Semi-Automated Construction of Sense-Annotated Datasets for Practically Any Language

Jai Riley, Bradley Hauer, Nafisa Sadaf Hriti, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei
Hadi Sheikhi, Mahvash Siavashpour, Mohammad Tavakoli, Ning Shi, Grzegorz Kondrak
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
jai.riley@ualberta.ca,gkondrak@ualberta.ca

Abstract

High-quality sense-annotated datasets are vital for evaluating and comparing WSD systems. We present a novel approach to creating parallel sense-annotated datasets, which can be applied to any language that English can be translated into. The method incorporates machine translation, word alignment, sense projection, and sense filtering to produce silver annotations, which can then be revised manually to obtain gold datasets. By applying our method to Farsi, Chinese, and Bengali, we produce new parallel benchmark datasets, which are vetted by native speakers of each language. Our automatically-generated silver datasets are of higher quality than the annotations obtained with recent multilingual WSD systems, particularly on non-European languages.

1 Introduction

Word sense disambiguation (WSD) is the widely studied NLP task of identifying the meaning of a word in context. High-quality sense-annotated datasets are vital for evaluating and comparing WSD systems. However, such datasets are either limited to a small set of high-resource languages or have resource and coverage constraints that limit their generality. Manual sense annotation is a challenging and costly task, and there is no standard, empirically tested multilingual procedure for the production of such "gold" datasets. The use of automatically-generated "silver" datasets is also an area of interest, and its potential for assisting and accelerating manual annotation has not been sufficiently explored. In general, as lexical semantics explores an increasingly diverse set of languages, more research is needed on the creation of multilingual semantically tagged datasets.

Without gold standard sense-annotated texts in a given language, it is not possible to automatically evaluate the performance of models or methods for WSD on that language, and therefore impossible to monitor the progress of the field or relative merits of different methods. For example, AMuSE (Orlando et al., 2021) can perform WSD in 40 languages, but for some of these languages, such as Farsi, there are no usable evaluation datasets. Therefore, the reported multilingual results provide no information on the reliability of AMuSE-generated sense tags on texts in those languages. Similarly, zero-shot methods can theoretically cover a large number of languages, but their quality can only be measured on those for which gold-standard evaluation datasets are available. For low-resource languages, even silver datasets, which are generated via high-precision methods, can be helpful as a source of training data (Barba et al., 2020), or, as we demonstrate in this paper, as a starting point for the efficient creation of gold datasets.

Prior efforts to create sense-annotated data for WSD systems are limited to small datasets for high-resource European languages. Automatic corpus tagging systems (Hauer et al., 2021; Procopio et al., 2021) for producing silver datasets are difficult to deploy in practice because of the availability, complexity, and dependencies of the software. Gold multilingual WSD datasets were created for SemEval tasks by leveraging crowd-sourcing and projection of gold English senses (Moro and Navigli, 2015; Navigli et al., 2013). Pasini et al. (2021) extract individual gold sense annotations from example sentences in multilingual wordnets, which are of varying quality. Moreover, multilingual datasets are generally not parallel or comparable, precluding analysis of relative performance across languages.

In this paper, we present a novel approach to creating sense-annotated datasets, which is applicable to any language for which MT models exist.

Proceedings of the 31st International Conference on Computational Linguistics, pages 6270–6284
January 19–24, 2025. ©2025 Association for Computational Linguistics

We begin with the creation of high-precision silver datasets by translating English sense-annotated datasets, and projecting the sense annotations onto the translations. Our method is easy to deploy, and makes relatively modest assumptions about the available tools and resources for the target language. By creating parallel datasets, our work facilitates the comparison of WSD system performance across different languages. The silver datasets can be revised and expanded by manual annotation into gold-standard datasets via an easy-to-prepare and easy-to-use interface, with less time and effort than the preparation of gold data from scratch.

We apply our auto-tagging approach to five languages, and find that it performs well compared to recent multilingual WSD systems on non-European languages. For instance, we obtain an improvement of over 50% F-score compared to the best available WSD system applicable to Bengali. We subsequently apply our gold annotation procedure to Farsi, Chinese, and Bengali, producing new benchmark datasets, manually vetted by native speakers, and parallel to an existing multilingual dataset. To the best of our knowledge, these are the only publicly-available sense-annotated gold datasets for Farsi and Bengali with an accessible sense inventory, which make it possible to evaluate modern WSD systems on these languages. We make our code and data available on GitHub.

To summarize, the following are the main contributions of this paper: (1) a novel approach for automatically creating semantically tagged text in any language; (2) an annotation framework that uses the above to facilitate manual sense annotation; (3) new parallel WSD benchmark datasets for Farsi, Chinese, and Bengali; (4) empirical validation of our method on the new gold datasets.

2 Related Work

In this section, we discuss approaches to creating both silver and gold WSD datasets. The former are typically produced automatically, while the latter require laborious manual annotation by native speakers, often via specialized instructions and interfaces. We emphasize multilingual approaches, although we also occasionally discuss English-specific work.

In the standard setting, a WSD system must tag all content words in a reference text with a correct sense; failing to disambiguate a designated word incurs a penalty. For example, Orlando et al. (2021) present AMuSE, a system that can sense-tag text in 40 languages, but beyond English, their evaluation is limited to aggregate results over multiple datasets, which are dominated by European languages. In presenting their XL-WSD dataset, Pasini et al. (2021) deploy a multilingual zero-shot WSD method which we refer to as XL-mBERT; again, most results pertain to European languages, and the results are highly variable. Both methods depend on transformer-based language models for sense tagging, which differ in availability and quality across languages; in contrast, our method avoids such dependencies.

