Leveraging Translations for Lexical Semantics

by

Arnob Mallik

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

We leverage multilingual translations from parallel corpora to improve sense annotations, build end-to-end Word Sense Disambiguation pipelines and detect cross-lingual lexical entailment. Based on theories of translational equivalence, we propose novel algorithms capable of correcting noisy sense annotations on a parallel corpus. We show that, when applied to bilingual slices of a parallel corpus, these algorithms can rectify noisy sense annotations and thereby produce multilingual sense-annotated training data of improved quality. Furthermore, we propose novel end-to-end pipelines which can produce high-quality sense annotations from scratch in a fully unsupervised manner. Our methods achieve state-of-the-art results on standard WSD datasets for unsupervised approaches in several languages. Additionally, by exploring the generalization property of translations, we develop novel approaches to detect cross-lingual lexical entailment by leveraging word embeddings along with translations. We evaluate our methods on a standard shared task dataset and achieve encouraging results constituting a strong proof-of-concept. In summary, our results in three different tasks of lexical semantics confirm the utility of translations in this field.

Preface

The work presented in chapter 3 of thesis is adapted from a research article in submission (Mallik and Kondrak, 2021). I was the principal contributor who implemented different methods that are included in this chapter, and conducted all related experiments.

The work presented in chapter 4 of this thesis is adapted from a research article in submission (Hauer et al., 2021). I was the principal contributor who implemented different methods that are included in this chapter, and conducted all related experiments.

The work presented in chapter 5 of this thesis is part of a system description article published as B Hauer, AA Habibi, Y Luan, A Mallik, G Kondrak, "UAlberta at SemEval-2020 Task 2: Using translations to predict cross-lingual entailment," proceedings of the Fourteenth Workshop on Semantic Evaluation, Pages 263–269 (Hauer et al., 2020). I was responsible for conducting experiments for all the results that are included in this chapter, was involved in subsequent analysis and in manuscript composition.

Acknowledgements

I would like to thank my supervisor, Professor Greg Kondrak, for his generous support and guidance throughout the thesis. I would also like to acknowledge Professor Lili Mou for his assistance with computational resources.

I would like to thank Bradley Hauer, Yixing Luan and Amir Ahmad Habibi for their guidance, suggestions and research advice.

Finally, I am grateful to my parents for their continuous support throughout my graduate studies.

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

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we first provide some background on the fundamental concepts related to our thesis. Then we declare our thesis statement and briefly describe the three tasks that constitute this thesis. Finally, we provide an outline of the rest of the thesis.

1.1 Background

Word Sense Disambiguation

In human language, content words are generally ambiguous in the sense that a word may have multiple meanings in the dictionary. The correct sense of a word can be inferred from its context. For example, consider the following sentences :

(a) The *bat* is feeding on fruit.

(b) He hit the ball with the *bat*.

In these sentences, it is clear to us that the word *bat* conveys different meanings: a nocturnal mammal with wings and a club used for hitting a ball in various games. However, this task of identifying the correct sense of a word in context is a complicated one for machines, as it involves analysis and processing of unstructured textual information. In the field of lexical semantics, the task of associating a word with its sense chosen from a fixed dictionary is known as word sense disambiguation (WSD) (Navigli, 2009), which is one of the primary focuses of this thesis.

WSD, one of the central problems in natural language processing, is typ-

ically configured as an intermediate task and is useful for numerous applications, such as text processing, information retrieval, and machine translation. For example, a machine translation model that is aware of word senses could translate the English word *bat* into Italian as *pipistrello* (a nocturnal mammal) or *mazza* (a club for hitting a ball), depending on the context.

The WSD task involves choosing the appropriate sense of a word from a predefined sense inventory. For English, the most widely utilized sense inventory is WordNet (Miller, 1995), a manually built lexical database where words are grouped into sets of synonyms, i.e. *synsets*, each expressing a distinct concept. The current version of WordNet [1] covers over 155,287 unique words or phrases, grouped into 117,659 synsets. WordNet also facilitates inter-synset relationships, such as hypernymy-homonymy (ISA relation) and meronymy (part-whole relation), and hence can be thought of as a semantic network. The synsets that a word belongs to are considered to be the senses that the word can represent. Hence, the WSD task can also be thought of as predicting the appropriate synset of a word in context.

Due to the lack of lexical resources in non-English languages, WSD was originally approached as a monolingual task, specific to English only. The paradigm shifted with the advent of BabelNet (Navigli and Ponzetto, 2012), an automatically built multilingual knowledge resource, which can be thought of as a multilingual expansion of WordNet. BabelNet integrates information from WordNet, Wikipedia and leverages machine translation to group synonymous words from various languages into multilingual synsets, and thereby facilitates multilingual WSD. We have used BabelNet version 4.0 for our research, which covers lexicalizations from 284 languages.

Lexical Entailment

Lexical entailment (LE) is a lexico-semantic relation that holds between lexical elements when the meaning of one element can be inferred from the meaning of the other. It is an asymmetric relation, which can also be called a hyponym-hypernym relation. For example, *canary* is a hyponym of *bird*, which means

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

canary entails *bird*. However, *bird* does not entail *canary*, as all birds are not canaries.

Zhitomirsky-Geffet and Dagan (2009) present a formal definition of lexical entailment in terms of substitutability of words: word w entails word v, if a sense of w implies v, and if w can substitute for v in a sentence such that the meaning of the modified sentence entails the meaning of the original one. Vyas and Carpuat (2016) extend this definition to the cross-lingual space by modifying the second condition in terms of translations: word w of one language entails word v of another language if a sense of w implies v, and if w can substitute for v in the translation of a sentence containing v, such that the meaning of the modified sentence entails the meaning of the original sentence.

LE relations are fundamental building blocks of semantics networks such as WordNet and BabelNet. They are essential in many fields of natural language processing, such as taxonomy induction and natural language inference. For example, if we know that *footballer* entails *sportsperson* and if we have the fact "Zidane is a footballer", then we can also imply that "Zidane is a sportsperson". Moreover, cross-lingual lexical entailment provides us the opportunity to infer relations or facts from texts in different languages, which makes it even more intriguing, simply because we have more text sources to work with.

Word Alignment

Word alignment is a fundamental problem in NLP, and is central to any research work concerned with the use of parallel multilingual corpora, i.e., sentence-aligned bitexts. Word alignment tools are employed on a sentence-aligned bitext to retrieve translations of individual words or phrases. All of our methods in this thesis are dependent upon accurate word-level alignments.

In this thesis, we have primarily utilized BabAlign (Luan et al., 2020), a high-precision knowledge-based alignment algorithm to word-align bitexts. BabAlign improves upon the generated output of a base aligner by leveraging translational information from a multilingual knowledge base, BabelNet. We employ FastAlign (Dyer et al., 2013) as the base aligner for BabAlign.

BabAlign augments the input corpus with lexical translation pairs, to

bias the base aligner towards aligning words which are mutual translations. BabAlign also corrects alignments via post-processing to maximize the number of aligned words which are translations. This emphasis on recovering word-level translation information makes BabAlign particularly well-suited to our methods which are reliant on mining translation pairs.

