Towards Sense-level Bilingual Dictionary Induction

Lydia Körber[1] **Katja Markert**[1] **Wei Zhao**[2]
[1]Heidelberg University [2]University of Aberdeen
{lastname}@cl.uni-heidelberg.de
wei.zhao@abdn.ac.uk

Abstract

Updating bilingual dictionary entries is a tedious, time-consuming, and highly subjective task, especially when a new sense in the source language requires identifying an appropriate translation equivalent. To date, there have been no attempts to automatize the discovery of new bilingual sense entries. Related tasks such as Word-level Bilingual Dictionary Induction and cross-lingual embedding alignment do not account for polysemy and are not applied to lexicographic data. In contrast to their monolingual counterparts, bilingual dictionaries fall short in terms of completeness, detail with respect to examples and glosses, and diachronic information. We introduce a novel NLP task, Sense-Level Bilingual Dictionary Induction (SENSEBDI), at the intersection of lexical semantics, cross-lingual, and diachronic NLP. We construct a dataset of time-stamped sense-level bilingual dictionary entries by aligning two bilingual dictionaries, two monolingual dictionaries, and the multilingual resource BabelNet, thereby enriching bilingual entries with monolingual source-language information. We propose a baseline based on nearest-neighbor search over cross-lingual embeddings of glosses and usages. We find that usages contribute more strongly than glosses, with substantial variation across language pairs and discuss task-specific challenges with regards to target language polysemy and future directions such as transfer to real-world scenarios. The code for reproducing the dataset construction is available on GitHub: https://github.com/lykoerber/towards-sense-bdi.

1 Introduction

Change is inherent to language. Like all linguistic resources, bilingual dictionaries need to be updated frequently. A bilingual dictionary can even be seen "as a mirror of language change" (Yermolovich, 2002), with societal changes mirrored in linguistic changes manifested in bilingual dictionaries.

The lexicographic workflow of drafting new bilingual entries is as follows: The basis for new bilingual entries are usually new words, senses, and phrases that are added to a monolingual dictionary in the source language, the reference work. Next, language professionals who are native speakers of the target language search translation equivalents for the new items in different resources in the target language, such as corpora and monolingual dictionaries. Once a suitable equivalent is found, the new bilingual entry is drafted.

This manual process poses several issues: It is time-consuming (Lorentzen and Trap-Jensen, 2016), tedious, and, like many natural language tasks, highly subjective[1]. In addition to the general difficulties of keeping dictionaries up-to-date, bilingual lexicography faces particular diachronic challenges: While new words and senses are fastly translated in multiple ways, it can take some time until a consensus is found and one equivalent is established; often, several translation options coexist in parallel (Williams, 1959).

NLP methods can help automatize this workflow of finding translation equivalents of new senses.

The task of automatically inducing translations, i.e., Word-level Bilingual Dictionary Induction (BDI), is well established in NLP (Denisová and Rychlý, 2024). However, it has been subject to different points of criticism (Kementchedjhieva et al., 2018, 2019; Søgaard et al., 2018): Benchmark datasets such as MUSE and XLING (Conneau et al., 2018; Glavaš et al., 2019) are often created using machine translation, which does not ensure that all possible senses and their translations are captured. They have been shown to contain a large portion of proper nouns, which are less informative for evaluating actual system performance and may lead to overestimation. Word-BDI outputs

[1]See Iamartino (2017) for a more detailed discussion about lexicographers' biases.

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 275–289

usually come in form of aligned source-target word lists, with no information on POS and senses. Inflected as well as base forms are included, which poses further challenges to evaluation. Denisová and Rychlý (2024) highlight the importance of incorporating such lexicographic knowledge.

Word-BDI does not take polysemy into account. Polysemous headwords like *marketplace* acquire several meanings over time.[2] An example of a WordBDI entry of this headword for the language pair English-German, here in the MUSE dataset (Conneau et al., 2018)[3], is a set of translation equivalents *{marktplatz, markt}*. When a new meaning, such as *the arena of commercial dealings* (see Table 7), emerges, the WordBDI output does not provide information about which of these equivalents is appropriate in this new meaning. A sense-level task, in contrast, can give insight into which translation equivalent to use for a specific sense.

We therefore introduce a new task, SENSEBDI, that performs BDI on a sense-level instead and finds translation equivalent word senses. It differs from Word-BDI in the following: The basis of our dataset are expert-curated dictionaries. This comes with several advantages:

a. This is a **sense-level** task: We exploit dictionary sense entries for a better coverage of polysemous words.

b. Sense entries contain detailed linguistic information: POS tags, glosses, usages, and in some cases even additional information on domain (e.g. *Physics*), grammar (e.g. *transitive*), register (e.g. *informal*), or place (e.g. *British English*).

c. Headwords are in the base form, inflected variants are not considered. This makes them easier to compare across resources and eliminates the need for lemmatization in the evaluation.

We develop new bilingual entries, making this a diachronic task: Given a word in the source language that developed a new meaning, the aim is to find the translation equivalent in the target language. In contrast to Word-BDI, we do not use large parallel or comparable corpora, but sense-annotated sentences with sense-specific usages. The application is the field of lexicography, with the aim of automating the lexicographic workflow. It further differs from existing bilingual dictionaries in the availability of timestamps, extended usages and glosses on the source language side.

We aim to answer the following research questions:

RQ1: How can we set up the SENSEBDI task and construct a dataset with sense-level, bilingual, and temporal information for each instance (§3)?

RQ2: Which components of bilingual dictionary entries and other lexical resources are most effective to solve SENSEBDI (§4)?

We introduce the new task of sense-level bilingual dictionary induction (SENSEBDI). We create a benchmark dataset with bilingual sense entries extracted from two bilingual online dictionaries, Cambridge English-Chinese (Simplified) and Collins English-German. In order to enrich these bilingual entries with further sense-level monolingual information such as glosses, usages, and timestamps, they are connected to a diachronic (OED) and a further monolingual dictionary (ODE), as well as to a cross-lingual semantic network, BabelNet (Navigli and Ponzetto, 2010), via a large multilingual sense-annotated dataset, XL-WSD (Pasini et al., 2021). We also propose a baseline approach to solve this problem, based on Nearest-Neighbor Search of embeddings of a cross-lingual word-in-context model, XL-LEXEME (Cassotti et al., 2023), and multilingual SentenceBERT models (Reimers and Gurevych, 2020).