Sense projection is an established technique for automatic sense tagging, in which a target word token is assigned the same sense as the corresponding source word, with English typically playing the role of the source language. For example, consider the English sentence *"the **interest** rate is very high"* and its French translation *"le taux d'**intérêt** est très élevé"*. The aligned words (in bold) exhibit *parallel polysemy* (Hauer and Kondrak, 2023) – both can refer to "return paid on a debt" or to "affinity for a particular subject." If the English *interest* has been annotated with its financial sense, we can likewise tag *intérêt* with the same sense, under the assumption that the aligned words have the same meaning. However, this assumption is not guaranteed to hold in every case (Hauer and Kondrak, 2020).

Unlike conventional WSD, systems for *corpus tagging* select and disambiguate only a subset of content words in a corpus, with the goal of producing sense-tagged data for training or fine-tuning WSD models. For example, MultiMirror (Procopio et al., 2021) aligns a sense-annotated English text with its translations, and projects English senses onto a target language, subject to a set of filters. LabelProp (Hauer et al., 2021) involves a similar approach, with an embedding-based WSD method serving as a supplementary check. Other non-projection-based methods include the selective application of an unsupervised WSD method (Delli Bovi et al., 2017; Pasini and Navigli, 2017), transforming Wikipedia text into sense-annotated data by mapping internal links to senses (Scarlini et al., 2019), trivially annotating monosemous words (Loureiro and Camacho-Collados, 2020), generating example sentences from definitions (Barba et al., 2021), and leveraging sense-translation mappings on raw parallel corpora (Hauer and Kondrak, 2023). Our work follows the cross-lingual sense projection paradigm, but sur-

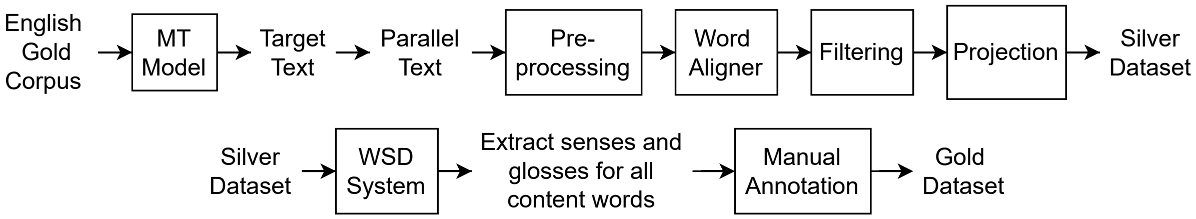

Figure 1: The modules involved in the creation of silver and gold datasets (Sections 3.2 and 3.3, respectively).

passes both MultiMirror and LabelProp in terms of practicability of application to low-resource languages.

Multilingual datasets that are manually annotated with BabelNet (Navigli and Ponzetto, 2012) synsets remain crucial in evaluating WSD systems. Datasets such as SemEval-2013 Task-12 (Navigli et al., 2013) and SemEval-2015 Task-13 (Moro and Navigli, 2015) are scarce and relatively small, even for high-resource European languages. For example, the latter dataset is composed of 138 parallel sentences containing fewer than 3000 sense tags, and incorporates a back-off to sense projection when annotators can not agree. More comprehensively, Pasini et al. (2021) created datasets for 18 languages (XL-WSD), but of those only three originate outside of Europe. The datasets were created by extracting example sentences from multilingual wordnets; therefore, each such sentence contains only a single annotated word token.

For many non-European languages, including Farsi and Bengali, no datasets annotated with BabelNet senses are available, which may be due to their low coverage in BabelNet itself. Datasets available for Farsi (Rouhizadeh et al., 2022) and Bengali (Das Dawn et al., 2022) are created with senses from language-specific wordnets (Rahit et al., 2018; Shamsfard et al., 2010) that have no mapping to BabelNet. Our approach differs from prior work in that it has minimal assumptions about the availability of resources, facilitating application to such low-resource languages, and maximizing the number of languages and domains to which it can be applied.

3 Method

In this section, we describe our two-stage approach to producing parallel WSD datasets. We begin by outlining the theoretical assumptions on which our method is based. Next, we describe our automated approach to creating high-precision silver datasets. Finally, we describe how, using targeted human annotation, these datasets can be efficiently enhanced to gold-standard quality.

3.1 Theoretical Assumptions

Our method is based on three basic assumptions articulated by Hauer and Kondrak (2020). We strive to elucidate theoretical assumptions that are often implicit in prior work and formulate them in a principled manner. These assumptions underlie our innovative and efficient filtering process, which improves the quality of the annotations.

1. Each content word in a text is intended by the writer to express exactly one lexical concept. This assumption holds for the vast majority of word instances in well-written texts, although in some exceptional cases, such as puns, multiple meanings may be intentionally conflated.

2. We assume access to a multilingual wordnet, a resource that organizes words into synonym sets or *synsets*, such that words share a synset if and only if there is a concept both can express. Each synset is assumed to uniquely correspond to a single concept, and contains exactly those words which can express that concept.

3. If two words from different languages are aligned in a sentence and its translation, and if those words are synonyms, then they express the same concept in that context. (In practice, aligned synonyms can be assumed to be mutual literal translations if they are listed as such in a bilingual dictionary, or if they appear to be cognates.)

Taken together, these properties imply that aligned words in a parallel text that share a synset have the same meaning in that context. In cases where those properties hold, we can confidently project senses across bitexts. More specifically, since senses are identified by synsets, any literal translation of a word in context can be correctly tagged with the same synset, *even if that synset fails to include the target lemma.*

Algorithm 1 PROJECTION+

```
1: for each sentence pair $(S, T)$ do
2:     for each word token $w_t$ in sentence $T$ do
3:         if $\exists\, w_s$ s.t. $aligned(w_s, w_t)$ and $LexCategory(w_s) = LexCategory(w_t)$ then
4:             if $same\_synset(w_s, w_t)$ or $translations(w_s, w_t)$ or $cognates(w_s, w_t)$ then
5:                 project the sense tag of $w_s$ onto $w_t$
```

3.2 Automatic Silver Dataset Creation

The pseudo-code shown in Algorithm 1 defines our method for silver dataset creation. It takes as input a sense-annotated dataset in a given source language, which we refer to as the *source dataset*. The output consists of sense annotations on content words in the corresponding *target dataset*.

