOLF	46.7	30.2	36.6	42.1	59807	59087	92.3%
	Universal Wordnet	48.2	27.8	35.3	42.1	39491	72009	74.9%
	EOMWN	**69.2**	19.2	30.0	45.4	20447	27150	63.16%
	AWCWE	42.1	40.5	41.3	41.8	53203	93121	91.5%
	LEXICOMATIC$_{\beta=0.7}$	55.6	39.1	45.9	**51.3**	30505	38595	93.4%
	LEXICOMATIC$_{\beta=0.9}$	50.2	42.5	**46.1**	48.5	30505	48743	93.4%
	LEXICOMATIC$_{\beta=1.0}$	43.0	**45.9**	44.4	43.5	30505	64916	93.4%
German	PWN + MT	29.6	25.0	27.1	28.6	117653	154979	99.6%
	Universal Wordnet	41.5	22.1	28.8	35.3	50488	110496	76.0%
	EOMWN	**61.7**	14.8	23.9	37.8	19673	29616	63.6%
	LEXICOMATIC$_{\beta=0.7}$	48.0	31.5	38.1	**43.5**	29543	37156	93.9%
	LEXICOMATIC$_{\beta=0.9}$	43.3	35.0	**38.7**	41.3	29543	45931	93.9%
	LEXICOMATIC$_{\beta=1.0}$	35.8	**38.3**	37.0	36.3	29543	60107	93.9%
Italian	PWN + MT	36.8	39.8	38.3	37.4	117653	155323	99.6%
	Universal Wordnet	55.8	24.0	33.6	44.1	37638	59805	72.6%
	EOMWN	**77.2**	17.6	28.7	46.1	14603	18710	52.8%
	LEXICOMATIC$_{\beta=0.7}$	56.6	44.4	**49.8**	**53.6**	31215	37676	95.0%
	LEXICOMATIC$_{\beta=0.9}$	51.4	48.1	49.7	50.7	31215	45679	95.0%
	LEXICOMATIC$_{\beta=1.0}$	43.9	**52.0**	47.6	45.3	31215	61364	95.0%
Korean	PWN + MT	8.3	23.4	12.3	9.6	117653	183471	94.4%
	Universal Wordnet	27.6	14.1	18.7	23.2	37940	69080	31.5%
	EOMWN	**39.4**	10.5	16.6	25.4	6287	9268	52.1%
	LEXICOMATIC$_{\beta=0.7}$	32.6	21.0	25.6	29.4	28564	30973	71.2%
	LEXICOMATIC$_{\beta=0.9}$	30.7	27.0	**28.7**	**29.9**	28564	46462	71.2%
	LEXICOMATIC$_{\beta=1.0}$	26.7	**29.8**	28.2	27.3	28564	65216	71.2%
Russian	PWN + MT	24.9	26.8	25.8	25.3	117653	156500	99.7%
	Universal Wordnet	37.3	16.3	22.7	29.7	30009	57479	67.0%
	EOMWN	**44.3**	17.2	24.8	33.7	19980	33716	64.3%
	AWCWE	33.8	24.9	28.7	31.5	50844	102605	91.5%
	LEXICOMATIC$_{\beta=0.7}$	37.4	33.2	**35.2**	**36.5**	26065	33961	89.0 %
	LEXICOMATIC$_{\beta=0.9}$	32.1	37.3	34.5	33.0	26065	43054	89.0 %
	LEXICOMATIC$_{\beta=1.0}$	24.9	**39.9**	30.7	27.0	26065	57192	89.0 %
Spanish	PWN + MT	42.8	32.5	36.9	40.3	117653	155486	99.7%
	Universal Wordnet	61.7	19.1	29.2	42.7	33920	53497	65.4%
	EOMWN	**72.4**	15.8	25.9	42.1	18428	27868	60.2%
	LEXICOMATIC$_{\beta=0.7}$	59.8	34.3	43.6	**52.0**	31268	37629	94.3%
	LEXICOMATIC$_{\beta=0.9}$	55.0	37.4	**44.5**	50.3	31268	45749	94.3%
	LEXICOMATIC$_{\beta=1.0}$	47.1	**40.8**	43.7	45.7	31268	59954	94.3%
Swedish	PWN + MT	14.8	30.1	19.9	16.5	117653	161793	94.4%
	Universal Wordnet	31.1	28.0	29.5	30.4	23848	33264	61.9%
	EOMWN	**37.3**	19.6	25.7	31.6	11999	16226	50.1%
	LEXICOMATIC$_{\beta=0.7}$	32.5	33.4	**33.0**	**32.7**	28801	35720	93.5%
	LEXICOMATIC$_{\beta=0.9}$	29.7	35.9	32.5	30.8	28801	37119	93.5%
	LEXICOMATIC$_{\beta=1.0}$	26.0	**38.7**	31.1	27.8	28801	50599	93.5%