2 Related Work

The relationship of lexicography and NLP is shaped by bi-directional influence: The field of electronic lexicography benefits from advancements in NLP methodologies used for computer-assisted dictionary making such as automatic compilation of lexicographic information like usages and collocations (Klosa-Kückelhaus and Tiberius, 2024). Lexicographic data, on the other hand, provide a valuable resource for the NLP community in terms of high-quality hand-crafted word sense relations, definitions and usages. These are leveraged for lexical semantics tasks such as Definition Modeling and Reverse Dictionary (Mickus et al., 2022), the tasks of generating embeddings from definitions, and definitions from embeddings. Also Word Sense Disambiguation and Induction can be used for lexicographic purposes: Škvorc and Robnik-Šikonja

[2] see https://www.oed.com/dictionary/marketplace_n?tab=meaning_and_use

[3] https://dl.fbaipublicfiles.com/arrival/dictionaries/en-de.txt

Dictionary	Languages	Diachrony	Format	Sense Depth	HS	HPS	Examples
Cambridge	en-zh, (en-ru)	synch.	XML	1	1.7 (1.73)	1.41 (1.54)	bl (ml)
Collins	en-de	synch.	XML	1-2	2.54	2.22	bl
ODE	en	synch.	HTML	1-2	3.95	2.87	ml
OED	en	diach.	HTML	1-4	10.34	6.66	ml, τ

Table 1: Overview over mono- and bilingual dictionaries used. HS describes the average number of senses per headword, HPS is the average number of senses per headword grouped by POS. These are computed based on the original sample. Examples are classified by bilingual (bl), monolingual (ml), and time-stamped (τ).

(2025) investigate leveraging dictionary examples from monolingual dictionaries of Slovene for solving word-sense tasks.

2.1 Lexicographic Perspective

Bilingual lexicography can be seen as a bridge between linguistics and translatology, the theory and practice of conveying meaning across languages (Adamska-Sałaciak, 2010, p. 389). Bilingual dictionaries vary in form and function depending on the intended user groups and purposes (Hausmann and Werner, 1991). Adamska-Sałaciak (2022) identifies three main reasons for issues in compiling bilingual dictionaries: vagueness of meaning, polysemy, and anisomorphism, which describes the lack of 1-to-1 correspondence between translation equivalents. The difference between vagueness, i.e. one broad, general meaning, and polysemy, i.e. several distinct meanings, is not clear-cut.

2.2 NLP Perspective

SENSEBDI is located at the intersection of three broad areas of NLP research (Cross-lingual NLP, Lexical Semantics, and Diachronic Language Modelling), since it combines linking of time-stamped word senses across languages. The majority of BDI approaches do not take polysemy into account (Denisová and Rychlý, 2024). However, there are a few exceptions: Early approaches include Varga and Yokoyama (2009) who applied a pivot-based method via intermediate language using WordNet and Kaji (2003) who assigned word meaning from a bilingual comparable corpus and a bilingual dictionary by clustering translation equivalents. Fišer et al. (2012) address polysemy in bilingual lexicon extraction from comparable corpora by combining several word sense disambiguation systems. More recently, Ding et al. (2025) proposed filtering polysemous words with multiple translations from seed lexicons. In sum, these use sense information to improve word-level BDI, but do not create a sense-level mapping.

Sense- and translation-related tasks intersect on different levels and can benefit from the information of each other: Translations have been integrated into WSD approaches, following the idea that different translations indicate different senses of a word (Resnik and Yarowsky, 1997; Luan et al., 2020; Kang et al., 2024). In the opposite direction, Hauer et al. (2021) improve sense annotation with translation.

3 SENSEBDI

3.1 Dictionaries & Resources

The basis of our dataset are two bilingual dictionaries, both with English as a source language[4]. The Cambridge Advanced Learner's English–Chinese Dictionary (**cam-en-zh**)[5] is recommended for intermediate to advanced learners of English. Examples are based on the Cambridge English Corpus.

The Collins Unabridged German to English and English to German online dictionary (**col-en-de**)[6] is aimed at both English and German learners.

We further include two monolingual dictionaries of the source language: The Oxford Dictionary of English (**ODE**) is a synchronic dictionary of English with fine-grained sense distinctions. Currently, it is printed in the third edition (Stevenson, 2010), which builds upon the earlier New Oxford Dictionary of English (NODE) (Pearsall and Hanks, 1998), and a second edition (Soanes and Stevenson, 2003). It is particularly useful for providing numerous usages for each sense and subsense. Senses are clustered into main senses (prototypical, core) with more specific subsenses which build upon them (Lew, 2022). Often, the core sense depicts the most general, literal use, whereas subsenses extend the core sense in specific and figurative usage (Van der

[4]The target languages were chosen upon availability in the resources and knowledge of the authors.

[5]https://dictionary.cambridge.org/dictionary/english-chinese-simplified/

[6]https://www.collinsdictionary.com/dictionary/english-german

Meer, 2000).

The Oxford English Dictionary (**OED**), a diachronic dictionary, aims to describe the English vocabulary since 1150 AD (Benbow, 1987). Its edition and publication history dates back to the second half of the 19th century[7]: Originally, it was published in fascicles between 1884 and 1928. Currently, an updated third print edition is in preparation. Sense structures in an entry are based on chronology, in order to "show the evolution of meanings and uses over time"[8]. It is considered to have the most detailed sense definitions of all English dictionaries (Landau, 2001).

For target language data, we retrieve glosses from BabelNet (**BN**)(Navigli and Ponzetto, 2010; Navigli et al., 2021), a semantic network and lexical database enriched with encyclopedic and multimodal knowledge by connecting multilingual synonym sets to Wikipedia, Wiktionary, and Wikidata, and usages from **XL-WSD** (Pasini et al., 2021), a multilingual dataset for Word Sense Disambiguation, annotated with BabelNet senses.

Table 1 gives an overview of the mono- and bilingual dictionaries used. Sense hierarchy depth differs across dictionaries: Cambridge bilingual dictionaries usually describe senses of depth one only, whereas OED entries structure homonyms and senses across up to four levels into overarching divisions, semantic branches, sections, and subsections (Lew, 2022).