1.2 Thesis Statement

In this thesis, we demonstrate that *multilingual translations extracted from parallel corpora can be leveraged to improve the quality of sense annotations, build WSD pipelines, and detect cross-lingual entailment.* To this end, we exploit two distinct properties of translations: equivalence and generalization. The equivalence property implies that a word and its translation should, in most cases, represent the same concept. This idea leads us to propose algorithms capable of making annotation corrections on an automatically sense-annotated parallel corpus. Furthermore, using this idea, we develop novel pipelines to annotate both sides of a bitext from scratch. On the other hand, the generalization property of translations implies that Words, in some cases, can be translated into more general concepts. We explore this phenomenon to develop methods for detecting cross-lingual entailment.

1.3 Contributions

In this section, we briefly describe the three tasks that are our primary contributions in this thesis.

Improving Automatic Sense Annotations

Acquiring large amounts of high-quality annotated data is an open issue in WSD, which has become more critical recently with the advent of neural network based supervised WSD models. As supervised systems consistently outperform their knowledge-based counterparts, recent research has focused on making these systems applicable for multilingual WSD, by producing large sense-annotated corpora automatically. Some of these automated annotation

approaches operate by exploiting a parallel corpus (Taghipour and Ng, 2015; Bovi et al., 2017).

In our thesis, we propose two algorithms, MultiWordNet and Bipartite, which leverage theories of translational equivalence to make selective corrections on an automatically sense-tagged parallel corpus. The MultiWordNet algorithm operates on each word alignment link individually, while the Bipartite algorithm takes into consideration all the alignments in the corpus and makes corrections based on frequency. We apply our algorithms to an existing sense-annotated parallel corpora and perform both intrinsic and extrinsic evaluations. We compare our results to those obtained by the original corpora, and thus demonstrate the utility of our algorithms in reducing noise in sense annotations.

Unsupervised Corpus Annotation Pipelines

We propose novel end-to-end pipelines, LabelSync and LabelGen, which can be applied to unannotated bitexts to automatically produce multilingual training data for WSD systems in a fully unsupervised manner. Both of our approaches are independent of manual annotation efforts and are scalable to any language or domain. We employ an off-the-shelf unsupervised WSD model as our baseline and refine the initial annotations provided by the model using translations.

We use the multilingual sense-tagged data produced by the pipelines to train supervised WSD systems and perform evaluations on standard English and multilingual benchmark datasets. We not only achieve state-of-the-art results among unsupervised approaches in several languages, but also rival the performance achieved by training on a manually annotated corpora.

Detecting Cross-Lingual Lexical Entailment

To detect cross-lingual lexical entailment, we leverage translations mined from bitexts to construct an initial set of entailment pairs, and vectorized representation of words, i.e. embeddings, to expand the set of entailment pairs.

We first propose a simple baseline approach, BiText, solely based on translation pairs mined from parallel corpora. Then, we devise an improved method, Vectors, by exploiting similarities between monolingual embeddings of words. We evaluate our methods on a shared task dataset. Our experimental results confirm the utility of translations in detecting entailment and thereby constitute a solid proof-of-concept.

1.4 Outline

The thesis is organized as follows: In chapter 2, we first revisit prior works which have leveraged translations to sense annotate words. Then we review existing WSD approaches. Lastly, we show prior attempts to detect cross lingual LE by using translations. In chapter 3, we describe our annotation correction algorithms: MultiWordNet and Bipartite, which can be applied to noisy sense-annotated bitexts to make error corrections. We subsequently present our experimental settings, evaluation results and analysis. In chapter 4, we introduce our corpus labelling pipelines: LabelSync and LabelGen, which can be applied to unannotated bitexts to automatically produce multilingual training data for WSD systems. We train existing WSD systems using our data, evaluate on standard test datasets, and compare our results to those obtained by previous unsupervised approaches. In chapter 5, we present our cross-lingual lexical entailment detecting methods: BiText and Vectors. We carry out our experiments in low-resource and high resource settings, and report results on a standard shared task dataset.

Chapter 2

Related Work

In this chapter, we provide an overview of prior approaches that are related to our work in this thesis. In section 2.1, we show prior attempts of acquiring sense annotations by leveraging translations, which relate to our work of improving annotations in chapter 3. Then in section 2.2, we describe various existing WSD approaches including supervised, semi-supervised and knowledge-based methods, and also point out similarities of these approaches to our WSD pipelines described in chapter 4. Finally in section 2.3, we discuss prior works of detecting cross-lingual lexical entailment, and also describe their relation to our work in chapter 5.

2.1 Sense Annotations from Parallel Corpora

Over the years, various approaches have been proposed to fully or partially automate the process of acquiring high quality multilingual sense annotations by exploiting parallel aligned bilingual corpora. Resnik (1997) proposed that different translations of an ambiguous source word in a target language could serve as sense-tagged training examples. Due to the lack of large scale parallel corpora at that time, they could not validate their intuition experimentally. However, they conjectured that supervised systems would eventually leverage bilingual corpora for sense distinctions. This idea was put into practice by Ng et al. (2003) and then on a large scale by Chan and Ng (2005), as they implemented an approach of disambiguating English nouns using distinct Chinese translations, leveraged from an English-Chinese parallel corpora. Their

approach was not fully automated, as they had to manually select target translations for each sense of the English nouns to be disambiguated. More recently, Taghipour and Ng (2015) used the same semi-automatic approach to create a publicly available WSD training set based on the WordNet sense inventory by leveraging the Chinese-English part of the MultiUN corpus (Eisele and Chen, 2010). Bovi et al. (2017) removed the bottleneck of manual intervention, as they proposed a fully automated approach of producing multilingual sense-tagged corpora by jointly disambiguating multiple languages of a parallel corpus.

In contrast to these approaches, our work in chapter 3 is focused on leveraging translations to improve the quality of an already sense-tagged parallel corpus rather than to annotate the corpus from scratch. Nonetheless, our proposed algorithms are inspired by the central idea of the aforementioned research works, that translations may provide the necessary information to disambiguate an ambiguous word.

2.2 WSD Approaches

We can divide the primary WSD approaches into supervised and knowledge-based methods. Supervised WSD systems, which rely on sense-annotated corpora, have historically achieved the best overall results on standard WSD datasets (Raganato et al., 2017). Recent supervised approaches use deep neural models to achieve state-of-the-art results (Kumar et al., 2019; Huang et al., 2019; Bevilacqua and Navigli, 2020). However, their utility is limited by the high cost and difficulty associated with manually creating large sense-annotated corpora. In particular, there is a severe lack of high-quality sense-annotated corpora for languages other than English, which is known as the knowledge acquisition bottleneck problem (Pasini, 2020). This limitation is the principal motivation behind our work in chapter 4.

The most common alternative to supervised WSD systems are knowledge-based (KB) WSD systems. Rather than depending on labelled training data, KB WSD methods rely on a lexical knowledge base (LKB), such as WordNet

or BabelNet. An LKB can be viewed as a graph, where nodes are concepts and edges are semantic relations such as hypernymy, homonymy, meronymy and holonymy. Babelfy (Moro et al., 2014) is an example KB WSD system that operates by applying random walks with restarts to BabelNet. Similarly, UKB (Agirre et al., 2014) performs WSD by applying personalized PageRank algorithm on WordNet. Maru et al. (2019) proposed SyntagNet, a manually curated resource of semantic relations, which can be integrated on top of a baseline LKB (e.g. WordNet). SyntagNet enables UKB to achieve better results for both English and multilingual WSD, rivalling the performance of supervised systems. We have used UKB in conjunction with SyntagNet as the baseline WSD model for our approaches described in chapter 4.