More specifically, our algorithm operates under the following assumptions: (1) The *bitext* is a set of aligned sentence pairs in the source and target languages. (2) Each sentence in the target side of the bitext is a *translation* of the corresponding sentence in the source language part. (3) Each content word token in the source side of the bitext is correctly annotated with a single "gold" *sense tag*, which is represented by a synset in a multilingual wordnet. (4) Both sides of the bitext are tokenized, lemmatized, and POS-tagged. (5) The sentences on both sides of the bitext are *aligned* at the word level.

Step 1: Translation The source dataset is translated into the target language by either a machine translation system or a human expert. Modern MT models are surprisingly effective even for low-resource languages that are severely underrepresented on the web, partially thanks to their use of multilingual representations. The translation step yields a sentence-aligned bitext on which our algorithm operates (line 1).

Step 2: Tokenization We deploy automated tools to tokenize the translated target text (line 2). Since each language presents its own set of challenges for tokenizers, this step is language-dependent. High-quality tokenizers are available for many languages. If a language-specific tokenizer is not available, a general, character-based tokenization heuristic can be applied instead.

Step 3: Part of speech tagging Since word senses are specific to four different lexical categories (coarse-grained parts of speech), POS tagging is an essential preprocessing step prior to WSD. For example, the word *bank* has distinct noun and verb senses in Princeton WordNet. We consider a possible sense projection across an alignment link *only* if the two words correspond to the same lexical category (noun, verb, adjective, adverb), as shown in line 3 of Algorithm 1.

Step 4: Lemmatization Wordnets represent word senses by the inclusion of lemmas within synsets. Since inflected forms, such as plural forms of nouns, are generally not explicitly represented, the lemma of the token must be identified prior to sense tagging. We use automatic tools to identify the lemma of each content word in the target text, using the tokenization and POS tags produced in the previous steps. This step can be omitted for languages without inflected forms, such as Chinese. If a language-specific lemmatizer is not available, an unsupervised lemmatization model could be applied.

Step 5: Alignment The goal of this step is to associate each content word in the target text with a content word in the source text. We rely on unsupervised word alignment systems, keeping our method free of any dependence on language-specific alignment training data. The alignment is performed at the level of tokens, rather than words or characters.

Step 6: Filtering out non-literal translations Aligned words may not express exactly the same concept, due to translation errors, stylistic choices, or the absence of a suitable word in the target lexicon (a *lexical gap*). We consider an alignment link to correspond to a literal translation (line 4) if the aligned lemmas satisfy at least one of the following conditions: (1) they are in the *same synset* in a multilingual wordnet; (2) they are listed together as mutual *translations* in a bilingual dictionary; or (3) they appear to be *cognates*. To prevent rejection of correct projections of rare words, both *same_synset* and *translations* also return TRUE if either of the aligned lemmas are not found in the corresponding lexical resources. We consider two words to be cognates if the first three characters of each word are identical (e.g., *interest* – *intérêt*); however, more sophisticated methods for cognate detection could be used instead, including phonetic-based methods that can operate on distinct writing scripts.

Step 7: Sense Projection For each aligned word pair where the translation is deemed to be literal, we *project* the sense of the source word onto the target word. That is, we tag the target language word with the same synset as the source word. For example, if an instance of the English word token *medicine* is annotated with BabelNet synset `bn:00054128n`, and aligned with its translationally-equivalent Spanish word *medicamento*, then the Spanish token will be tagged with `bn:00054128n`, as they have the same lexical category (noun) and satisfy the cognate filter condition. This step completes the silver tagging of the target text.

3.3 Focused Gold Annotation Creation

Our next objective is to convert the projection-based silver annotation into a gold dataset suitable for evaluating WSD systems. To this end, we propose a *focused annotation* procedure, wherein native speakers correct and complete the sense tagging of the target language text, using the silver tags as a starting point.

We first apply a WSD system to tag any words that have no projected sense, e.g., due to not being aligned with a source-language token. The intuition is that correcting automatically generated translations, alignments, and annotations is faster and easier than manually annotating a text from scratch.

For each token to be annotated, we provide the annotators with the English glosses of each possible sense of the target word, as found in a multilingual wordnet, In addition to the automatically projected or generated sense tag. Annotators are instructed to decide whether the definition of the automatic sense tag accurately captures its meaning, in which case the tag is to be marked as correct. Otherwise, the annotator is asked to identify the correct sense, again by comparing the gloss of each sense against the meaning of the target word in context. If the annotators find that none of the provided glosses accurately describe the meaning of the word, they are instructed to consult the glosses of the source language word, and choose from among those.

In addition to gloss matching, we encourage the annotators to apply the following three tests for verifying the correct sense of a word in context (Hauer and Kondrak, 2020): (1) A *substitution test* verifies that the focus word can be replaced with another lemma from the same synset without changing the meaning of the sentence. (2) A *translation test* – a cross-lingual analogue of a substitution test – verifies that the translation of the focus word can be substituted by each target-language lemma from the corresponding multilingual synset. (3) Finally, an *entailment test* verifies that the focus word can be replaced with a lemma from a hypernym synset, to verify that the resulting sentence is in an entailment relation with the original sentence. The applicability of the tests may depend on the presence of multiple lemmas in a given synset. While a negative test result demonstrates that the tested sense is incorrect, the converse is not necessarily true, as multiple senses may correspond to synsets containing the same lemmas.