Table 2: Results on our evaluation framework. In bold the best scores per language for precision, recall, F_1 and $F_{.5}$.

		DE	ES	FR	IT	KO	ZH
LEXICOMATIC	# lexemes	21k	21k	23k	20k	21k	16k
	# instances	5.2M	6.0M	6.0M	6.3M	2.0M	3.2M
	# synsets	30k	31k	31k	31k	15k	20k
XL-WSD	# lexemes	16k	22k	18k	24k	-	-
	# instances	185k	393k	253k	385k	-	-
	# synsets	16k	32k	22k	30k	-	-

Table 3: Statistics of the training sets from LEXICOMATIC (top) and XL-WSD (bottom) in terms of number of lexemes, instances and distinct synsets.

Model	DE	ES	FR	IT	KO	ZH
MCS	76.0	55.6	59.3	52.8	52.5	29.6
∅-shot	83.2	75.8	83.9	77.7	**64.2**	**51.6**
T-SC+WNG	73.8	77.3	71.4	77.7	-	-
ConSeC	**84.2**	77.4	84.4	79.3	-	-
LEXICOMATIC	80.0	**78.5**	**85.3**	**79.4**	62.2	49.1

Table 4: F1 comparison of LEXICOMATIC against the MCS, ∅-shot and T-SC+WNG systems reported in Pasini et al. (2021). We highlight the best system in bold.

disambiguated sentences $t_1, \ldots, t_n$ produced for each language. Therefore, in this section, we examine the performance on multilingual WSD that a reference architecture achieves once trained on the silver corpora that our process generated for the experiments detailed in Section 4.

5.1 Experimental Setup

As the evaluation framework, we consider XL-WSD, a benchmark recently proposed by Pasini et al. (2021) that includes a set of language-specific development and test sets in different languages. In particular, we now focus on German (DE), Spanish (ES), French (FR), Italian (IT), Korean (KO) and Chinese (ZH)[15] and use the corresponding resources to assess the performances of LEXICOMATIC on multilingual WSD. In what follows, we illustrate the architecture of our reference model and provide an analysis of the silver training corpora under consideration.

Model Architecture To be comparable with the resources evaluated in Pasini et al. (2021), we use the same architecture as the WSD classifier, that is, a Transformer-based encoder, namely XLMR-Large (Conneau et al., 2020), followed by a 2-layer feedforward network with swish activation function, batch-normalization and a softmax layer on top. We represent each subword in the input sentence as the sum of the last 4 layers of the Transformer encoder and each word as the average of the vectors corresponding to the subwords it was split into. Finally, the model is trained to assign each instance to its corresponding synset in PWN.

Training Data For each language in our evaluation suite, we use as the training corpus its automatically-disambiguated sentences $t_1, \ldots, t_n$.

We show in Table 3 the number of lexemes, instances and distinct synsets for both our datasets and the silver training resources included in Pasini et al. (2021) for the four European languages.[16] To counter possible excessive skewness towards the most common sense that might occur for some word sense distributions, we limit the number of occurrences for each sense and randomly select up to 10 000 instances when preprocessing this data. Interestingly, besides the difference on the number of instances caused by the bigger collection of sentences we consider, our corpora cover an amount of synsets that is either on par with their reference counterparts (Spanish and Italian) or significantly higher (German and French). Finally, note that, as our purpose here is to further assess the quality of our process, we do not perform any kind of *inventory filtering*, that is, we do not employ the mapping from word to its possible synsets included in XL-WSD in order to avoid the inclusion of incorrect instances that our process might have generated.