Multilingual WSD can also be performed by leveraging sense-level representations, on which proximity based algorithms (e.g. k-nearest neighbors) are applied. Scarlini et al. (2020a) proposed SensEmBERT, an unsupervised approach of producing BERT (Devlin et al., 2018) based synset embeddings by leveraging lexical-semantic information in BabelNet and Wikipedia. A similar approach is ARES (Scarlini et al., 2020b), which constructs synset representations by utilizing SemCor annotations, sense definitions and syntagmatic information from SyntagNet. This approach is termed "semi-supervised" in the sense that it utilizes the manual annotations on English, but generalizes to other languages as well.

Another approach to tackle multilingual WSD is to automatically construct multilingual sense-annotated data that can be used to train supervised systems. The net result is a WSD system which is not supervised, but rather is either knowledge-based or semi-supervised, depending on whether the method for automatically generating the training data is itself unsupervised or semi-supervised. The current state-of-the-art corpus labelling method is MuLaN (Barba et al., 2020), a semi-supervised approach which propagates sense annotations from SemCor and WordNet Gloss Corpus (WNG) (Langone et al., 2004), to semantically similar contexts in Wikipedia corpora using contextual word representations from multilingual BERT.

Unsupervised methods of generating WSD training data produce sense

annotations "from scratch", with no dependence on existing tagged corpora. Train-O-Matic (Pasini and Navigli, 2020) annotates Wikipedia in multiple languages by applying PPR to BabelNet. OneSeC (Scarlini et al., 2019) combines Wikipedia categories and BabelNet synset representations to produce WSD training data, and outperforms Train-O-Matic. However, both Train-O-Matic and OneSeC annotate nominal instances only, and hence are not applicable to all-words WSD. The most similar prior work to our unsupervised corpus labelling approaches (Chapter 4) is that of Bovi et al. (2017), where they jointly disambiguate a parallel corpus using a knowledge-based WSD system, Babelfy, and subsequently refine the annotations using distributional similarity. In contrast, we refine our initial annotations by leveraging the translational information already existing in the parallel corpus.

2.3 Translations as Broader Concepts

It has been observed in prior work on cross-lingual lexical semantics that translations may be broader in meaning than the original text. In an attempt to quantify the lexical gaps between English and Italian lexica, Bentivogli and Pianta (2000) introduced the notion of denotation differences, cases where the translational equivalent of a source language word is either a cross-lingual hyponym or a cross-lingual hypernym. Rudnicka et al. (2012) formulated a set of inter-lingual semantic relations in a bid to map the Polish WordNet (Maziarz et al., 2012) onto the Princeton WordNet, and found that inter-lingual hyponymy and hypernymy accounted for half of all the inter-lingual relations. We describe a simple alignment based method to detect entailment in section 5.1.1, which is motivated from these findings.

Our semantic expansion method described in 5.1.2 is inspired from the nearest neighbor method of Qiu et al. (2018), which was the best performing system for discovering hypernyms in Spanish in SemEval-2018 Task 9 (Camacho-Collados et al., 2018). Their method is purely monolingual and is based on the assumption that hyponyms which are close to each other in terms of cosine similarity in the embedding space, often share the same hy-

pernyms. One major limitation of their system is that they cannot extract hypernyms that do not exist in the training set. In contrast, we attempt to predict cross-lingual hypernyms using nearest neighboring hyponyms and also address the dependency on the training set by creating a much larger pseudo training list resulting from the alignment of a parallel corpus.

Chapter 3

Correcting Sense Annotations using Translations

In this chapter, we propose two algorithms: MultiWordNet and Bipartite, both of which make use of translation information to reduce the noise of automatically sense-annotated parallel corpora. Both these algorithms are based on unifying theories of synonymy and translational equivalence (Hauer and Kondrak, 2020).

3.1 MultiWordNet Algorithm

Algorithm 1 MultiWordNet Algorithm

Input : Aligned Sense Pair (s,t).
$w(s) \leftarrow$ Word of which s is a sense
$M(s) \leftarrow$ Multi-Synset that contains sense s
$M(w) \leftarrow$ Set of multi-synsets that contain word w.

$C \leftarrow M(w(s)) \cap M(w(t))$
if $M(s) \neq M(t)$ **then**
  if $M(s) \in C$ and $M(t) \notin C$ **then**
    $t \leftarrow (w(t), M(s))$
  end if
  if $M(t) \in C$ and $M(s) \notin C$ **then**
    $s \leftarrow (w(s), M(t))$
  end if
end if

The MultiWordNet algorithm operates on a single aligned sense pair and attempts to make corrections in selective cases where aligned words are

not annotated with the same synset. To this end, we define a *common multi-synset* between two aligned words to be a multilingual synset that contains both the words. For each alignment link, there can be three possible cases in terms of the number of common multi-synsets :

No common multi-synset : The two words forming an alignment link represent two different concepts and neither of the two concepts can be represented by both words. In such cases, we suspect multiple errors in annotations or in the MultiWordNet and hence we do not attempt corrections.

Two common multi-synsets : The two words forming an alignment link represent two different concepts and both the concepts can be represented by both words. In such cases, the annotation on either side could be incorrect, and hence we do not make any correction.

One common multi-synset : Only one of the aligned concepts can be represented by both the aligned words. In such cases, we posit that both the aligned words should represent this common concept and we make the correction.

Algorithm 1 shows our approach of making annotation corrections using the algorithm.

3.2 Bipartite Algorithm

The Bipartite algorithm takes all the alignment links of the given bitext as input and attempts to make annotation corrections based on the most frequent links. To this end, we construct an undirected bipartite graph whose vertices can be divided into two disjoint sets, each containing the synsets of a language represented by the bitext. Every edge in this graph corresponds to an alignment link between two synsets. Only the links that are most frequently observed from both directions are taken as edges in the graph. This process results in a graph where every node has a degree of 1. The goal of this process is to create mappings between identical concepts across languages. For our English-Italian experiments, 75.3% of the mappings returned by our algorithm are of identical concepts. At the next step, we make annotation corrections based on the edges of the constructed bipartite graph.

Algorithm 2 BIPARTITE Algorithm

Input: Alignment links involving synset pairs (p_1,q_1),(p_2,q_2),....,(p_m,q_m), where $p_i \in$ *Language P* and $q_i \in$ *Language Q*
Input : Frequency Threshold α

1: *candidate_edges_P* $\leftarrow \varnothing$, *candidate_edges_Q* $\leftarrow \varnothing$
2: Initialize Graph G (E), where Edges $E \leftarrow \varnothing$

3: **for** each language L in (P,Q) **do**
4: **for** each synset x in Language L **do**
5: $n \leftarrow$ total alignment links involving x
6: **for** each synset y aligned to x **do**
7: $a \leftarrow$ total alignment links involving (x,y)
8: **if** $a \div n > \alpha$ **then**
9: *candidate_edges_L* $\leftarrow$ *candidate_edges_L* $\cup (x, y)$
10: **end if**
11: **end for**
12: **end for**
13: **end for**

14: $E \leftarrow$ *candidate_edges_P* $\cap$ *candidate_edges_Q*

15: **for** each edge (p,q) in E **do**
16: **for** each synset q_i aligned to p **do**
17: $w_q \leftarrow$ associated word of Language Q
18: **if** $q_i \neq q$ and $w_q \in q$ **then**
19: CORRECT: Alignment Link $(p, q_i) \implies (p, q)$
20: **end if**
21: **end for**
22: **end for**

