Leveraging Translations for Lexical Semantics

by

Hongchang Bao

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

Master of Science

Department of Computing Science

University of Alberta

Abstract

Computational lexical semantics is the study of word meanings which involves algorithms and ontologies. Computation of semantic similarity plays an important role in various applications of natural language processing, including information retrieval, machine translation, and question answering. Most prior work on computing meaning similarity focus on sense definitions or the relational structure of lexical resources. Lexical translations constitute another important component in lexical resources such as BabelNet and CLICS. In this thesis, we explore the idea of leveraging multilingual translations to compute semantic similarity. In particular: (1) we posit and investigate the hypothesis that there are no universal colexifications, (2) propose an algorithm to align concepts across lexical resources, and (3) develop novel approaches to detect sense synonymy across different contexts. Our results in these three tasks confirm the utility of translations in computational lexical semantics.

Preface

Chapters 3, 4, and 5 in this thesis are adapted from the following research papers: Bao et al. (2021), Bao et al. (2022), and Hauer et al. (2021), respectively. The three papers were written collaboratively. I implemented all methods, and performed all experiments that are described in this thesis. Parts of the papers have been included throughout the thesis to provide relevant contextual information consistent with the content of those papers.

Acknowledgements

I would first like to thank my supervisor, Professor Greg Kondrak, for his support and guidance throughout this thesis. I would also like to thank Bradley Hauer for his contributions and assistance to this thesis.

I would like to thank Arnob Mallik for his contributions to the Chapter 5 of this thesis.

Thanks to Yixing Luan and Amir Ahmad Habibi for their suggestions and research advice.

This thesis was completed with the funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

Contents

1 Introduction **1**
- 1.1 Background . . . 3
- 1.2 Colexification Hypothesis . . . 5
- 1.3 Mapping Lexical Resources . . . 6
- 1.4 Detecting Sense Synonymy . . . 7
- 1.5 Outline . . . 7

2 Resources **9**
- 2.1 WordNet . . . 9
- 2.2 BabelNet . . . 10
- 2.3 Open Multilingual WordNet . . . 10
- 2.4 CLICS . . . 11
- 2.5 OmegaWiki . . . 11

3 On Universal Colexifications **13**
- 3.1 Related Work . . . 14
- 3.2 Colexification . . . 15
  - 3.2.1 Formalization . . . 15
  - 3.2.2 Hypothesis . . . 15
- 3.3 Method . . . 16
- 3.4 Results . . . 18
  - 3.4.1 Analysis . . . 18

4 Lexical Resource Mapping via Translations **21**
- 4.1 Related Work . . . 22
- 4.2 Methods . . . 24
- 4.3 Experiments . . . 26
  - 4.3.1 Comparison Methods . . . 26
  - 4.3.2 Evaluation Measures . . . 27
  - 4.3.3 Language Selection . . . 28
  - 4.3.4 Aligning WordNet and CLICS . . . 29
  - 4.3.5 Aligning WordNet and OmegaWiki . . . 31

5 Determining Sense Synonymy via Translations **34**
- 5.1 Related Work . . . 35
- 5.2 Theoretical Solution . . . 36
  - 5.2.1 Substitution Test . . . 37
  - 5.2.2 Translation Criss-Cross . . . 38
  - 5.2.3 Multi-Synset Intersection . . . 39
- 5.3 Methods . . . 40
  - 5.3.1 IDENT and CVAL . . . 40
  - 5.3.2 Synonymy Check . . . 40
  - 5.3.3 SUB and CSUB . . . 41

5.4 Experiments 41
5.4.1 Translation and Lemmatization 42
5.4.2 Word Alignment 42
5.4.3 Contextual Embeddings 43
5.4.4 Development Results 43
5.4.5 Test Results and Discussion 44

6 Conclusion **46**

References **48**

List of Tables

2.1 Statistics of the lexical resources. 10

3.1 The empirical validation of our hypothesis. 18
3.2 The concept pairs colexified by the most languages. 19
3.3 The concept pairs with the ratio of 1 represent possible exceptions to our hypothesis. 20

4.1 Results of the LangVote method on the development set for aligning WordNet and CLICS. 29
4.2 Results on the test set for aligning WordNet and CLICS. . . . 30
4.3 Results on the test set for aligning WordNet and OmegaWiki. 32