5.2 Results

We report in Table 4 the results attained in terms of F_1 score by LEXICOMATIC over the languages considered. For comparison, we consider three systems reported in Pasini et al. (2021), namely i) MCS, where words are always disambiguated to their most common sense, ii) ∅-shot, where the reference architecture is trained on English sense–tagged resources and tasked to *zero-shot* over the test languages, and iii) T-SC+WNG, where the training is performed, instead, over the silver resources released in the reference paper, i.e., an automatically-translated version of SemCor and the Princeton WordNet Gloss Corpus. Furthermore, we also include ConSeC (Barba et al., 2021b), a recent extractive approach to WSD that is trained

[15] The choice of this language set is the result of the intersection between the languages we considered in Section 4 and those included in XL-WSD.

[16] No silver resource was released for Korean and Chinese.

on the silver resources released in XL-WSD and that represents the current state of the art in this benchmark.

As Table 4 highlights, training upon the corpora which LEXICOMATIC generates results in performances that are at least on par with the alternatives reported. Specifically, we achieve a new state of the art on 3 languages, namely Spanish, French and Italian, and significantly close the gap between ∅-shot and silver resources on German. On Korean and Chinese, we attain performances inferior to plain zero-shot, but still competitive and significantly higher than the MCS baseline. Therefore, the first two steps of our process do generate high-quality corpora and this finding has interesting ramifications. Indeed, on the one hand, it suggests that the aggregation strategy is similarly expected to generate high-quality lexicalizations and, on the other hand, the fact that our process produces a competitive sense-tagged corpus on each language it is applied upon, is a significant result in its own right.

6 Conclusions

In this work, we introduce LEXICOMATIC, a novel resource-independent approach to the automatic construction and expansion of multilingual semantic dictionaries. By leveraging recent advances in WSD and word alignment, we frame this task as an annotation projection problem over parallel corpora and find this strategy to be particularly effective. Using the wordnet creation task as our main test case, we find that LEXICOMATIC surpasses its alternatives by a large margin, in terms of both F_1 score and $F_{0.5}$ score, when testing against a new evaluation suite covering 9 languages which we put forward. Crucially, our new benchmark is intended to address the current lack of a comprehensive multilingual alternative for this task.

As future work, we plan to further develop our evaluation benchmark, especially so as to expand the number of low-resource languages covered, and investigate the applicability of LEXICOMATIC to other resources beyond PWN.

Limitations

In this section we discuss some limitations that we believe our work currently presents.

First, our method requires a Machine Translation system to generate a translated silver corpus and human annotators to create the training data for word alignment. This might constrain its applicability for some mid-to-low resource languages.

Second, our evaluation requires the availability of human annotators and a coverage with good recall from lemmas to BabelNet synsets in order to scale over a new language. Depending on the language under consideration, the availability of these resources might be limited and, paired with the overall complexity of the task, this implies that expanding our evaluation to span over more languages will be a costly process that requires time.

Finally, our framing as annotation projection might not be applicable to languages that present a high number of translation divergences (e.g., the English adverb *usually* in *John usually goes home* corresponds to the Spanish verb *suele* in *Juan suele ir a casa* (Blloshmi et al., 2020)) as senses would be paired with lemmas that have different syntactic properties from their English counterparts.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487, of the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme, and of the PNRR MUR project PE0000013-FAIR. The authors thank Babelscape for supporting and performing the annotation and evaluation work in many languages.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Feras Al Tarouti and Jugal Kalita. 2016. Enhancing automatic wordnet construction using word embeddings. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 30–34.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. 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.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. 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*, pages 2854–2864.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Terra Blevins and Luke Zettlemoyer. 2020. 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*, pages 1006–1017, Online. Association for Computational Linguistics.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online. Association for Computational Linguistics.

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

Francis Bond and Kyonghee Paik. 2012. A Survey of Wordnets and their Licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. 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.