Algorithm 2 shows our approach of making corrections on language Q, using base language P. Firstly, we initialize the sets *candidate_edges_P*, *candidate_edges_Q*, and the undirected bipartite graph G (lines 1-2). In lines 3-12, we select the most frequent alignment link involving each synset of language P and add it to the set of *candidate_edges_P*. We follow the exact same procedure from the opposite direction and add the most frequent alignment link involving each synset of language Q to the set of *candidate_edges_Q*. In these iterations, we leave out alignment links, which have a lower relative frequency than a frequency threshold α, which is configured to be greater than 0.5. Then, in

Language	Bitext Used	Valid Aligned Sense Pairs	Corrections by MWN Algorithm	Corrections by bipartite Algorithm
EN	EN-IT	4,713,589	235,087	89,798
IT	EN-IT	4,713,589	541,326	82,685
FR	EN-FR	5,219,146	664,253	106,023
DE	EN-DE	3,083,325	179,400	59,446
ES	EN-ES	5,015,140	518,488	92,634

Table 3.1: Dataset and Correction Statistics

line 14, we select the edges that are present in both *candidate_edges_P* and *candidate_edges_Q*, and add them to our graph G. Each node in the graph will have a degree of 1. Each edge in the graph will represent an alignment link that is observed most frequently for the involved synsets. In lines 15-22, we leverage this graph to make annotation corrections on language Q. For every undirected edge (p,q) in G, we iterate over every alignment link in the bitext involving synset p. If p is aligned to a synset q_i that is not equal to q and if the involved word in the link w_q belongs to synset q, then we make the correction and annotate w_q with synset q.

3.3 Evaluation

In this section, we provide the details of our experiments, including the source corpus, pre-processing steps. extrinsic and intrinsic evaluations and corresponding results.

3.3.1 Experimental Setup

Both of our algorithms take annotated translation pairs as inputs. To retrieve these pairs, we leverage the EuroSense high-precision corpus (Bovi et al., 2017), an automatically constructed sense-annotated resource based on the EuroParl parallel corpus (Koehn, 2005). In EuroSense, the words are annotated with multilingual synsets from BabelNet and annotated words are also accompanied by their respective lemma forms. We extract 4 sentence-aligned bitexts from EuroSense, by considering 4 different language pairs: English-Italian (EN-IT), English-German (EN-DE), English-French (EN-FR) and English-Spanish

Training Set	Test Set							
	SemEval 2015			SemEval 2013				
	EN	IT	ES	EN	IT	FR	DE	ES
EuroSense	64.3	56.3	54.3	65.3	56.5	45.4	58.8	53.9
ES + MultiWordNet	65.1	57.1	55.3	65.5	58.3	48.0	60.0	56.7
ES + Bipartite	64.5	57.2	55.3	65.4	56.7	45.9	59.1	54.1

Table 3.2: WSD F-score of IMS trained on different corpora

(EN-ES). We employ BabAlign to word-align the bitexts and the aligned word or phrase of each annotated token is treated as its translation.

Before using as inputs to the algorithms, we filter these annotated translation pairs based on the following three cases:

Invalid Senses: A translation pair is filtered out if any of its senses are not valid with respect to the current version of BabelNet.

Entailment Pairs: Although a word and its translation are synonymous in most cases, a subset of the translation pairs may involve the hypernymy relation (Hauer et al., 2020). Since our algorithms are focused on exploiting the synonymy between translation pairs, we filter out a pair if one of the synsets is a hypernym of the other. We detect such cases using hypernymy links in BabelNet.

Non-Literal Translations: A translation pair is filtered out if the involved words do not have any synset in common. We consider these cases as non-literal translations, i.e., translations that do not occur in bilingual dictionaries.

In our experiments, we have found that around 3% of the pairs contain invalid senses, 4.8% involve entailment pairs and 12.9% are cases of non-literal translations. After this filtering procedure, the remaining translation pairs are used as inputs to the algorithms to retrieve corrections for each language separately. For IT, DE, FR and ES corrections, we use EN as the base language. To retrieve EN corrections, we use IT as the base language. Table 3.1 contains dataset and correction statistics for each of the five languages.

Language	Algorithm	# instances where original annotation is correct	# instances where new annotation is correct	# instances where neither annotation is correct
EN	MultiWordNet	6	<u>18</u>	26
	Bipartite	12	18	20
ES	MultiWordNet	11	<u>33</u>	6
	Bipartite	17	20	13

Table 3.3: Results of Intrinsic Evaluation of Annotation Corrections.

3.3.2 Extrinsic WSD Evaluation

We apply the MultiWordNet and Bipartite algorithms separately on the EuroSense corpus to make annotation corrections. We extrinsically evaluate the corrections by providing the corrected corpora as training data for a supervised Word Sense Disambiguation (WSD) system, which is then evaluated on standard benchmark datasets.

To this end, we employ IMS (Zhong and Ng, 2010), a lexical feature based supervised WSD system. To keep the corpus at a reasonable size, we consider a maximum of 10,000 randomly sampled training examples per sense for our experiments. The multilingual WSD evaluation is performed on benchmark parallel datasets from SemEval-2013 task 12 (Navigli et al., 2013) and SemEval-2015 task 13 (Moro and Navigli, 2015). We apply the most frequent sense (MFS) backoff strategy for English, in cases where the system fails to make a prediction. For all languages, any monosemous words are automatically tagged with their single possible sense.

Table 3.2 presents the WSD results of IMS models trained on the corrected corpora, along with the results of models trained on the original EuroSense corpus. The underlined results indicate a statistically significant improvement ($p < 0.05$ using McNemar's test) over the results obtained by the original corpus. It is evident that the supervised system consistently achieves better results when trained on the corrected corpora. This verifies the utility of the annotation corrections made by two algorithms.

3.3.3 Intrinsic Evaluation

To intrinsically evaluate the quality of the annotation corrections made by the algorithms, we designed a manual sense annotation task involving a small random sample of these corrections. The annotators were Computing Science graduate students and were native speakers of the language they were annotating. For each correction, the annotators were asked to examine the corresponding sentence from EuroSense containing the word in focus. They were given only two possible senses of the word, the original sense in the EuroSense corpus and the corrected one, which were both described using BabelNet glosses and synonyms. The order of the senses was randomized.

For each instance, the annotators had to decide which one of the given senses is correct or neither is correct. Each annotator was given 100 correction instances to annotate, 50 from each algorithm. The annotators also had the option to provide comments regarding any particular instance. The evaluation was done for English and Spanish, and the results are presented in table 3.3. The underlined results indicate a statistically significant improvement ($p < 0.05$ using McNemar's test) over the results obtained by the original corpus. It is evident from the results that both the algorithms improve the overall quality of the annotations for both languages.

3.4 Analysis

As apparent from the intrinsic evaluation results in table 3.3, some corrections made by our algorithms are wrong. These cases can be traced back to the following causes :

Incompleteness of BabelNet: Some BabelNet synsets do not contain all possible lexicalizations of the concept that it represents, which often leads to wrong corrections made by the algorithms. For example, the BabelNet synset *bn:00109131a* (Gloss: "Of or concerned with or related to the future") contains the Spanish lemma *"futuro"*, but not the English translation "future". Such cases affect the MultiWordNet algorithm, as it is based on the number of common multi-synsets in an alignment link.