3.2 Dataset Construction

We extracted sense-level entries and usages from different sources into a unified format. The pipeline is depicted in Figure 1: We randomly sampled 1000 English headwords[9] (1) from the English headword list of Wiktionary[10], a large collaborative online dictionary. Since the coverage of Wiktionary is considerably larger than that of the dictionaries used, the condition for headword selection was that a headword is present in Cambridge bilingual dictionaries for the language pairs `en-ru`[11] and

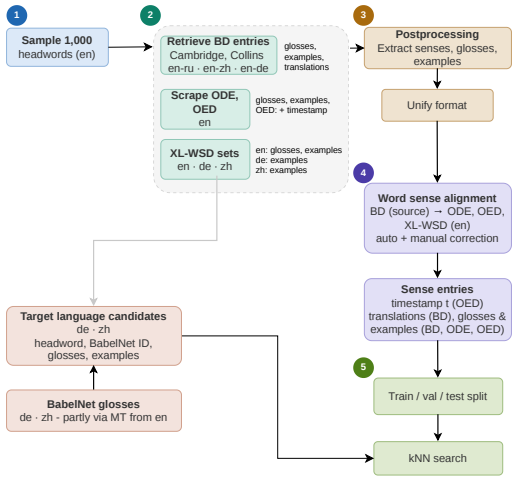

Figure 1: Dataset construction pipeline, colors indicate the four stages: (1) headword sampling, (2) dictionary entry retrieval, (3) postprocessing and unification, (4) word sense alignment, and (5) data splitting.

`en-zh`.[12] We then retrieved all dictionary entries from the respective APIs (2), including examples if available. In a subsequent step, we extracted dataset instances from raw XML and HTML files into a unified format (3): one instance per sense, including headword, POS, and optionally a gloss, other information, and examples. Senses were then linked across dictionaries (4): Each bilingual entry is enriched with usages in source and target language, as well as an English timestamp retrieved from OED. The Word Sense Alignment approach is described in Appendix A. The different components of a sense entry and the source dictionary of each are listed in Table 2.

Component	Source
headword w^s, pos p	all (originally BD)
glosses G	BD, ODE, OED, XL-WSD
usages U^s	BD, ODE, OED, XL-WSD
usages U^t	XL-WSD
usages $U^{s,t}$	BD
translations T	BD
timestamp τ	OED

Table 2: Dataset components and their sources: Bilingual Dictionary (BD) refers to `cam-en-zh` or `col-en-de`, respectively. U^s comprises usages in the source, U^t usages in the target language. In some cases, the bilingual dictionaries provide bilingual examples $U^{s,t}$. OED usages U^s_{OED} are time-stamped as well.

[7] https://www.oed.com/information/about-the-oed/history-of-the-oed/oed-editions/?tl=true

[8] https://www.oed.com/information/understanding-entries/oed-terminology/

[9] We report POS statistics in Table 5.

[10] https://en.wiktionary.org/w/api.php?action=query&list=categorymembers&cmtitle=Category:English_lemmas&cmlimit=max&format=json, retrieved on 2024-11-18 11:25am CET.

[11] The `en-ru` language pair was excluded at a later point due to limited availability of suitable sense-annontated target language data.

[12] Other dictionaries (Collins, ODE, OED) were only included at a later point. In a potential extension of the dataset, availability in other dictionaries could be incorporated as additional sampling conditions.

A running example of a dataset instance is shown in Appendix C. We split the dataset into training and evaluation sets by time stamp (5). Details and statistics are reported in Appendix B. The small dataset size is due to the limited availability of bilingual sense-level data and a result of the different filtering steps.

3.3 Notation

A bilingual dictionary BD maps words of the same meaning in the source language to the target language. Depending on the resource, this mapping relation can operate on a headword, or sense level - mapping words w from the dictionary's vocabulary $\mathcal{W}_{BD}$ or senses σ from its sense inventory $\mathcal{S}_{BD}$, on the source and target language side, illustrated in Equations (1a)-(1d)

$$BD_{\text{ww}} : \mathcal{W}^s_{BD} \rightarrow \mathcal{W}^t_{BD} \quad \text{(1a)}$$
$$BD_{\text{w}\sigma} : \mathcal{W}^s_{BD} \rightarrow \mathcal{S}^t_{BD} \quad \text{(1b)}$$
$$BD_{\sigma\text{w}} : \mathcal{S}^s_{BD} \rightarrow \mathcal{W}^t_{BD} \quad \text{(1c)}$$
$$BD_{\sigma\sigma} : \mathcal{S}^s_{BD} \rightarrow \mathcal{S}^t_{BD} \quad \text{(1d)}$$

4 Experimental Setup

Our baseline approach performs Nearest-Neighbor Search on cross-lingual contextual embeddings to find a target translation for a given source sense. The pipeline is depicted in Figure 2. For each dictionary sense entry x in the dataset portion, we create a sense embedding by averaging gloss and usage embeddings from different sources. On the target language side, likewise, candidate sense embeddings consist of averaged embeddings of XL-WSD examples and BN glosses for each ⟨headword, synset⟩ pair. We then perform k Nearest-Neighbor Search based on cosine similarity to predict a translation for the given word sense.

4.1 Cross-lingual Models

We experiment with three cross-lingual models for generating token- and sentence-level sense embeddings:

XLL XL-LEXEME (**XLL**) is the current state-of-the-art model for Lexical Semantic Change Detection (Cassotti et al., 2023), a crosslingual Siamese Network based on XLM-RoBERTa (Conneau et al., 2020), fine-tuned on WiC datasets in different languages: MCL-WiC (Martelli et al., 2021), XL-WiC (Raganato et al., 2020), and AM^{2}ICO (Liu et al., 2021).

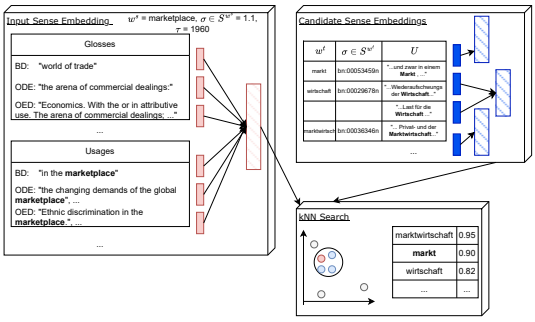

Figure 2: Pipeline of the baseline approach with an example sense entry: Colors indicate which components regard the source language (en, red), target languages (de zh, blue).