The procedure outlined above results in a complete sense annotation of the target text, in which each content word is annotated with a sense that is either (1) automatically generated and vetted by a native speaker of the target language, or (2) manually selected by a native speaker. In either case, a native speaker of the target language has determined that the sense tag accurately matches the meaning of the word in context. Therefore, the ultimate product is a gold-standard semantically tagged dataset in the target language, which is parallel to the source dataset.

4 Experiments

This section describes the details of our experiments, including the language-specific tools and resources used for each language, followed by our evaluation procedure, results, and analysis.

4.1 Languages

The source language in all our experiments is English due to the relative abundance of existing sense-annotated datasets. We test the generality of our method by applying it to five target languages, which represent different families, writing scripts, and levels of resource availability. Spanish (ES) and Italian (IT) are high-resource European languages written in Latin script, with existing gold sense-annotated datasets. Farsi[1] (FA), Bengali (BN), and Chinese (ZH) have their own unique writing systems, relatively poor coverage in BabelNet, and few (if any) gold datasets annotated

[1] Our method was developed on Farsi.

with BabelNet synsets. For the evaluation of these three languages, we employ newly-created parallel gold datasets created using the method described in Section 3.3.

4.2 Resources

The English dataset we use as our source for translation and projection originates from SemEval 2015 Task 13 (Moro and Navigli, 2015), henceforth **SE15**. The dataset is composed of a set of parallel sentences from OPUS (Tiedemann, 2012), in English, Spanish, and Italian (Table 1). While the English part of SE15 serves as our source dataset, the Spanish and Italian datasets serve as gold standards against which we evaluate our silver data creation procedure. For Spanish and Italian, we use the provided SE15 translations, and align the sentences with BertAlign[2].

Content word instances in SE15 are annotated with at least one BabelNet 2.5.1 synset. BabelNet incorporates semantic knowledge from many heterogeneous sources, including Princeton WordNet, Wikipedia, Wiktionary, OmegaWiki, and Wikidata. Therefore, some tokens in SE15 have multiple annotations associated with different resources, some of which may represent the same concept. For example, an instance of the word *cancer* is annotated with a BabelNet sense, a Wikipedia link, and two WordNet sense keys.

We ensure that, at most, one sense is projected from each source token. If a single word token is assigned multiple distinct senses, we select the first BabelNet sense listed as the one to be projected. We also prefer to project the sense of entire multi-word expressions rather than their individual words, where possible. For example, we sense tag *lung cancer* as a single token, rather than tagging *lung* and *cancer* separately. The intuition is that multi-word expressions may be non-compositional, and therefore convey information that can not be inferred from the individual words.

4.3 Experimental Setup

In this section, we provide the details of our experiments, including the language-specific tools and resources used for each language.

Step 1: Translation In the development phase, we experimented with several MT systems for English-to-Farsi translation including Google Cloud Translation, the Google Sheets translation function, Google Translate library, MBART50$_{MTM}$ (Tang et al., 2021), and M2M100$_{418M}$ (Fan et al., 2021). We selected Google Translate Cloud based on its performance on two English-Farsi test sets from MIZAN (Kashefi, 2020) and OPUS(GV) (Tiedemann, 2012), respectively. We used the same MT system for English-to-Chinese translation. For English-to-Bengali, we instead used BanglaNMT (Hasan et al., 2020), as it outperformed Google Translate Cloud on the RisingNews test set (Hasan et al., 2020). For Spanish and Italian, we used the translations from SE15.

	EN	**ES**	**IT**	**FA**	**ZH**	**BN**
Total Sentences	138	137	135	138	138	138
Total Tokens	2759	2973	2851	2386	2519	2283
Unique Lemmas	702	756	786	805	797	805
Total Tags	1112	1096	1097	1226	1364	904
Unique Tags	577	549	564	627	635	487
Synsets/Tag	1.3	1.6	1.5	1.0	1.0	1.0
Synsets/Lemma	2.0	2.3	2.0	1.5	1.7	1.1
Senses in BN	1.0	.920	.972	.518	.681	.145

Table 1: Statistics for the datasets. “Tags” refers to the number of sense annotated tokens. EN, ES, and IT refer to the original SE15 datasets. FA, ZH, and BN refer to the newly annotated gold datasets that we release.

Steps 2-4: Preprocessing We used language-specific tools for tokenization, POS tagging, and lemmatization, For Farsi, we perform all three tasks with **HAZM**[3]. We tokenize and POS-tag Chinese text with **HanLP** (He and Choi, 2021); lemmatization is not necessary in Chinese, as it has no inflected forms. For Bengali, we use separate tools for each task: **BNLP tokenizer** (Sarker, 2021), the POS tagger of Alam et al. (2016), and **BanglaNLP Lemmatizer** (Chakrabarty et al., 2017).

Step 5: Alignment We experimented with several word alignment programs, including FastAlign (Dyer et al., 2013), AwesomeAlign (Dou and Neubig, 2021), MultiMirror (Procopio et al., 2021), and SimAlign (Jalili Sabet et al., 2020) with either mBERT or XLM-RoBERTa. We settled on SimAlign with XLM-RoBERTa, using the intermax algorithm and 8-layer embeddings, based on feedback from native speakers, and its lack of dependence on training data. SimAlign creates contextualized embeddings of words using XLM-RoBERTa, which are then used by the Itermax algorithm to generate alignment links. If adjacent tokens are aligned to

[2] github.com/bfsujason/bertalign

[3] github.com/roshan-research/hazm

the same word, we treat them as a single token.

Step 6: Translation filtering For all languages, we used PanLex[4] for dictionary filtering. For Farsi, we added two English-Farsi dictionaries found on GitHub: VahidN[5] and 0xdolan Aryanpour[6]. When testing word synonymy, we consider two words to be synonyms if they share any synset in BabelNet.