Overall quality of the source corpus: The English-German bitext slice of EuroSense contains a total of 19,230 distinct English synsets, among which only 10,661 (55.44%) have matching German synsets existing in the dataset. This points to a significant amount of noise in the parallel corpus, as a large portion ($\sim$44.5%) of concepts represented in English, does not exist on the German side. Hence, the BIPARTITE algorithm, which is based on the most frequent alignment links, fails to link similar concepts from both languages.

Fine granularity of senses: In our intrinsic evaluation, some of the annotations that were deemed wrong by the annotator involved a choice between fine-grained senses. For example, one instance involved a choice between the synsets *bn:00019918n* (Gloss : "The temporal end; the concluding time") and *bn:00019966n* (Gloss : "A concluding action"). The annotator commented that "Both senses are really close in meaning".

In summary, the limitations of existing lexical resources have a negative impact on the performance of our algorithms. Nevertheless, our extrinsic and intrinsic evaluation results constitute a strong proof-of-concept that translations can be leveraged to make effective annotation corrections on a sense-annotated bitext.

Chapter 4

Unsupervised Corpus Labelling using Translations

In this chapter, we propose two novel unsupervised methods for generating sense-tagged corpora: LabelSync and LabelGen, both of which make use of semantic information obtained from word-level translations and a lexical knowledge base. Both of our approaches are independent of any pre-existing sense-annotated corpora. LabelSync is an unsupervised language-independent approach that produces sense-annotated corpora in two languages at once by independently applying a knowledge-based WSD system to each side of a raw bitext. LabelGen is a modified version of LabelSync, which leverages the advancements in English WSD to improve the quality of multilingual sense-annotations. In this approach, we employ the knowledge-based WSD system to the English side of a bitext and then apply a label propagation method to transfer the English annotations to the non-English side. In both approaches, the initial annotations are revised and filtered on the basis of confidence scores and multilingual information.

Experiments on standard WSD test sets demonstrate that both of our approaches outperform the best comparable methods for English and multilingual WSD. Furthermore, our results rival the performance obtained by training on a manually sense-annotated corpus.

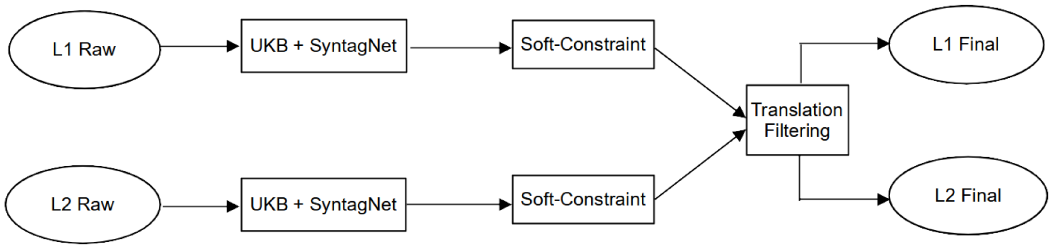

Figure 4.1: Summary of the LabelSync method. L1 and L2 refer to languages. A bitext representing L1 and L2 is used to produce training data in both L1 and L2 for a supervised WSD system.

4.1 LabelSync

LabelSync is a language-independent approach, which, given only a bitext, produces sense annotated corpora in both of the languages it represents. The method is composed of three steps: monolingual word sense disambiguation, multi-lingual post-processing, and translation-based filtering. Figure 4.1 summarizes our approach.

4.1.1 Word Sense Disambiguation

Our goal is to enrich both sides of the input bitext with sense tags. Since no sense-annotated corpus is available, we are limited to the use of knowledge-based WSD methods. Therefore, we employ a variant of UKB (Agirre et al., 2014) enhanced with SyntagNet (Maru et al., 2019), a language independent knowledge-based WSD system. Using UKB, we annotate each side of the bitext independently. Once the system has been applied, the result is two sense annotated corpora, one in each of the languages represented in the bitext.

4.1.2 Multi-Lingual Post-Processing

Now that we have annotated both sides of the bitext independently, we leverage the lexical translation information inherent in the bitext to increase the

accuracy of the sense annotations. We apply the language independent WSD post-processing approach of Luan et al. (2020). In particular, we employ their SoftConstraint method. This method is applicable to any base WSD system which assigns a numerical score, such as a probability, to each sense of a disambiguated word; most modern WSD systems, including UKB, satisfy this property.

The SoftConstraint method depends upon word-level translations of each annotated word, as well as translation information from BabelNet (Navigli and Ponzetto, 2012). We use BabAlign to word-align the bitext, and the aligned word or phrase for each annotated token is treated as its translation.

The SoftConstraint method can also incorporate sense frequency information, to bias the annotations toward more frequent, and so more probable senses. In our development experiments, we found that the inclusion of sense frequency information does not substantially improve the quality of the annotations. We therefore exclude frequency information from this step.

4.1.3 Translation-Based Filtering

In the final step, we aim to further reduce the noise in our sense-annotated corpora by employing a translation-based filtering method. This filtering approach is based on the following constraint: aligned words should be annotated with the same synset, i.e. should be semantically equivalent. Since we annotate each side of the bitext independently, this third step aims to enforce this constraint by synchronizing the sense annotations across both sides of the bitext in order to increase the overall precision.

We favor precision over recall, removing questionable sense annotations, following the example of Bovi et al. (2017). Rather than leverage embeddings of concepts to filter annotations which are "less semantically coherent", we adopt a binary alignment-based criteria based on the assumption of semantic equivalence of lexical translations.

We leverage the word-alignment of bitexts from the previous step to get word-level translations. For each sense-annotated taken, if the sense annotation of its translation does not refer to the same multilingual synset as the

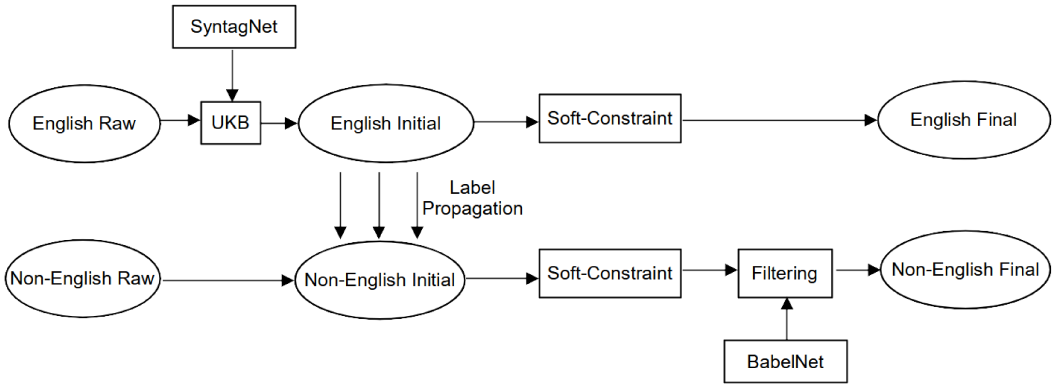

Figure 4.2: Summary of the LabelGen method. A bitext representing English and another language is used to produce training data in both languages for a supervised WSD system.

token itself, then we remove the annotation. Sense annotations are not filtered out if the token cannot be aligned or if its translation is not annotated.

4.2 LabelGen

To improve the quality of multilingual annotations, we have devised a modified version of LabelSync, LabelGen, which requires us to employ the unsupervised WSD base system on the English side of the bitext only. This enables us to influence the non-English side of the bitext with information from high-precision lexical resources that are exclusively available in English. The LabelGen approach, however, is only applicable if one side of the given bitext is English.