SBP is a multilingual version of `paraphrase-MiniLM-L12-v2` (Reimers and Gurevych, 2019).[13] It is based on the MINILM (Wang et al., 2020) distillation framework, and trained on paraphrase pairs to capture meaning across different phrasings.

SBD is a distilled version of the multilingual Universal Sentence Encoder (mUSE) (Yang et al., 2020) based on the DistilBERT (Sanh et al., 2020) distillation framework.[14] Unlike the other two models, it uses BERT's WordPiece tokenization (Devlin et al., 2019).

Candidates in the target languages German and Simplified Chinese are extracted from XL-WSD: For each sense-annotated headword in XL-WSD, one sense embedding per headword and BN-ID is created by averaging over all examples, statistics are shown in the Appendix 8.

We extract target language glosses for all candidates via the BabelNet API. Once again, we encounter problems of data sparsity: The availability of target glosses varies strongly, with many synsets missing a zh or de gloss. We therefore create silver data via machine translation: We retrieve English glosses additionally and use Google Translate[15] to automatically translate English glosses from Ba-

[13] `paraphrase-multilingual-MiniLM-L12-v2` https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[14] `distiluse-base-multilingual-cased-v1` https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[15] https://translate.google.com/

belNet into the target languages.[16] We use the Wikipedia gloss, if available. Otherwise, we average all other available glosses, sources include different WordNet projects, as well as Wikidata, Wiktionary, and Wikipedia.

4.2 Evaluation Metrics

BDI is typically evaluated as a ranking task in order to take the ordering of output translations into account, using many metrics inspired by Information Retrieval (IR). The common metric to describe system performance is Precision@k (P@k, 3). An instance is counted as positive if at least one true translation appears within the top k predictions. The common values reported are $k = \{1, 5, 10\}$ (Sharoff et al., 2023). Since P@k does not take the outputs below k into account, and all systems ranking the correct translation below k are assigned the same scores, Glavaš et al. (2019); Laville et al. (2022) argue for reporting Mean Average Precision (MAP, 6) instead. Since not all gold translations T of the sense entries appear among the candidates due to data sparsity, we also report an adjusted MAP score, only considering the gold translations that can be actually found. We further report Mean Reciprocal Rank (MRR, 7). Detailed formulae are noted in Appendix E.

5 Experimental Results

5.1 Main Results

The results of the base configurations on the evaluation set are reported in Table 3. XLL proves to be the best choice for the en-zh LP, whereas for the en-de LP, higher results are achieved with the SBERT models. SBD has the disadvantage of having a less language-agnostic tokenization approach and mapping some Chinese tokens to [UNK], but performs well for the en-de language pair.

Embedding Level. For usage embeddings, we experimented with token and sentence level. Since XLL is fine-tuned for the WiC task, the target token embedding has a major influence on the overall sentence embedding. Using token instead of sentence embeddings only leads to minor improvements in en-de and a minor decline in en-zh in MRR and MAP, as observed in Table 3 when comparing the configurations with usage embeddings U on sentence- and token-level (*). The multilingual SBERT models (SBP, SBD), on the other hand, show strong performance improvements on token-level embeddings across all configurations.

Language Pairs. Across all models and combinations, results in en-zh are lower than in en-de: One possible reason for this are the typological differences, with en-de being closely typologically related. A further explanation is the possible effect of unbalanced candidate sizes on the target size: As depicted in Table 8, also for the shared POS (NOUN), the candidate set size for zh is considerably larger than for de, which makes the task harder to solve from a statistic point of view. Further differences between the pairs regard the underlying dictionary data, the Cambridge entries providing more and better glosses than Collins, and the Collins entries displaying more parallel polysemy. Also, the domain differences in the XL-WSD portions should be noted, with zh sentences stemming from the environmental domain, and de from news data.

5.2 Error Analysis

We conduct a manual error analysis and specifically investigate hard instances where the first correct translation appeared at a very low rank in the prediction. We identify three common error patterns and sources, with examples listed in Table 4.

Sense mismatch. Some errors can be traced back to the sense-word BDI setup which lacks sense disambiguation on the target language side. This concerns wrong predictions where the gold translation headword appears in a different sense in the candidates (A): An example from the training portion is the translation pair *square*, *Platz*, where the sense entry regards the *town square* sense, whereas the XL-WSD candidate example for *Platz* is the sense *rank/position*: *...kommt damit auf den ersten Platz innerhalb seiner Gruppe...* (...thereby attains the first place within its group....). This is reflected in the predictions, with the gold translation appearing at low ranks and with low confidence scores overall. The top 3 predictions here, *wall_street, viertel, demokratie* (Wall Street, quarter, democracy), relate to the *town square* sense. This underlines the need for polysemy disambiguation on the target side, addressed in 6.2.

Domain mismatch. Other cases of wrong predictions can be traced back to domain mismatch

[16] We manually inspected the de translations, which overall are of high quality. Nevertheless, we acknowledge that this synthetic data creation step is prone to errors.

Configuration			en-zh						en-de					
Model	U	G	P@1	P@5	P@10	MRR	MAP	MAP_{adj}	P@1	P@5	P@10	MRR	MAP	MAP_{adj}
XLL	✓	✓	16.7	25.0	29.2	20.7	9.0	20.6	14.3	21.4	21.4	19.3	19.3	19.3
		✓	8.3	20.8	29.2	15.3	7.6	14.9	21.4	35.7	35.7	29.6	29.6	29.6
	✓		**37.5**	45.8	50.0	**42.8**	**18.7**	**42.3**	35.7	42.9	50.0	42.1	38.0	41.6
	✓*		29.2	**50.0**	50.0	39.0	17.0	38.0	35.7	42.9	57.1	42.9	38.8	42.3
SBD	✓	✓	-	-	-	-	-	-	**50.0**	**57.1**	**71.4**	**56.3**	**49.2**	**52.8**
		✓	20.8	41.7	45.8	29.5	14.9	29.3	35.7	**57.1**	64.3	46.8	39.8	43.4
	✓		-	-	-	-	-	-	14.3	35.7	42.9	23.6	20.1	23.6
	✓*		-	-	-	-	-	-	35.7	50.0	50.0	44.9	41.1	44.6
SBP	✓	✓	16.7	**50.0**	**58.3**	27.8	13.8	27.6	35.7	50.0	64.3	45.3	38.2	41.7
		✓	16.7	33.3	**58.3**	25.7	12.5	25.7	42.9	**57.1**	64.3	50.6	43.5	47.1
	✓		4.2	20.8	33.3	13.0	6.8	13.0	14.3	21.4	21.4	18.7	16.9	18.7
	✓*		16.7	37.5	41.7	23.8	11.4	23.8	35.7	50.0	50.0	43.5	40.0	43.5

Table 3: Results of the baseline approach on the evaluation set. Columns U and G show whether usage, or gloss embeddings, or both, were used. * indicates target token embeddings of usages. Scores are shown $\times 100$.