WSD In addition to the tools mentioned above, we also use two WSD systems, AMuSE and XL-mBERT to provide additional suggestions to annotators.

AMuSE While AMuSE (Orlando et al., 2021) can perform WSD in 40 languages, the authors do not report evaluation results on individual languages other than English. AMuSE assigns senses to almost all content words in each dataset. However, AMuSE is not restricted to the BabelNet sense inventory, and can tag tokens with synsets that do not include the target word. For example, occurrences of the drug name *Cerenia* are tagged with various synsets though none of these synsets contain that word. We found the percentage of such "new senses" to be approximately 10% for the European languages, 23% for Farsi, and 36% for Chinese.

XL-mBERT We applied the pre-trained mBERT-based unsupervised WSD method previously used by Pasini et al. (2021). We used the provided sense inventories for Spanish, Italian, and Chinese. While XL-mBERT offers scalability and broad applicability, given its coverage of 18 languages, its performance varies dramatically across languages; reported F-scores for mBERT range from 42.4% for Basque to 81.6% for French. When we applied XL-mBERT in our experiments, we observed relatively low F-scores on some languages. To verify that these results were not due to methodological errors on our part, we validated our procedure for extracting sense inventories from BabelNet by replicating the extraction of the Spanish and Italian inventories provided by the XL-mBERT authors[7].

	ES	IT	FA	ZH	BN
AMuSE	70.5	68.2	27.1	40.1	–
XL-mBERT	69.8	69.0	35.6	50.5	12.4
LabelProp	68.9	72.7	–	–	–
MultiMirror	70.4	73.7	–	–	–
Projection+	63.4	61.6	**75.4**	68.4	68.4
+AMuSE	74.6	73.3	68.1	70.2	–
+XL-mBERT	**74.9**	**73.7**	75.3	**75.3**	**69.8**

Table 2: WSD F-score on the gold datasets.

4.4 Results

We evaluate sense tags a produced by our Projection+ method (Algorithm 1) against two different types of gold datasets: the Spanish and Italian SE15 benchmarks, and our new Farsi, Chinese, and Bengali annotations which have been manually verified by native speakers.[8] We also report results obtained by incorporating two WSD systems into our silver annotation procedure as a back-off to improve coverage.

We use standard WSD evaluation metrics. We calculate the F-score with the evaluation script of Raganato et al. (2017). F-score is the harmonic mean of precision and recall. Precision is the number of correct sense tags divided by the number of sense tags produced by the method, and recall is the number of correct sense tags divided by the number of instances.

We were unable to deploy either LabelProp or MultiMirror on our test sets due to their complex dependencies, and the lack of availability of required code and resources. However, both of those papers do provide mBERT-based models fine-tuned for WSD using data produced by their respective corpus tagging systems. Thus, while we cannot test these systems directly, we can test them indirectly by evaluating models trained on the semantically tagged data they produce. For simplicity, we refer to these models as LabelProp and MultiMirror.

Table 2 presents the results on five languages. On the languages covered by SE15, Spanish and Italian, the two stand-alone WSD systems, AMuSE and XL-mBERT, as well as the mBERT models trained with data created by MultiMirror and LabelProp, perform relatively well.[9] However, on the other three languages, our Projection+ method substantially outperforms the WSD systems. Furthermore, incorporating XL-mBERT as a back-off produces consistent, highly competitive results on all five languages.

[4] old.panlex.org/

[5] github.com/VahidN/EnglishToPersianDictionaries

[6] github.com/0xdolan/AryanpourDictionary

[7] sapienzanlp.github.io/xl-wsd

[8] We had four native speaker annotators for Farsi, two for Chinese, and one for Bengali; all of them are co-authors of this paper.

[9] We note that the accuracy of the WSD systems correlates well with the fraction of gold senses that are present in BabelNet (Table 1).

4.5 Ablation Analysis

To evaluate the impact of our projection filtering criteria (lines 3 and 4 in Algorithm 1), we conduct an ablation study with several variants of our method. The results are summarized in Table 3. Naive Projection, in which senses are unconditionally projected across all alignment links, serves as a baseline. Enforcing the lexical category matching (line 3 in Algorithm 1) yields a consistent improvement of 2-3% F-score across all languages. In fact, we use this configuration for non-European languages in our experiments due to their poor coverage in BabelNet. The application of All Filters, which corresponds to the complete implementation of Algorithm 1, achieves similar improvements over the baseline.

The final three lines in Table 3 show the results obtained by withholding individual filters. The cognate filter slightly improves F-score on the European languages. The BabelNet filter has no influence on the SE15 datasets, which are tagged exclusively with existing BN senses; however, its substantial impact on the other languages demonstrates our method's potential for improving BN coverage. The dictionary filters similarly improve the results on those languages, but surprisingly have detrimental effects on Spanish and Italian.

In order to explain this counter-intuitive finding, we performed a manual analysis of the 16 Spanish "errors" that are introduced by the PanLex dictionary verification. All 16 instances involve "incorrect" projections that are validated by the dictionary. We found that the SE15 Spanish gold annotations appear to be constrained by the Spanish sense inventory in BabelNet, which unavoidably causes our perfectly reasonable target projections to be classified as errors. For example, the English phrase about a medication that *"can be given"* to a patient is translated into Spanish as *"se puede administrar."* However, since the source synset `bn:00088826v` with which the token *give* is tagged does not include the Spanish lemma *administrar*, it is instead tagged with the synset `bn:00082365v`. This again shows that our projection-based approach combined with dictionary filters could be effectively used to improve existing multilingual wordnets.