The base system of our LabelSync approach, UKB, operates on a semantic network where nodes are concepts, which are language-independent by definition. In order to retrieve sense candidates for words in a language, these concepts need to be mapped to lexicalizations in that particular language. For English, this candidate retrieval process is done using WordNet, a manually built lexical resource where the lexicalization information of all concepts is available along with the frequency information of the senses. On the other hand, BabelNet is exploited to retrieve sense candidates for non-

English languages. This process is "sub-optimal" as BabelNet lexicalizations are automatically derived from various resources and the frequency information of the senses is not directly available (Scozzafava et al., 2020). Therefore, the non-English annotations provided by UKB are expected to be of lower quality compared to the English ones. To address this issue, we have designed the LabelGen method to bypass the requirement of running UKB on the non-English side of a bitext representing English.

The LabelGen method is composed of four steps : English word sense disambiguation, label propagation, multilingual-processing and knowledge-based filtering. Figure 4.2 summarizes our approach.

4.2.1 English Word Sense Disambiguation

We employ the same unsupervised WSD system as in section 4.1.1, UKB enhanced with SyntagNet, on only the English side of the bitext to produce a English sense-annotated corpus. This allows UKB to solely depend on the much reliable WordNet for lexicalization of concepts and sense frequency information.

4.2.2 Label Propagation

To annotate the non-English side of the bitext, we first word-align it using BabAlign and consider the aligned word or phrase of each sense-annotated English token as its translation. Then, we annotate each translation with the same BabelNet synset as the aligned English word. In fact, we propagate the confidence scores of each sense of the annotated English word to the other side, which are used by the subsequent post-processing step. This procedure is based on the assumption that lexical translations are semantically equivalent, and therefore express the same concept (Hauer and Kondrak, 2020). At the end of this step, we have sense-annotations in each of the two languages of the bitext.

	LabelSync			LabelGen		
	Annotated Tokens	Annotated Word Types	Sense Types	Annotated Tokens	Annotated Word Types	Sense Types
EN	1,783,334	9,509	16,748	2,447,676	9,850	18,195
IT	2,083,741	10,910	22,211	1,372,876	8,355	16,046
ES	1,692,232	10,549	25,181	1,326,244	7,926	17,335
FR	1,458,588	7,776	11,529	1,433,647	7,712	17,980
DE	645,289	2,139	2,756	821,552	6,589	9,121

Table 4.1: Corpora Statistics

4.2.3 Multi-Lingual Post-Processing

At this step, we re-rank the senses on each side of the bitext individually using the SoftConstraint method, as described in section 4.1.2. We apply the same settings as before, taking into account the probability scores provided by UKB and translation information from BabelNet.

4.2.4 Knowledge-Based Filtering

At this stage, we may have a non-English word annotated with a BabelNet synset which does not actually contain that word. These invalid sense annotations may occur due to non-literal translation (i.e. the word and its translation do not express the same concept), errors in translation or alignment, or omissions in BabelNet.

Therefore, we apply a BabelNet based filtering method to the non-English side to enforce the following constraint : a word should only be annotated with a synset that contains the word. Following Barba et al. (2020), we discard the annotation if a word is annotated with a synset of which it is not an element.

4.3 Evaluation

Following prior work, we extrinsically evaluate our corpus construction approaches by providing the generated annotations as training data for reference WSD systems, which are then evaluated on standard benchmark datasets.

4.3.1 Reference WSD Systems

We perform experiments with two reference supervised WSD systems. For each of these reference systems, we train models on sense annotations produced by our LabelSync and LabelGen methods. The hyperparameters of each system are held constant throughout all experiments.

The first reference WSD system is a transformer-based method, built on multilingual BERT (Devlin et al., 2018), as described by Barba et al. (2020). We refer to this system as *mBERT*. We use the default parameter settings and number of training epochs.[1] Following prior work, we use the SemEval-2007 dataset (Raganato et al., 2017) as our validation set for the English experiments. Due to the lack of standard validation sets for non-English languages, we use random samples of 1000 sentences from our training corpora.

The second reference WSD system is IMS (Zhong and Ng, 2010), an SVM-based model that relies on multiple lexical features. In cases where the system fails to provide a prediction, we apply the most-frequent-sense (MFS) back-off for English. For all languages, any monosemous words are automatically tagged with their single possible sense.

4.3.2 Test Data

We test the reference WSD models on standard benchmark datasets for WSD. For multilingual experiments, we use the datasets from SemEval-2013 task 12 (Navigli et al., 2013), which contain data for Italian, Spanish, French, and German, and SemEval-2015 task 13 (Moro and Navigli, 2015), which covers Italian and Spanish. The SemEval-2013 datasets contain only nominal instances, while the SemEval-2015 datasets cover nouns, verbs, adjectives, and adverbs. We use the latest version of the datasets[2], which are annotated with synsets from BabelNet version 4.0.

For the experiments on English, we use the five standardised test sets provided by Raganato et al. (2017)[3]. This dataset contains five English all-

[1] https://github.com/edobobo/transformers-wsd
[2] https://github.com/SapienzaNLP/mwsd-datasets
[3] http://nlp.uniroma1.it/wsdeval

words test sets from five shared tasks: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015). We also report the results on the concatenation of all five test sets. We refer to this result as "ALL".

4.3.3 Experimental Setup

For both of our approaches, we employ UKB (Agirre et al., 2014) to perfrom the initial labelling of a bitext. UKB operates by applying the personalized page rank (PPR) algorithm (Haveliwala, 2003) on a lexical knowledge base (LKB). Following Maru et al. (2019), we apply WordNet as the LKB, further enriching it with information from WordNet Gloss Corpus (WNG) (Langone et al., 2004), and syntagmatic information from SyntagNet. BabelNet is utilized as the source of multilingual lexicalization information for the LabelSync approach.

Both LabelSync and LabelGen approaches take an unannotated bitext as input. We randomly sample 200k sentences with English, French, German, Italian, and Spanish translations from EuroSense (Bovi et al., 2017), discarding its existing sense annotations. Thus, we create a 200k sentence five-language parallel corpus, from which bitext slices are provided as input to our methods. The SoftConstraint method employed to refine the initial annotations requires lexical translations. Translations in all four languages are considered for this purpose. Table 4.1 presents the statistics of the LabelSync and LabelGen corpora [4].

4.3.4 Comparison Systems

For comparisons, we have selected approaches that are entirely unsupervised, i.e., independent of any manual sense annotations. We compare our results against three unsupervised approaches. Our direct competitor is OneSeC (Scarlini et al., 2019), which automatically produces sense-annotated data by

[4] A small sample of word types, covering the test sets were annotated to limit the running time of UKB.

	SemEval-2013				SemEval-2015		
Model	IT	ES	FR	DE	IT	ES	AVG
MCS	44.2	37.1	53.2	70.2	44.6	39.6	48.2
UKB+SyntagNet	72.1	74.1	70.3	76.4	69.0	63.4	70.9
SensEmBERT	69.8	73.4	77.8	**79.2**	-	-	-
OneSeC	63.5	61.6	65.1	75.8	-	-	-
LabelSync	75.7	78.2	72.4	75.3	**70.8**	**66.3**	73.1
LabelGen	**77.8**	**80.5**	**80.7**	75.4	68.7	66.1	**74.9**

Table 4.2: mBERT F score (%) comparison of LabelSync and LabelGen corpora against other competitors on standard multilingual WSD datasets.

leveraging the semantic information within Wikipedia categories. Since the resulting multilingual corpus covers only nominal instances, we are unable to test OneSeC on the SemEval-2015 datasets. We also compare against two other unsupervised WSD systems: UKB with SyntagNet (Maru et al., 2019) and SensEmBERT (Scarlini et al., 2020a).