Ex.	w^s	T	configuration	top-3 predictions	rank	description
A	square	Platz	SBP-U (token)	wall_street, viertel, demokratie	198	sense mismatch
B	call	Entscheidung	SBD-UG	spiel, sport, fußball	33	domain mismatch
C	side	Seite	SBD-U (sent)	linke, thema, konsens	68	low $\|U\|$

Table 4: Example error patterns with outputs in selected configurations. Rank refers to the rank of the first correct prediction. The top-3 predicted headwords are listed. Low $|U|$ refers to a low number of sentences for a candidate embedding.

between XL-WSD and dictionary data (B): An example in the German portion is the noun translation pair *call*, *Entscheidung* which on the source language side contains mostly examples from the sports domain, on the target language side from law and jurisdiction. This is especially visible in SBERT predictions, such as in SBD-UG with candidates related to sports predicted. However, this observation does not hold in all configurations, with the correct translation retrieved at high ranks in SBP-U and SBP-G. A related sense exists in the en-zh portion with the target translations {决定, 抉择}, also with the meaning of a general decision in BD, sports-specific in other dictionaries. But the performances are much better, and a correct translation appears among the top 10 predictions in most cases, very often even at rank 1, especially in usage-only configuration.

Low number of sentences. An interesting effect of the low number of usages (C) in some candidates is that sometimes, in a sentence-level embedding configuration, the predictions can be traced back to specific sentences. An example of this is the translation pair *side*, *Seite* in the sense of the *position or attitude of a group or party*. Here, the wrong prediction in SBD-U *linke, thema, konsens* (left, topic, consensus) can be traced back to a specific sentence in XL-WSD: *Über dieses Thema herrscht nämlich ein weitreichender Konsens zwischen Rechten und Linken* . (There is a broad consensus on this topic between the right and the left.). From this sentence, several candidates are derived that appear in the prediction and that each have a very low number of usages: *linke* (1), *thema* (2), *konsens* (2). So the sentence mentioned strongly contributes to these candidate embeddings. This sentence indeed broadly covers the sense of the input entry (*position or attitude of a group or party*) by mentioning a consensus between political parties, but does not explicitly contain the gold translation *Seite* (*side)*. In this example, the token-level embedding instead yields the correct prediction at rank 1, which underlines the improvement with contextual target word embeddings.

Overall, we observe strong differences across instances: for some instances, including glosses is very useful, for others not. The heterogeneity of lexicographic data and lexical semantic resources is visible in our dataset too. The differences across models and configurations suggest that a combination of different approaches might be useful.

6 Discussion

6.1 Research Questions

We first answer the research questions: RQ 1 was devoted to setting up a BDI task on a sense-level, and construct an appropriate dataset. This objective succeeded partially, with mapping from a sense-level on the source language side to a word-level on the target language side. We address these limitations and possibilities to achieve a fully sense-level mapping in more detail in 6.2 below.

RQ 2 asked which components and information can be useful to solve SENSEBDI. We have seen in Section 5.1 that the results of our baseline approach differ strongly, based on language pairs. Model-wise, we can generally recommend XL-LEXEME for the `en-zh` language pair, and SBERT models, especially SBD, for `en-de`. The major embedding component should be usages, whereas glosses can be included in addition, but not alone. An open direction to explore would be joined encoding of glosses and usages in bi- or cross-encoder architectures such as (Huang et al., 2019; Blevins and Zettlemoyer, 2020). Differences in performance across language pairs are not straightforward to explain: Typological relatedness might offer an explanation, which goes along with more instances of parallel polysemy. The differences in quality of the underlying bilingual dictionaries, on the other hand, see the `col-en-de` print version criticized in (Liberman, 1984), would rather suggest worse performances on the `en-de` pair. Also, the short and less informative glosses hindered automatic sense linking.

6.2 Target Side Polysemy

The bilingual dictionaries that our dataset is based on do not disambiguate polysemy on the target language side. We thus evaluate a BD(sense-word) mapping (1c) only, even though our XL-WSD candidates with BN glosses are ⟨headword, synset⟩ pairs, and do provide target-side sense information. A sense level mapping on source- and target side would mitigate the sense mismatch errors identified in Section 5.2. To adjust the evaluation to a BD(sense-sense) mapping (1d), gold annotations on the target sense side are needed: This could be resolved with manual annotation or a (semi-)automatic approach: Since we connected the source language sense entries to XL-WSD via semi-automatic WSA (see appendix A), each entry is mapped to one or more BN IDs. The BN relations between the gold sense entry synset(s) and the synsets of the respective candidates can be compared automatically in different ways. One option would be to compute the shortest path of semantic relations between the two synsets.[17] Alternatively, one could compute the synset member overlap, i.e., the number of shared headwords in several languages, though this approach might be prone to just capturing parallel polysemy. In sum, sense-level BDI is implemented here mostly on the source-side: Our bilingual dictionaries do not provide sense information on the target side. The ideal would be working with a truly reversible bilingual dictionary, or to leverage monolingual dictionaries in the target languages as well. Considering the scenario of the original motivation—given a new sense in the source language, how to find a translation in the target language—sense-word BDI can still be useful for applications in bilingual lexicography. This is supported by the fact that most bilingual dictionaries, including the ones we use for dataset construction, do not disambiguate on the target side.

6.3 Towards a Real-World Setup

This work models SENSEBDI in an artificial setting. We divide our data into train and evaluation portions based on timestamps, but we do not have access to actual new bilingual senses. For transfer to real-world scenarios, we would need access to up-to-date corpora, and ideally, time stamps of bilingual dictionary entries. Larger corpora would prevent the error patterns related to data sparsity (see Section 5.2), namely domain mismatch and low number of usages. In such a scenario, it could be interesting to explore the usage of LLMs for finding translation equivalents of (really) new senses that were not seen during pre-training. To further mimic the lexicographic process, Retrieval-Augmented Generation (RAG) could be leveraged to retrieve translation equivalents from up-to-date corpora such as social media data or lexicographers' data resources.