4.6 Error Analysis

Our method, which is composed of several modules, produces silver sense annotations in the target language. Each module may introduce errors, which affect the accuracy of the final output. We analyzed independent random samples of 50 errors in Spanish, Italian, and Chinese. For each such instance, we manually determined and categorized the cause of the error. The numerical results of the analysis are shown in Table 4. In the remainder of this section, we discuss the principal error categories, which collectively comprise the vast majority of errors that we observe.

Configuration	ES	IT	FA	ZH	BN
Naive Projection	59.0	58.0	72.8	66.3	66.9
LexCat Filter Only	61.1	59.9	**75.6**	**68.4**	**68.4**
All Filters	62.9	61.2	74.6	67.1	66.3
w/o Cognate Filter	62.8	61.1	74.6	67.1	66.3
w/o BabelNet Filter	62.9	61.2	70.1	66.1	57.4
w/o Dictionary Filter	**63.4**	**61.6**	71.2	63.4	59.6

Table 3: The effects of filtering on F-score.

Non-literal translations While modern MT systems rarely generate clearly incorrect translations, word translations that are aligned across a bitext are not always literal. In many cases, an entire phrase, rather than its individual component words, is translated instead, sometimes expressed using a different set of concepts and/or parts of speech. In such cases, source concepts cannot be effectively projected onto the target words. For instance, the English word *delayed* in the phrase *"treatment should be delayed or discontinued"* is translated into Italian as *posticipato* "postponed". The two translations have slightly different meanings, and therefore share no synset in BabelNet.

Tokenization issues Many tokenization errors involve multi-word expressions that express a single concept. This issue is particularly salient in Chinese, where word boundaries are not indicated by white space. For example, one projection error was ultimately traced to our Chinese tokenizer segmenting a word composed of three characters with the compound meaning of "pharmacist" into two tokens with the meaning of "pharmacy" and "worker". Another type of error that we observed was caused by individual words in multi-word expressions being aligned separately to different tokens in the other language.

Missing or incorrect alignment Our method projects senses across alignment links; therefore, alignment errors can lead to sense-tagging errors. Many-to-one, one-to-many, and many-to-many

Error	ES	IT	FA	ZH	BN	Avg.
Lack of alignment	17	15	16	21	31	20.0
Wrong alignment	3	4	4	2	3	3.2
Function word	2	0	7	3	1	2.6
Named entity	0	1	4	9	4	3.6
Lexical category	7	8	8	7	2	6.4
Tokenization	2	0	0	2	0	0.8
Lack of validation	7	5	1	0	1	2.8
Sense granularity	12	14	6	1	5	7.6
Source gold error	0	3	4	5	3	3.0

Table 4: Counts of error categories in samples of 50 errors across languages.

alignments complicate sense projection. In our manual error analysis, we found that many target words receive no tag because they are not aligned to any source word, which was responsible for up to 38% of sense projection errors. However, the number of errors caused by *incorrect* alignment links was only about 7%.

Lexical category mismatch In some cases, direct translations of words across languages can have different POS tags. For instance, the Spanish adverb *más* is translated into an English adjective *more*. Similarly, a natural-sounding translation may involve translation pairs with different parts of speech. For example, the English word *dehydration* is translated into Farsi as a phrase meaning "dehydrating the body". As a consequence, aligning the English noun with the Farsi verb could lead to an incorrect concept projection. Cases such as these demonstrate the importance of the POS filter.

Sense granularity issues Since manual sense annotation is quite difficult, it is not unusual to encounter annotation errors even in gold datasets such as SE15. Source annotation errors unavoidably lead to incorrect projections, while incorrect gold-standard sense annotations on the target side lead to correct projections being evaluated as incorrect. For example, in the phrase *"advanced or metastatic non-small cell lung cancer"*, the English word *advanced* is correctly tagged as having its "advanced illness" sense; however, its Italian translation *avanzato* is tagged with a different synset, meaning "further along in time". So, while the projected sense is correct, it is spuriously reported as an error. In our manual analysis, we found that approximately 25% of the tags produced by our method that are counted as incorrect in Spanish and Italian are in fact artifacts of sense annotation errors in the SE15 datasets.

5 Conclusion

We have presented a solution to the scarcity of sense-annotated data in low-resource languages. Our method produces sense-tagged texts of consistent quality across languages by projecting senses from English WSD datasets onto automatically-generated translations, and applying a flexible set of filters. The robustness of our method is unrivaled by competing methods: it is easy to run, makes few assumptions, and establishes a novel benchmark for multilingual WSD research. The results of our experiments, along with our detailed ablation study and error analysis, demonstrate the utility of our method even in the absence of manual annotation. In particular, the automated component of our method outperforms existing multilingual WSD systems on non-European languages. We hope that our work, including the novel benchmark datasets, will support further research on global WSD.

Limitations

While our method is intended to be as broadly applicable as possible, it does depend on some assumptions, such as access to reasonably accurate machine translation and alignment systems. Some very low-resource languages may not satisfy these requirements; caution should be taken when applying our method to languages without first consulting the relevant literature on machine translation and alignment for that language, or where such literature is limited or unavailable. Similar limitations apply to highly specific domains, where otherwise high-quality tools may become unreliable.

While our method is fairly robust with respect to BabelNet coverage of low-resource languages, BabelNet undergoes frequent revisions, which may result in senses being added, removed, or changed. This may affect the reliability of the annotations produced by our method. In particular, the extremely poor performance of XL-mBERT on Bengali may be due to its low coverage in BabelNet. However, even if a target language is completely missing from BabelNet, our method is able to project source-side sense tags directly onto target-side word tokens. Thus, it implicitly creates a target language sense inventory that is grounded in English senses. Such inventory can then be expanded via manual annotation of the target text.