4.3.5 Multilingual Results

Table 4.2 presents the multilingual WSD results obtained with mBERT trained on various corpora, along with reported results of UKB + SyntagNet and SensEmBERT, and also the baseline results obtained by using the most common sense (MCS) as the prediction. We have included the macro average F scores across all sets. With the consistent exception of German, the results of mBERT trained on the corpora produced by LabelSync and LabelGen are substantially better than those obtained using the corpus generated by OneSeC, which is the previous state-of-the-art for unsupervised corpora tagging. Unlike OneSeC, LabelSync and LabelGen are applicable to all parts of speech, and can therefore be applied to the SemEval 2015 datasets.

As expected, the LabelGen method achieves better multilingual results on average than LabelSync, as it influences the non-English side with sense information from WordNet, which provides high-quality coverage on English. LabelGen also outperforms both knowledge-based WSD systems, UKB + SyntagNet and SensEmBERT.

	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15	ALL
MFS	72.1	72.0	65.4	63.0	66.3	67.6
OneSeC	74.2	67.1	62.9	68.8	74.2	70.2
LabelSync	76.8	70.8	66.0	71.4	**75.7**	73.0
LabelGen	76.4	70.7	**67.9**	**71.6**	73.6	72.7
SemCor	**79.7**	**75.4**	**67.9**	69.9	75.0	**74.0**

Table 4.3: English WSD F-score (%) results on nominal instances obtained with mBERT trained on various corpora.

	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15	ALL
MFS	72.1	72.0	65.4	63.0	66.3	67.6
OneSeC	73.2	68.2	63.5	66.5	70.8	69.0
LabelSync	76.1	70.0	68.6	**71.7**	**72.1**	**72.3**
LabelGen	75.8	71.0	**69.2**	70.6	69.7	71.8
SemCor	**76.8**	**73.8**	67.3	65.5	66.1	70.4

Table 4.4: English WSD F-score (%) results on nominal instances obtained with IMS trained on various corpora.

The domain of EuroSense, which consists of parliamentary proceedings, may have a negative impact on our results. For example, the SemEval-2013 datasets for French and German contain 123 and 79 polysemous lemmas, respectively, that are not present in EuroSense. We posit that these multilingual results could be improved by annotating a corpus with broader domain coverage, or by matching the domain of the source corpus to the domain of the data to be disambiguated. We leave this as a direction for future work.

4.3.6 English Results

We have conducted noun-only experiments for English to facilitate a fair comparison against our primary competitor, OneSec. Table 4.3 and 4.4 present the WSD results on English nominal instances obtained with mBERT and IMS respectively. The most frequent sense (MFS) baseline results along with results obtained using SemCor are also listed for comparison. For both mBERT and IMS, LabelSync and LabelGen again are clearly better than OneSeC across all five test datasets.

As our approaches are independent of parts-of-speech, we have also con-

	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15	ALL
MFS	66.8	66.2	55.2	63.0	67.8	65.2
EuroSense+SC	-	-	-	66.4	69.5	-
LabelSync	69.4	64.5	57.4	**71.7**	**72.9**	**68.4**
LabelGen	68.9	65.7	57.8	70.6	71.1	68.1
SemCor	**71.3**	**69.1**	**61.5**	65.1	68.3	68.3

Table 4.5: English WSD F-score (%) results on all instances obtained with IMS trained on various corpora.

ducted English all words WSD evaluation on the standard test sets. We have used IMS as the reference system to facilitate comparison against the reported results obtained using the annotations of our base corpus, EuroSense. The results are presented in table 4.5. Both of our approaches outperform the results obtained by training EuroSense on top of SemCor.

In all the English experiments, we have included results obtained by training on SemCor, a manually tagged corpus. It is remarkable that the corpora generated in an unsupervised manner by LabelSync and LabelGen yield results that are comparable, and in some cases better on average than those obtained by training the same reference system directly on SemCor. These results not only confirm the quality of our approaches, but is surprising and impressive in itself, considering that no manual annotation is used in our corpus construction procedure, and therefore not subject to the knowledge acquisition bottleneck.

We conclude that our unsupervised approach to corpus annotation, which rivals the performance obtained with a manually sense annotated corpus, represents a step towards overcoming the knowledge acquisition bottleneck.

Chapter 5

Using Translations to Predict Cross-Lingual Entailment

In this chapter, we focus on predicting cross lingual binary lexical entailment, which was introduced as the task of "detecting whether the meaning of a word in one language can be inferred from the meaning of a word in another language" (Vyas and Carpuat, 2016). Our principal objective is to provide evidence for the hypothesis that translations are useful in predicting cross-lingual entailment. For example, from the English phrase "you gave me the bottle", and its Italian translation "mi hai dato il contenitore", it can be inferred that *bottle* entails *contenitore* ("container") [1]. We are interested in leveraging this phenomenon to perform unsupervised LE prediction.

5.1 Methods

In this section, we provide detailed descriptions of our translation based approaches to predict cross-lingual LE.

5.1.1 Entailment via Alignment

Our baseline method, which we call the BiText method, is based on the intuition that translations may result in more general concepts. We leverage automatic word alignment of raw, sentence-aligned bitexts to mine translation pairs. We make the following assumption: a word and its translation in a

[1] This is an example from the OpenSubtitles corpus (Lison and Tiedemann, 2016).

corpus either (a) represent the same concept, or (b) represent two distinct concepts, one entailing the other. We expect that in most cases the relation between aligned word pairs is synonymy or equivalence, but a subset of the word pairs may involve hypernymy instead. The direction of such entailments is an open question.

Based on the above intuition, we developed the following method: given a sentence-aligned bitext representing the languages on which the prediction is to be performed, we extract translation pairs by performing word-alignment of the corpus. Given a test instance, we simply check whether the two words are among the aligned pairs to obtain a binary classification.

5.1.2 Semantic Expansion

The coverage of our baseline method is constrained by the coverage of translation pairs in the aligned bitext. For example, if *ankle* is never translated as *gelenk* ("joint") in our English-German bitext, the BiText method will fail to identify the entailment between the two words. However, the bitext may contain another word, such as *knee*, that is aligned to *gelenk* and semantically similar to *ankle*. This motivates the development of our second method, which we refer to as the Vectors method.

The general intuition behind this method is that semantically similar words tend to share entailments. This method requires an automated way of measuring semantic similarity. To this end, We use the well known technique of computing the cosine similarity of monolingual word embeddings. These embeddings are trained independently on each side of the bitext. If the cosine similarity of two words is not less than a tunable threshold, the words are deemed to be semantically similar. Note that the search for similar words is performed only with respect to the first of the words in a given instance, which may entail the second word.

5.2 Experiments

In this section, we describe the details of our experiments, including the dataset, tools, settings and results. We focus primarily on the English-German subtask.

5.2.1 Dataset

For our experiments, we use the datasets provided by SemEval (Glavaš et al., 2020). The English-German set contains 75, 418 and 2,149 trial, development and test instances respectively. We utilize the trial set for tuning purposes.