Simone Conia and Roberto Navigli. 2021. Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. In *Proceedings of the EACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Gerard De Melo and Gerhard Weikum. 2009. Towards a Universal Wordnet by Learning from Combined Evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Ahmed El Sheikh, Michele Bevilacqua, and Roberto Navigli. 2021. Integrating Personalized Pagerank into Neural Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9092–9098.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. 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)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating phrase translations for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170, Ann Arbor, Michigan. Association for Computational Linguistics.

Mikhail Khodak, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. 2017. Automated Wordnet Construction Using Word Embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 12–23.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140, Avignon, France. Association for Computational Linguistics.

Khang Nhut Lam, Feras Al Tarouti, and Jugal Kalita. 2014. Automatically Constructing Wordnet Synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 106–111.

Changki Lee, Geunbae Lee, and Seo Jung Yun. 2000. Automatic Wordnet Mapping Using Word Sense Disambiguation. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the*

Association for Computational Linguistics-Volume 13, pages 142–147.

Abelardo Carlos Martinez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741.

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

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

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308.

Mortaza Montazery and Heshaam Faili. 2010. Automatic persian wordnet construction. In *Coling 2010: Posters*, pages 846–850.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (present and future) use. In *IJCAI*, pages 5697–5702.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Steven Neale. 2018. A survey on automatically-constructed wordnets and their evaluation: Lexical and word embedding-based approaches. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Antoni Oliver. 2014. Wn-toolkit: Automatic generation of WordNets following the expand model. In *Proceedings of the Seventh Global Wordnet Conference*, pages 7–15.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 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*, pages 3915–3921. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. 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*, pages 99–110.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. *OntoLex 2008 Programme*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135m Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 37–46. Association for Computational Linguistics.

Nasrin Taghizadeh and Hesham Faili. 2016. Automatic Wordnet Development for Low-Resource Languages Using Cross-Lingual WSD. *Journal of Artificial Intelligence Research*, 56:61–87.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Piek Vossen. 1998. EuroWordNet: A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

A Scaling to Other Dictionaries

In this Appendix, we show how LEXICOMATIC can be used effectively to construct and expand multilingual semantic dictionaries when adopting a sense inventory other than PWN.

Differently from PWN for which a considerable amount of manually-annotated data is available, most dictionaries do not have this kind of resource

at their disposal. In these settings, LEXICOMATIC has two possible strategies to work around this obstacle. On the one hand, it can rely upon the zero-shot capabilities of the WSD model under consideration. On the other hand, as most dictionaries contain examples for each meaning enumerated, it may use these to generate silver training data. In this section, we compare these two approaches and highlight the key differences between them, conducting our studies on Wiktionary. Wiktionary presents a few major differences compared to PWN, besides the lack of sense-tagged corpora. Most importantly, it is less coarse-grained, with an average of 2.70 senses per lemma in contrast to 1.67 in Princeton WordNet[17], and it has significantly more senses (752 473 compared to the 206 941 in PWN).

A.1 LEXICOMATIC Setup

Word Sense Disambiguation Depending on the strategy chosen, the disambiguation model, which is the only dictionary-dependent component in LEXICOMATIC, needs to be adapted as follows. When resorting to sense inventory zero-shot, the underlying model remains identical and the changes only pertain to the definitions provided at inference time, which originate from Wiktionary rather than PWN. Conversely, when using its meaning examples, we first need to convert these into WSD silver data. To this end, we retrieve the examples provided for each sense s and process them so as to tag the words corresponding to s. This operation results in 66 570 annotated instances, for as many sentences, and we partition them into 3 datasets, namely, *train*, *validation* and *test*, amounting to 60 570, 3000 and 3000 instances, respectively. Then, we replace the underlying disambiguation model with ESCHER trained on these data. We report in Table 5 the F_1 score ESCHER achieves when trained with this configuration (Wiktionary), along with the score achieved, instead, when trained on PWN and tasked to zero-shot on Wiktionary (Wiktionary^{zs}). To better contextualize these results, we further show the performances ESCHER attains when trained and tested on PWN, reporting the F_1 score on the framework proposed by Raganato et al. (2017).[18] We note that, despite the training data being silver and available in a smaller quantity, ESCHER reaches a significant 84.6 F_1 score. As for Wiktionary^{zs}, although it exhibits a significant drop, the overall F_1 score is still remarkable, especially taking into consideration the differences between the two inventories. This finding is particularly promising for our setting, as Wiktionary^{zs} effectively provides an effective estimate of how well LEXICOMATIC can be applied to dictionaries where neither sense-tagged data nor examples are available.