With timestamps of dictionary edits available, it would be highly interesting to take the edit history into account and mimic the diachronic dynamics of the lexicographic workflow more concretely. For a user-generated dictionary, this could be explored using Wiktionary. For an expert-curated dictionary, especially a bilingual one, this type of data is not

[17] Semantic relations between synsets in BN cover hypernymy/hyponymy, meronymy/holonymy, and other.

freely available, and research cooperation with a dictionary publisher would be required.

7 Conclusion

This work has introduced the novel task of Sense-Level Bilingual Induction, SENSEBDI, inspired by the lexicographic workflow of drafting new bilingual dictionary articles: Given sense entries linked across bilingual, monolingual, and a diachronic dictionary, translations are automatically induced. It further provided a way to enrich bilingual dictionary entries with detailed source language information with regard to usages, glosses, sense distinctions, and diachronic information. Our baseline approach to solve this task embeds components of the linked dictionary entries with cross-lingual transformer models, and finds a mapping of source language senses and target language headwords on the basis of nearest neighbor embeddings cosine similarities. This work has been a first attempt to model SENSEBDI on the source side, target side polysemy remains to be explored.

Limitations

This work has focused on two high-resource language pairs with English as a source language, and German and Simplified Chinese as target languages, respectively. Whether the proposed approach is extensible to language pairs with fewer resources remains an open question. Previous work has shown that Word-BDI methods that perform well on high-resource LPs are not necessarily transferable to low-resource pairs (Denisová and Rychlý, 2024). The experiments have been carried out on a simple kNN-based baseline approach to serve as a starting point. We leave the exploration of more advanced LLM-based approaches to future work.

Data The dataset creation process relies heavily on access to and availability of expert-created bilingual dictionary resources. English as a source language is enriched with further information not present in the bilingual dictionaries (usages from ODE, diachronic information from OED).

The WSA approach relies on manual postcorrection, which is not ideal for several reasons. Firstly, due to the subjectivity of the task, it would have been better to include more than one annotator. Secondly, this restriction hinders scaling up the dataset construction to create more dataset instances. The semi-automatic approach presented here serves as a starting point, but would require to be revised.

The setup generally suffers from data sparsity on several levels: In some sense entry instances, not all glosses are available, in others, no usages for the target words, or usages only for a different sense of the target language translation - the size of the resulting dataset is therefore small. This is due to the complexity of sense-level tasks on the one hand and to the heterogeneity and complexity of dictionary data on the other. The intersection of polysemy and cross-linguality in NLP poses challenges, and it is difficult to capture both aspects (Giunchiglia et al., 2023; Goworek and Dubossarsky, 2025).

A major issue in methodologies that rely on pre-trained LMs is data contamination (Sainz et al., 2023): Overlap between (pre-)training and test data results in overestimation of system performance. In our case, data contamination may appear on a more abstract level: While it is ensured that fine-tuning data of the employed model XL-LEXEME does not overlap with XL-WSD and BN glosses, the overall artificial task setup has the weakness that finding "new" translation equivalents is easier since they have been seen in multilingual pre-training data, which would not be the case in a real-world scenario.

Target Side Polysemy The bilingual dictionaries that our dataset is based on do not disambiguate polysemy on the target language side. We thus evaluate a BD(sense-word) mapping (1c) only, even though the XL-WSD candidates with BN glosses are ⟨headword, synset⟩ pairs, and do provide target-side sense information. To adjust the evaluation to a BD(sense-sense) mapping (1d), gold annotations on the target sense side are needed: This could be resolved with manual annotation or a (semi-)automatic approach: Since we connected the source language sense entries to XL-WSD via semi-automatic WSA (A), each entry is mapped to one or more BN IDs. The BN relations between the gold sense entry synset(s) and the synsets of the respective candidates can be compared automatically in different ways. One option would be to compute the shortest path of semantic relations between the two synsets.[18] Alternatively, one could compute the synset member overlap, i.e., the number of shared headwords in several languages, though this approach might be prone to just capturing parallel polysemy. In sum, sense-level BDI is implemented here mostly on the source-side: Our bilingual dic-

[18] Semantic relations between synsets in BN cover hypernymy/hyponymy, meronymy/holonymy, and other.

tionaries do not provide sense information on the target side. The ideal would be working with a truly reversible bilingual dictionary, or to leverage monolingual dictionaries in the target languages as well. Considering the scenario of the original motivation - given a new sense in the source language, how to find a translation in the target language - sense-word BDI can still be useful for applications in bilingual lexicography. This is supported by the fact that most bilingual dictionaries, including the ones we use for dataset construction, do not disambiguate on the target side.

Ethical Considerations

Reproducing the dataset construction pipeline requires access to the dictionary APIs used in this work. Some of these APIs are licensed services and may require an institutional or paid subscription for use. API access for research purposes is required for the bilingual dictionaries: Cambridge English-Chinese (Simplified)[19] and Collins English-German[20]. For scraping the monolingual Oxford University press dictionaries, we accessed the entries via a research account for the Oxford Dictionary of English[21] and institutional access for the Oxford English Dictionary[22].

Acknowledgments

This work constitutes part of the first author's Master's thesis at Heidelberg University. We thank the Institute for Computational Linguistics for providing access to their CPU servers used in this work. We are grateful to Oxford University Press, Cambridge University Press, and Harper Collins Publishing for providing API access for research purposes. We thank Andreas Witt and Dominik Schlechtweg for helpful comments and insightful recommendations on literature and datasets, and Melis Çelikkol for help with dataset construction. We thank the anonymous reviewers for their thorough feedback.

References

Arleta Adamska-Sałaciak. 2022. *The Bloomsbury handbook of lexicography*, chapter Issues in compiling bilingual dictionaries. Bloomsbury Publishing.

Arleta Adamska-Sałaciak. 2010. Examining equivalence. *International Journal of Lexicography*, 23(4):387–409.

Omar Adjali, Emmanuel Morin, Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2022. Overview of the 2022 BUCC Shared Task: Bilingual Term Alignment in Comparable Specialized Corpora. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022*, pages 67–76, Marseille, France.