The trial and development sets provided by SemEval have three columns in each line. The first column denotes *concept_1*, which is the first concept in the ordered pair. The second column denotes *concept_2*, which is the second concept in the ordered pair. Both *concept_1* and *concept_2* come with language prefixes. The third column denotes *gold score*, which is 0 or 1 in case of binary entailment. The test set has two columns in each line, denoting *concept_1* and *concept_2*.

5.2.2 Tools and Resources

Our bitexts are from the OpenSubtitles project. The English-German corpus has 22.5M aligned sentence pairs. We lower-case all text, and tokenize by white space and punctuation. We employ TreeTagger (Schmid, 1999, 2013), a modified ngram tagger, to lemmatize the corpus.

Both of our methods are dependent on alignment accuracy. Therefore, in addition to FastAlign, we also employ the more accurate BabAlign, to word-align the bitexts,

For the purpose of computing word similarity in the Vectors method, we generate word embeddings using the skip-gram model of word2vec (Mikolov et al., 2013). We set the vector dimensions to 200, the context window size to 10, and run word2vec for 25 iterations. All other parameters affecting the vectors are left at their default values.

Method	Alignment	Setting	Trial	Dev	Test
BiText	FastAlign	LR	17.4	20.1	24.7
	FastAlign	HR	17.0	29.0	31.2
	BabAlign	HR	49.2	49.5	52.4
Vectors	FastAlign	LR	62.9	60.1	63.1
	FastAlign	HR	61.0	61.7	65.0
	BabAlign	HR	**71.0**	**65.6**	**70.7**

Table 5.1: F-score (%) on the English-German trial, development and test sets.

5.2.3 Experimental Setup

We perform experiments in two settings: low-resource (LR) and high-resource (HR). In the LR setting, the bitext is restricted to randomly selected 1M sentence pairs, and no lemmatization is used. In the HR setting, we utilize the full bitext and also lemmatize the corpus before alignment. The knowledge-based alignment method BabAlign can only be applied in the HR setting, as BabelNet contains only lemmas and no lemmatization is done for the low resource setting.

5.2.4 Results

Table 5.1 shows the English-German results on the trial, development and test sets. Our complete system is the Vectors method combined with the knowledge-based alignment BabAlign, which demonstrates the best results in all three sets. The incorporation of word embeddings in the Vectors method yields significant improvement over the baseline BiText method. Also, the knowledge based alignment method BabAlign clearly outperforms the standard alignment tool FastAlign for both BiText and Vectors methods.

5.3 Analysis

The Vectors method is an expansion of the BiText method. For any test instance, if BiText returns a positive classification, then Vectors does so as well. Thus the set of entailment relations reported by the former is a subset of the entailment relations reported by the latter. Consequently,

VECTORS reduces the number of false negatives, at the cost of a higher number of false positives. Overall, the result is a substantial net gain over the baseline BITEXT method in LE prediction accuracy. For example, in the English-German development set in the LR setting, the precision drops from 60.5% to 48.7%, but the recall increases from 12.0% to 78.5%.

One weakness of our methods is the inability to distinguish the direction of an entailment relation. This can lead to false negatives, an issue which the VECTORS method sometimes exacerbates. For instance, in the English-Italian bitext, Italian *creatura* ("creature") is not aligned with English *wolf*. However, VECTORS incorrectly predicts an entailment, because *creatura* and *animale* ("animal") are semantically similar, and *animale* is found to be aligned with *wolf*.

Another source of errors are non-literal translations. For example, the English phrase *automobile key* is translated into Italian as *chiave di accensione* ("ignition key"). This leads to the incorrect conclusion that *automobile* is entailed by *accensione* ("ignition") and similar words.

Overall our results constitute a proof-of-concept and demonstrate a strong connection between translations and cross-lingual entailment.

Chapter 6

Conclusion

In this thesis, we have successfully exploited two different properties of word translations to address important tasks in lexical semantics. Based on translational equivalence, we have proposed novel algorithms for correcting sense annotations on a bitext and achieved statistically significant improvements over the results obtained by the original corpora. We have also introduced new approaches to address the problem of knowledge acquisition bottleneck in word sense disambiguation in an unsupervised manner. Our methods are largely language-independent, and hence we have evaluated them in both English and multilingual WSD settings. Results demonstrate that our pipelines for automatic sense tagging, LabelSync and LabelGen, can produce annotated corpora for arbitrary languages and domains, which approach the quality of manual annotations.

Furthermore, using the generalization property of translations, we have developed effective methods of detecting cross-lingual lexical entailment. In our BiText method, we alleviate the sparsity of the translation data by leveraging semantic similarity between word embeddings.

There are several directions of further research related to our approaches. We have only applied our corpus labelling methods to a parallel corpus from the political domain. We plan to investigate how the results would change if we include a corpus from a different domain, such as movie subtitles from the OpenSubtitles corpus. Also, it would be interesting to investigate how our pipelines would perform beyond the standard WSD test sets. To this end, we

would like to approach domain specific WSD, by applying our methods to automatically produce WSD training data from a specific domain. Furthermore, we would like to investigate the impact of our annotation corrections algorithms, MultiWordNet and Bipartite, when integrated into the pipelines.

It may appear to the reader that our corpus labelling approaches are constrained by the availability of bitexts. However, with recent advances in machine translation (MT) technology, high quality parallel texts can be produced given a corpus from a single language. Hence, instead of using a manually translated bitext, it would be interesting to investigate how the results would change if we used translations from a state-of-the-art MT model. This modification could increase the generalizability of our methods, as we would then only require text from a single language as input.

References

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

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proc. of IJCAI*, pages 3837–3844.

Luisa Bentivogli and Emanuele Pianta. 2000. Looking for lexical gaps. In *Proceedings of the ninth EURALEX International Congress*, pages 8–12. Stuttgart: Universität Stuttgart.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

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.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. Semeval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018); 2018 Jun 5-6; New Orleans, LA. Stroudsburg (PA): ACL; 2018. p. 712–24.* ACL (Association for Computational Linguistics).

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *AAAI*, volume 5, pages 1037–1042.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

Goran Glavaš, Ivan Vulić, Anna Korhonen, and Simone Paolo Ponzetto. 2020. Semeval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 24–35.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. Ualberta at semeval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269.

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

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021. Semi-supervised and unsupervised sense annotation via translations. In *Submission*.

Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Helen Langone, Benjamin R Haskell, and George A Miller. 2004. Annotating wordnet. Technical report, PRINCETON UNIV NJ COGNITIVE SCIENCE LAB.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

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.

Arnob Mallik and Grzegorz Kondrak. 2021. Correcting sense annotations using translations. In *Submission*.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. Syntagnet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3525–3531.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. Approaching plwordnet 2.0. In *Proceedings of the 6th Global Wordnet Conference, Matsue, Japan*, pages 189–196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

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.

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

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

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.

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.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 455–462.

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-20, Yokohama, Japan*.

Tommaso Pasini and Roberto Navigli. 2020. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103215.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 87–92.

Wei Qiu, Mosha Chen, Linlin Li, and Luo Si. 2018. Nlp_hz at semeval-2018 task 9: a nearest neighbor approach. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 909–913.

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.

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

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping polish wordnet onto princeton wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048.

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.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Helmut Schmid. 2013. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154.

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

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 338–344.

Yogarshi Vyas and Marine Carpuat. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1187–1197.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational linguistics*, 35(3):435–461.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 system demonstrations*, pages 78–83.