Sense Inventory	Training Instances	Dev	Test
PWN	226036	76.3	80.7
Wiktionary	60570	84.5	84.6
Wiktionary^{zs}	226036	71.5	70.9

Table 5: F_1 scores of ESCHER when trained on different sense inventories, i.e., PWN and Wiktionary.

Test Set & Comparison System To evaluate the resources LEXICOMATIC creates, we leverage additional in-house datasets for each language in our evaluation suite. As comparison systems, since no other work attempts to *translate* Wiktionary to the best of our knowledge, here we report only the performance of our MT baseline (Wiktionary + MT).[19]

A.2 Results

Table 6 shows the overall scores on the test sets. As a first result, we note that, even in this setting, the baseline has competitive performances and, interestingly, reaches F_1 and $F_{.5}$ scores even higher than when *translating* PWN. This is likely due to the longer definitions[20] Wiktionary provides for each sense, which help the translation system better contextualize the lexicalizations.

This trend is reflected on LEXICOMATIC, with Arabic, Chinese and Korean being slight exceptions, where, compared to the rest of the board, LEXICOMATIC achieves a significantly lower recall.[21] Furthermore, as in PWN, decreasing β increases the precision and lowers the recall, even if, in this case, the precision gain is more substantial than the recall loss on average.

Nevertheless, arguably the most interesting finding is the behavior of LEXICOMATIC when used in zero-shot (LEXICOMATIC$^{zs}_{\beta=0.7}$).[22] Indeed, while

[17] Average senses per lemma computed on the intersection of the lemmas in the two inventories.

[18] Results taken from Barba et al. (2021a).

[19] See Section 4.4.

[20] Wiktionary has 12.19 tokens on average, whereas PWN has 10.02.

[21] We believe this phenomenon is the result of the different behavior the annotators had: significantly less candidates were produced for each element compared to the other languages.