One limitation of our experimental setup is our use of sense annotations produced using the procedure outlined in Section 3.3 as gold standard data for our evaluation. All sense annotations in this data are validated by native speakers; however, the outputs of our projection method, as well as those of the WSD systems, AMuSE and XL-mBERT, are provided to the annotators as suggestions. Therefore, there is a possibility that these annotations may exhibit bias in favor of the suggested senses, inflating the performance of methods which incorporate these systems. However, as this bias does not specifically favor any single method, we believe that this issue does not compromise our conclusions.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Firoj Alam, Shammur Absar Chowdhury, and Sheak Rashed Haider Noori. 2016. Bidirectional LSTMs — CRFs networks for bangla POS tagging. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 377–382.

Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844. International Joint Conferences on Artificial Intelligence Organization. Main track.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*.

Abhisek Chakrabarty, Onkar Arun Pandit, and Utpal Garain. 2017. Context sensitive lemmatization using two successive bidirectional gated recurrent networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491, Vancouver, Canada. Association for Computational Linguistics.

Debapratim Das Dawn, Abhinandan Khan, Soharab Shaikh, and Rajat Pal. 2022. A dataset for evaluating bengali word sense disambiguation techniques. *Journal of Ambient Intelligence and Humanized Computing*, 14.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, Vancouver, Canada. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy = translational equivalence. *Preprint*, arXiv:2004.13886.

Bradley Hauer and Grzegorz Kondrak. 2023. One sense per translation. In *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 442–454, Nusa Dua, Bali. Association for Computational Linguistics.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021. Semi-supervised and unsupervised sense annotation via translations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 504–513, Held Online. INCOMA Ltd.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Omid Kashefi. 2020. Mizan: A large persian-english parallel corpus. *Preprint*, arXiv:1801.02107.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3514–3520, Online. Association for Computational Linguistics.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. 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)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

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.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tommaso Pasini and Roberto Navigli. 2017. Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88, Copenhagen, Denmark. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):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.

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, Valencia, Spain. Association for Computational Linguistics.

K.M. Tahsin Hassan Rahit, Khandaker Tabin Hasan, Md. Al Amin, and Zahiduddin Ahmed. 2018. BanglaNet: Towards a WordNet for Bengali language. In *Proceedings of the 9th Global Wordnet Conference*, pages 1–9, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Hossein Rouhizadeh, Mehrnoush Shamsfard, and Vahide Tajalli. 2022. Sbu-wsd-corpus: A sense annotated corpus for persian all-words word sense disambiguation. *International Journal of Web Research*, 5(2):77–85.

Sagor Sarker. 2021. BNLP: Natural language processing toolkit for Bengali language. *Preprint*, arXiv:2102.00405.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "OneSeC" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709, Florence, Italy. Association for Computational Linguistics.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaei, Niloofar Mansoory, Ali Reza Gholi Famian, and Somayeh Bagherbeigi. 2010. Semi Automatic Development Of FarsNet: The Persian Wordnet.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Appendix

The appendix includes four tables. Table 5 shows three examples of information provided to annotators for manual annotation. Table 6 contains annotator instructions for the gold annotation procedure. Table 7 explains how to apply the three tests for verifying the correct sense of a word in context. Table 8 defines the categories of errors in the silver annotation output.

ZH Token	药物	输入	工人
Token ID	d001.s010.t066	d002.s051.t004	d003.s011.t030
ZH PoS	NN	VV	NN
ZH Context	另一种抗癌药物	如果您已经输入该功能请单击“确定”按钮	改善工作地点和社区中老龄工人的招聘
Proj. PoS	N	V	N
Proj. Context	Another anticancer medicine.	If you have entered the function click on the OK button.	improve the recruitment, training and development of ageing workers
Proj. Sense	bn:00054128n	bn:00089626v	bn:00081593n
Proj. Gloss	(Medicine) Something that treats or prevents or alleviates the symptoms of disease.	Bring in from abroad.	A person who works at a specific occupation.
AMuSE Sense	bn:00028872n	bn:00089626v	bn:00081593n
XL-WSD Sense	bn:00054128n	bn:00089626v	bn:00081593n
Potential Senses	bn:00054128n | bn:00028872n | bn:00053800n | bn:00061887n	bn:00089626v | bn:00089764v	bn:04801100n | bn:00047795n | bn:00081593n | ...
Senses Glosses	(Medicine) Something that treats or prevents or alleviates the symptoms of disease. | A substance that is used as a medicine | ...	Bring in from abroad | Enter (data or a program) into a computer	Person who works | Someone who works with their hands | A person who works at a specific occupation | A person who works at a specific occupation | ...
Substitution Test	bn:00054128n: 药品, 药, 药材, ...	bn:00089626v: 进口, 入, 口	bn:00081593n: 劳动者, 劳工, 工作者
Translation Test	bn:00054128n: medication, medicine, medicament, ...	bn:00089626v: import	bn:00081593n: worker
Hypernymy Test	bn:00054128n: 药物	bn:00089626v: 买卖	bn:00081593n: 人

Table 5: Examples of information provided to annotators for manual (gold) annotation.

Instructions for Gold Annotations

For each marked word pair in the automatically generated target dataset, your task is to verify or fix the target translation, POS tags, tokenization, and/or alignment of the target to the English. Next you must **find a single BabelNet synset for each target content token** and enter it in the column titled "*BN Synset*".
Most target tokens have already been annotated with a synset via automatic projection. You must verify whether the projected sense is correct or in the case that the projected sense is not correct, **find the correct synset for the token.**
The suggested approach consists of the following steps:

1. You should first examine the existing BN synsets and glosses in the columns titled "*Potential Synsets*" and "*Synset Glosses*". The idea is to choose the BN synset whose meaning matches the meaning of the target token.

2. Many of the glosses may seem very similar and difficult to differentiate between. We have also provided columns which include information to perform the three tests: *Substitution*, *Translation*, and *Entailment (Hypernymy)* test. The details of these tests are given in the document titled "*Instructions for Tests*".