Timothy Benbow. 1987. The new Oxford English dictionary project. In *Proceedings of Translating and the Computer 9: Potential and practice*, London, UK. Aslib.

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.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Preprint*, arXiv:1710.04087.

Michaela Denisová and Pavel Rychlý. 2024. A survey of neural-network-based methods utilising comparable data for finding translation equivalents. *Preprint*, arXiv:2410.15144.

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

Qiuyu Ding, Hailong Cao, Zihao Feng, Muyun Yang, and Tiejun Zhao. 2025. Enhancing bilingual lexicon induction via harnessing polysemous words. *Neurocomputing*, 611:128682.

[19] https://dictionary-api.cambridge.org/

[20] https://www.collinsdictionary.com/collins-api

[21] https://premium.oxforddictionaries.com/

[22] https://www.oed.com/information/get-help-with-access/

Darja Fišer, Nikola Ljubešić, and Ozren Kubelka. 2012. Addressing polysemy in bilingual lexicon extraction from comparable corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Fausto Giunchiglia, Gábor Bella, Nandu C. Nair, Yang Chi, and Hao Xu. 2023. Representing interlingual meaning in lexical databases. *Artificial Intelligence Review*, 56(10):11053–11069.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Roksana Goworek and Haim Dubossarsky. 2025. Multilinguality does not make sense: Investigating factors behind zero-shot transfer in sense-aware tasks. *Preprint*, arXiv:2505.24834.

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.

Franz Josef Hausmann and Rainer Otto Werner. 1991. Spezifische bauteile und strukturen zweisprachigen wörterbücher: Eine Übersicht. In Franz Josef Hausmann and 1 others, editors, *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, volume III, pages 2729–2769. Mouton de Gruyter, Berlin. (1989–1991).

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.

Giovanni Iamartino. 2017. Lexicography, or the gentle art of making mistakes. *Altre Modernità: Rivista di studi letterari e culturali*, (1):48–78.

Hiroyuki Kaji. 2003. Word sense acquisition from bilingual comparable corpora. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 111–118.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1562–1575, St. Julian's, Malta. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. 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 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium. Association for Computational Linguistics.

Annette Klosa-Kückelhaus and Carole Tiberius. 2024. The lexicographic process revisited. *International Journal of Lexicography*, 38(1):1–12.

Sidney I. Landau. 2001. *Dictionaries: The Art and Craft of Lexicography*, 2nd edition. Cambridge University Press, Cambridge, UK.

Martin Laville, Emmanuel Morin, and Phillippe Langlais. 2022. About evaluating bilingual lexicon induction. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 8–14, Marseille, France. European Language Resources Association.

Robert Lew. 2022. *The Bloomsbury handbook of lexicography*, chapter Identifying, ordering and defining senses. Bloomsbury Publishing.

Anatoly Liberman. 1984. Review essay: A new german-english dictionary. *The German Quarterly*, 57(2):280–287.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Henrik Lorentzen and Lars Trap-Jensen. 2016. What, when and how?–the art of updating an online dictionary. In *Proceedings of the XVII Euralex International Congress*, pages 6–10.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, Francesco Cecconi, and 1 others. 2021. Ten years of babelnet: A survey. In *IJCAI*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. 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.

Judy Pearsall and Patrick Hanks. 1998. *The new Oxford dictionary of English*. Clarendon Press, Oxford.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2023. *Induction of Bilingual Dictionaries*, pages 61–95. Springer International Publishing, Cham.

Catherine Soanes and Angus Stevenson, editors. 2003. *Oxford Dictionary of English*. Oxford University Press.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Angus Stevenson. 2010. *Oxford Dictionary of English*. Oxford University Press.

Geart Van der Meer. 2000. Core, subsense and the" new oxford dictionary of english"(node): on how meanings hang together, and not separately. In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000: Stuttgart, Germany, August 8th-12th, 2000*, pages 419–431.

István Varga and Shoichi Yokoyama. 2009. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 862–870, Singapore. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Edwin B. Williams. 1959. The problems of bilingual lexicography particularly as applied to spanish and english. *Hispanic Review*, 27(2):246–253.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Dmitry Yermolovich. 2002. Bilingual dictionary as a mirror of language change and modern lexicographic challenges. *Across Languages and Cultures*, 3(2):201–226.

Tadej Škvorc and Marko Robnik-Šikonja. 2025. Solving word-sense disambiguation and word-sense induction with dictionary examples. *Preprint*, arXiv:2503.04328.

A WSA

We implement a semi-automatic Word Sense Alignment (WSA) approach. We compute semantic similarities using sentence embeddings from Sentence BERT (Reimers and Gurevych, 2019). Specifically, we employ the RoBERTa Base cross-encoder model fine-tuned on the Semantic Textual Similarity (STS) benchmark dataset[23].

The algorithm is shown in Algorithm 1: Given a bilingual dictionary BD that is to be connected to a monolingual dictionary D, the intersection of their English vocabularies is iterated. For each common headword, only entries with the same POS are considered. The scoring function $\texttt{score}(x, y)$ computes semantic similarity between string representations of senses $\texttt{str}(\sigma)$. We experimented with two types of string representations: (i) concatenation of all information or (ii) a weighted combination of component-wise similarity scores ($i*$ `gloss_similarity`, $j*$`example_similarity`, $k*$`other_similarity`). The values $i = 0.5$, $j = 0.3$, and $k = 0.2$ were chosen to prioritize gloss similarities. Disadvantages of the component-wise scores (ii) are possibly very low or even 0 scores in the case where major components are not available; e.g. `col-en-de` in many cases does not provide glosses. XL-WSD on the other hand only provides glosses and usages, but no other information.

B Dataset Statistics

We report the POS statistics of the original 1000 headword list in Table 5.

The numbers of headwords, senses, and translations in the final data set are shown in Table 6. We also report the average number of translations per sense (ST). We split the data set of sense entries X into train X_{train}, validation X_{val}, and test sets X_{test} by timestamp, with 80% train, and 10% validation and test portions each. The validation and test sets can be grouped into an evaluation set X_{eval}. Due to the relatively small test set sizes, we report results on the merged test and validation sets in 5.1. We further control for a balanced POS distribution during splitting, which is the reason for timestamp overlaps in the `en-zh` portions in Table 6. The small dataset size overall is a byproduct of limited cross-lingual sense-level data and the filtering steps in the dataset construction process. It is also important to note that the focus is on word senses, not headwords.