[22] Due to space constraints, we only report LEXICO-

	Model	Precision	Recall	F_1	$F_{.5}$	Senses
Arabic	Wiktionary + MT	32.4	13.6	19.2	25.4	406112
	LEXICOMATIC$_{\beta=0.7}$	**46.6**	18.3	26.3	**35.6**	38317
	LEXICOMATIC$_{\beta=0.9}$	38.1	20.4	26.6	32.5	46768
	LEXICOMATIC$_{\beta=1.0}$	33.6	**23.1**	**27.4**	30.8	57626
	LEXICOMATIC$^{zs}_{\beta=0.7}$	46.4	17.9	25.8	35.2	33036
Chinese	Wiktionary + MT	41.6	17.3	24.5	32.5	471320
	LEXICOMATIC$_{\beta=0.7}$	52.3	16.6	25.2	**36.6**	40627
	LEXICOMATIC$_{\beta=0.9}$	45.2	18.3	**26.1**	34.9	67394
	LEXICOMATIC$_{\beta=1.0}$	31.4	**19.6**	24.2	28.0	89648
	LEXICOMATIC$^{zs}_{\beta=0.7}$	**52.5**	16.2	24.7	36.2	35133
French	Wiktionary + MT	48.9	44.1	46.4	47.9	573007
	LEXICOMATIC$_{\beta=0.7}$	**58.7**	43.0	**49.6**	**54.7**	73593
	LEXICOMATIC$_{\beta=0.9}$	48.6	44.7	46.6	47.8	87039
	LEXICOMATIC$_{\beta=1.0}$	37.3	**46.9**	41.6	38.9	102436
	LEXICOMATIC$^{zs}_{\beta=0.7}$	**58.7**	41.5	48.6	54.2	55762
German	Wiktionary + MT	46.3	33.14	38.5	42.7	603238
	LEXICOMATIC$_{\beta=0.7}$	61.0	35.9	45.2	53.5	71683
	LEXICOMATIC$_{\beta=0.9}$	54.0	39.4	**45.5**	50.3	83056
	LEXICOMATIC$_{\beta=1.0}$	44.7	**42.5**	43.6	44.3	96933
	LEXICOMATIC$^{zs}_{\beta=0.7}$	**62.8**	35.5	45.4	**54.5**	54128
Italian	Wiktionary + MT	52.6	41.4	46.3	49.9	578599
	LEXICOMATIC$_{\beta=0.7}$	68.3	43.7	53.3	61.4	73724
	LEXICOMATIC$_{\beta=0.9}$	63.0	47.0	53.9	59.0	84055
	LEXICOMATIC$_{\beta=1.0}$	51.0	**49.1**	50.0	50.6	98625
	LEXICOMATIC$^{zs}_{\beta=0.7}$	**70.0**	44.0	**54.0**	**62.6**	57577
Korean	Wiktionary + MT	22.4	17.0	19.3	21.1	403295
	LEXICOMATIC$_{\beta=0.7}$	**48.2**	15.9	23.9	**34.3**	54563
	LEXICOMATIC$_{\beta=0.9}$	44.3	17.2	**24.8**	33.7	78509
	LEXICOMATIC$_{\beta=1.0}$	39.8	**18.0**	**24.8**	32.0	103798
	LEXICOMATIC$^{zs}_{\beta=0.7}$	48.0	14.0	21.7	32.3	45853
Russian	Wiktionary + MT	46.8	20.3	28.3	37.1	449321
	LEXICOMATIC$_{\beta=0.7}$	**67.3**	30.7	**42.1**	**54.3**	63009
	LEXICOMATIC$_{\beta=0.9}$	57.9	32.7	41.8	50.2	75491
	LEXICOMATIC$_{\beta=1.0}$	48.5	**36.7**	41.8	45.6	90142
	LEXICOMATIC$^{zs}_{\beta=0.7}$	65.3	28.7	39.9	52.0	45007
Spanish	Wiktionary + MT	53.8	42.0	47.2	50.9	568727
	LEXICOMATIC$_{\beta=0.7}$	**67.2**	41.5	**51.4**	**59.8**	72842
	LEXICOMATIC$_{\beta=0.9}$	59.3	43.6	50.2	55.3	83121
	LEXICOMATIC$_{\beta=1.0}$	49.5	**45.7**	47.5	48.7	96875
	LEXICOMATIC$^{zs}_{\beta=0.7}$	68.0	40.3	50.6	**59.8**	57201
Swedish	Wiktionary + MT	52.5	35.9	42.6	48.1	623679
	LEXICOMATIC$_{\beta=0.7}$	63.7	43.8	**51.9**	58.4	64563
	LEXICOMATIC$_{\beta=0.9}$	57.9	45.8	51.2	55.0	72257
	LEXICOMATIC$_{\beta=1.0}$	48.2	**48.6**	48.4	48.3	80944
	LEXICOMATIC$^{zs}_{\beta=0.7}$	**64.5**	42.7	51.3	**58.5**	56865

Table 6: Results on Wiktionary for the 9 languages under consideration. We mark in bold the best scores per language for precision, recall, F_1 and $F_{.5}$.

MATIC^{zs} for $\beta = 0.7$.

Table 5 showed a significant gap between Wiktionary and Wiktionary^{zs}, the F_1 and $F_{0.5}$ scores of LEXICOMATIC$_{\beta=0.7}$ and LEXICOMATIC$^{zs}_{\beta=0.7}$ are almost identical for each language. We believe this is a consequence of the large amount of text disambiguated which, combined with the filtering heuristics adopted, helps the model fill the gap between the two systems. The only significant difference between these lies in the number of senses produced, with LEXICOMATIC$^{zs}_{\beta=0.7}$ emitting consistently less senses than LEXICOMATIC$_{\beta=0.7}$. Nonetheless, this result further backs our claim that leveraging the zero-shot capabilities of the disambiguation model considered is a viable option when translating resources with neither sense-tagged data nor examples.

Finally, the number of total senses covered by our approach compared to the baseline is as few as one-ninth when $\beta = 0.7$ and German is considered (and even lower if we take into account the zero-shot setting). The low percentage of senses covered is due to the large number of Named Entities that are present in Wiktionary but which are, instead, under-represented in parallel corpora.