3. If the sense is not there either, try to find the synset directly in BabelNet. Search for the right concept by looking up the synonyms in English which are translationally equivalent to the target word.

4. There may be circumstances where the token is translated, aligned and given the correct POS tag, but there is no correct BN sense tag. For this we want you to search for the token in WordNet and provide the correct WN sense tag.

5. In the event that a word really cannot be annotated, please add a note explaining why that is. We do not expect this to happen, except in a few strange cases.

Table 6: General instructions given to the annotators to facilitate the gold annotation procedure. The mentioned "Instructions for Tests" can be found in Table 7.

Instructions for the Three Tests

Sometimes when assigning a sense to a token, there are many different synsets to choose from. We have devised 3 tests to help disambiguate these similar synsets:

(1) Substitution test,
(2) Translation test, and the
(3) Entailment (Hypernymy) test.

Substitution Test
In the given sheet, there is a column titled "Substitution" with a list of synsets that contain the target token. For each given synset, there is a list of other lemmas in the synset in the target language. These are to be used for the Substitution Test: taking one token at a time, replace the original token with another synset lemma in the sentence. If this substitution changes the meaning of the sentence, this is unlikely to be the correct synset for the token.
For example, consider the word "***contains***" in the sentence "It ***contains*** many hit songs from this year". For this token we get this list of possible synsets including one with the lemma "***incorporate***" and one with the lemma "***control***". We then can substitute the given lemma in place of the original token (contains):
It ***incorporates*** many hit songs from this year. || It ***controls*** many hit songs from this year.
The lemma "***incorporate***" does not change the meaning of the original sentence, so it may be the correct synset. However, the lemma "***control***" from the second synset does change the meaning of the original sentence, and therefore the corresponding synset is not correct.

Translation Test
In the given sheet, there is a column titled "Translation" with a list of synsets that contain the target token. Additionally, there is a list of English tokens for each synset. These are to be used for the Translation Test: Given your knowledge of the target language and English, decide whether each English lemma is a translation of the given target token in the given context. If it is an incorrect translation, this is unlikely to be the correct synset for this target token.
For example, consider the token "***domaine***" in the sentence "J'étudie dans le ***domaine*** de l'informatique." (I study in the field of computer science.). For this token, we get a list of possible synsets, including one with the English lemmas "***discipline***" and one with the lemma "***plain***". Based on the context here, the word "***plain***" is not a correct English translation as its noun form refers to an "area of land" where "***domaine***" refers to an "area of study". Thus, the more likely synset is the one containing the lemma "***discipline***".
(Chinese version) For example, consider the token "生存" in the sentence "使用Alimta治疗后平均生存期为8.3个月" (The average survival after treatment with Alimta is 8.3 months.). For this token, we get a list of possible synsets, including one with the English lemma "***survival***" and one with "***life***". Based on the context here, "***survival***" best captures the meaning, as "生存" refers specifically to the duration of survival after treatment where "***life***" refers to existence or general life. Thus, the more likely synset is the one containing the lemma "***survival***".

Entailment (Hypernymy) Test
In the given sheet, there is a column titled "Hypernymy Test" which contains a list of synsets for the target token and the hypernym for that synset. The hypernym of a concept is a more general term for the concept. For example, the hypernym of "apple" is "fruit". These are to be used for the Entailment Test: Replace the original target token in the sentence with the given hypernym. If the modified sentence is implied by the original sentence, the synset passes the test. Otherwise, the corresponding synset is likely incorrect.
For example, consider the word "***cards***" in the sentence "I received many ***cards*** for my birthday". For this token, we get this list of possible synsets including one with the hypernym "***correspondence***" and one with the hypernym "***paper***". We then can replace the original token (cards) with the given hypernyms.
I received many ***correspondences*** for my birthday. || I received many ***papers*** for my birthday.
The new sentence created by replacing "***cards***" with the hypernym "***correspondences***" is implied by the original sentence. Thus this is a plausible synset for the original token. However, the new sentence created by replacing "***papers***" with the hypernym "***correspondences***" is not implied by the original sentence as "***papers***" (such as newspapers, or sheets of paper) are not typically something you receive for a birthday. This synset is unlikely to be correct.

Table 7: Instructions given to the annotators for the Substitution, Translation, and Entailment (Hypernymy) tests.

Instructions for Labeling Types of Errors

In some cases, a word requiring annotation either has no sense tag suggested, or the suggested tag is incorrect. As part of your annotation work, we ask you to provide further analysis on the nature of these errors and omissions. Below is a list of categories of errors you may encounter, each associated with a 2-3 letter code. Error labels have been automatically generated in some cases, but may be overly general or incorrect. **Your task is to analyze 50 errors we have selected** from the silver dataset and **assign each to one of the following nine categories:**

Lack of Alignment (LAE): There is no English token aligned to the target.
Alignment Error (AE): The target token was aligned with an English token, but the aligned tokens have completely unrelated meanings.
Function Word (FW): The target token is a function word that was incorrectly assigned a sense.
Named Entity (NE): The target token is a named entity that was assigned a sense in the target but not the source.
Lexical Category Mismatch (LCM): The target token and source token belong to different lexical categories (different POS tags).
Tokenization Error (TK): Tokenizers sometimes fail to recognize multi-word tokens, such as "make up"; this may lead to words with a multi-word token being aligned separately.
Projected Annotation Mismatch (PAM): The target token is similar in meaning to its aligned English token so the silver target sense is tentatively projected from the English. However, the sense has been removed by our filter due to lack of validation.
Sense Annotation Mismatch (SAM): The projected target sense and the gold target sense are related but not identical.
Source Sense Annotation Error (SSA): The source sense tag is wrong or missing.

Table 8: Instructions for labeling the types of errors seen in the silver output.