[23] https://huggingface.co/cross-encoder/stsb-roberta-base

UPOS	Number of headwords
NOUN	458
ADJ	179
VERB	118
X	22
ADV	21
ADP	3
NUM	3
DET	2
INTJ	2
PRON	2
several	190

Table 5: UPOS counts of the sampled 1000 headwords according to entries in `cam_en_zh`. *several* refers to headwords with entries of several POS tags, e.g. *NOUN* and *VERB*.

C Dataset Example

Table 7 shows a running example of a sense entry for the source language headword *marketplace*.

D Candidate Extraction

Candidate statistics in both target languages are shown in Table 8. The `de` portion of XL-WSD only contains sense annotations for nouns, while the `zh` portion provides annotation for all open-class POS. We only consider candidates with the same POS. As can be seen from the headword and synset counts, the Chinese data, apart from adverbs, exhibit more polysemy with more synsets than unique headwords, whereas the German data contains a few synonyms, with slightly more headwords than synsets.

E Metrics

For each instance x of the test set X, we compare the set of true translation equivalents T^x of length m to the ranked list of predictions of the candidate set $\text{pred}(x)$. $\text{pred}(x)$ is the list of can-

Algorithm 1 Word Sense Alignment across Dictionaries

Require: Dictionaries BD, D with headwords $\mathcal{W}_{BD}$, $\mathcal{W}_D$ and sense inventories $\mathcal{S}_{BD}$, $\mathcal{S}_D$

$\mathcal{W}_{\text{common}} \leftarrow \mathcal{W}_{BD} \cap \mathcal{W}_D$
$mapping \leftarrow \{\}$
for $w \in \mathcal{W}_{\text{common}}$ **do**
  for $\sigma_i \in \mathcal{S}^w_{BD}$ **do**
    $candidates \leftarrow \{\sigma_j \in \mathcal{S}^w_D \mid \text{pos}(\sigma_j) = \text{pos}(\sigma_i)\}$
    $similarities \leftarrow \{\text{score}(\text{str}(\sigma_i), \text{str}(\sigma_j)) \mid \sigma_j \in candidates\}$
    $mapping[\sigma_i] \leftarrow \text{argmax}(similarities)$
  end for
end for
  return $mapping$

LP	split	headwords	senses	translations	usages	ST	time range
en-de	train	45	52	52	1294	1.15	0-1779
en-de	val	7	7	8	97	1.14	1780-1834
en-de	test	6	7	8	104	1.14	1844-1960
en-de	eval	13	14	15	201	1.14	1780-1960
en-de	all	56	66	64	1495	1.15	0-1960
en-zh	train	72	87	213	2409	2.62	0-1813
en-zh	val	11	11	24	187	2.18	1770-1854
en-zh	test	12	13	34	247	2.62	1600-1924
en-zh	eval	23	24	57	434	2.42	1600-1924
en-zh	all	88	111	264	2843	2.58	0-1924

Table 6: Numbers of headwords, senses, translations and source language usages in the dataset splits. We also report the ratio of translations per sense (ST) and the time range; 0 refers to Old English or late Old English.

Component	Example Sense Entry
headword w^s	*marketplace*
pos p	NOUN
gloss g_{BD}	world of trade
gloss g_{ODE}	the arena of commercial dealings
gloss g_{OED}	Economics. With the or in attributive use. The arena of commercial dealings; the world of commerce or trade. Cf. market n. II.12.
usages U^s_{ODE}	• "the changing demands of the global *marketplace*" • "…this product will satisfy the commercial and domestic *marketplaces*." • "…unwillingness to face the commercial realities of the *marketplace*."
usages U^s_{OED}	• "Ethnic discrimination in the *marketplace*." • "…the models that happened to be most successful in the *marketplace*."
usages U^t	• "Die New_Yorker_Börse schloss am Freitag ohne Kurs ab und zwar in einem *Markt* , der zwischen besseren als in den USA vorgesehenen Indikatoren und einem Anstieg des Dollars hin- und hergerissen wurde …"
usages $U^{s,t}$	• "in the *marketplace*", "auf dem *Markt*"
translations T	*Markt*
timestamp τ	1960

Table 7: Example of a dataset instance with the different components. Source language usages U^s show selected sentences only (full sets contain additional examples).

didates ordered by descending cosine similarity of embeddings. We denote Top-k(pred(x)) as the headwords w^t of the top k extracted candidates.

$P@k$ is computed for each instance x (2) and then averaged over the whole dataset X (3).

$$P@k(x) = \begin{cases} 1 & \text{if } |\text{Top-}k(\text{pred}(x)) \cap T^x| \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

l	POS	headwords	synsets	candidates
de	NOUN	484	481	509
zh	NOUN	876	1244	1495
zh	ADJ	253	410	468
zh	ADV	74	59	84
zh	VERB	609	1019	1408

Table 8: Statistics of headwords and BN synsets of the candidates in target language l. A candidate is a ⟨headword, synset⟩ pair.

$$P@k = \frac{1}{|X|} \sum_{x \in X} P@k(x) \tag{3}$$

Mean Average Precision (MAP, 6) is defined as the Average Precision (AP, 5) averaged over all instances. With $m = |T^x|$ as the number of gold translations for an instance, at each rank k where a correct translation is found, denoted with R_k, the precision is computed (Adjali et al., 2022). Thus, $P(R_k)$ (4) is the precision at the rank of the k-th correct translation.

$$P(R_k) = \frac{|\text{Top-}k(\text{pred}(x)) \cap T^x|}{k} \tag{4}$$

$$AP(x) = \frac{1}{m} \sum_{k=1}^{m} P(R_k) \tag{5}$$

$$MAP = \frac{1}{|X|} \sum_{x \in X} AP(x) \tag{6}$$

The reciprocal rank is defined as the inverse rank of the first correct prediction $rank_i$ in pred(x), which is averaged over all instances to report the Mean Reciprocal Rank (MRR, 7). MAP and MRR are equal in cases where there is exactly one true translation, $|T^x| = 1$.

$$MRR = \frac{1}{|X|} \sum_{i=1}^{|X|} \frac{1}{rank_i} \tag{7}$$