5.1 Accuracy on the development set with different methods and languages of translation 43
5.2 Accuracy on the test set with different methods and languages of translation. 45
5.3 Detailed breakdown of the results of our best performing method. 45

List of Figures

1.1 Three synsets for the adjectives *nascent* and *right* in the WordNet 3
1.2 An example of a multi-synset 4
1.3 Three concepts that are colexified in Persian, English, and Chinese. 5

4.1 An example of concept lexicalization overlaps. 24

5.1 An example of Translation Criss-Cross. 37

Glossary

colexification
: The phenomenon that multiple concepts in the same language can be expressed by a single word.

concept
: A discrete meaning that can be expressed by at least one word.

contextual embeddings
: Real-valued vectors that are used to represent word meaning in context.

embeddings
: Real-valued vectors that are used to represent word meaning.

gloss
: The definition of a synset or a concept in a lexical resource.

hypernymy
: The converse of hyponymy.

hyponymy
: A semantic relation between a subtype (hyponym) and a supertype (hypernym).

sense
: A discrete meaning that a word can have. (Each sense of a given word corresponds to a different concept.)

synonyms
: Words that can have the same meaning.

synonymy
: The relation of sameness of meanings.

synset
: A set of all words that can express a given concept.

translational equivalents
: Words in different languages that can be mutual translations.

word sense disambiguation (WSD)
: The task of tagging a word in context with its sense.

Chapter 1

Introduction

Computational lexical semantics is the study of word (including non-compositional expressions, such as "single out") meanings which involves algorithms. It is common for a word to have multiple meanings. For example, the word *bat* can refer to *a nocturnal mouse-like mammal with wings* or *a piece of wood used for hitting the ball in various games*. Lexical semantics addresses the problem of identifying the meaning of a word in context, such as the word *bat* in the sentence: "The bat is drinking from an agave flower".

Accurately capturing word meanings plays a crucial role in natural language processing (NLP). First, words can be ambiguous, which makes interpretation more difficult. For example, is someone who types *bat* in a search engine looking for a mammal or a racket? Second, although it is easy for a human being to identify the meaning of the word *bat* in the above sentence, this is a complicated task for computers, as they need to process and analyze textual information before determining the underlying meaning. This task of tagging a word in a given context with its meaning chosen from a lexical resource is known as word sense disambiguation, which is one of the central problems in natural language processing (Navigli, 2018), and can be used to improve numerous applications, such as information retrieval, machine translation, and question answering.

In order to capture word meanings, a number of lexical-resource-based approaches have attempted to compute semantic similarity. These methods mainly leverage two different types of information in the lexical resources:

textual definitions (glosses) and the structure of lexical resources. Textual definitions are usually used for computing meaning similarity. These methods (Meyer and Gurevych, 2011; Navigli and Ponzetto, 2012) apply similarity measures on pairs of glosses and identify the most similar meaning by maximizing the similarity. Exploring the graph structure of lexical resources is another line of research for detecting semantic similarity. Most lexical resources can be viewed as graphs in which vertices represent word meanings, and edges represent relations between them. These methods (Pilehvar and Navigli, 2014, 2015) apply graph algorithms, such as Personalized PageRank algorithm, to the structure of lexical resources to measure meaning similarity. In addition to these two kinds of approaches, it is also possible to use translations to measure semantic similarity, since different meanings of a word are translated differently.

Translation information plays a significant role in NLP. One issue when applying NLP techniques to the same task in multiple languages is that people need to repeat annotating the data (i.e. providing training data) for each new language (Navigli et al., 2021). The annotation activity is time consuming and expensive. Moreover, if the data in some language is unavailable, then the annotation processing will be more difficult. Here, translation information can be used to generate the annotated data automatically in different languages. In addition to that, translation information has also been leveraged to improve the performance of word sense disambiguation systems (Luan et al., 2020).

In this thesis, we explore the idea of leveraging translations to compute semantic similarity, using sets of lexical translations from different languages. In particular, *we demonstrate that multilingual translations extracted from lexical resources can be leveraged to improve the accuracy on three semantic tasks: identifying universal colexifications, aligning concepts between lexical resources, and detecting sense synonymy.* Our results in these three tasks confirm the utility of translations in lexical semantics.

The rest of this chapter is structured as follows: We first provide some background knowledge related to this thesis. Then, we briefly discuss the three tasks that are the main contributions in this thesis. Finally, we provide

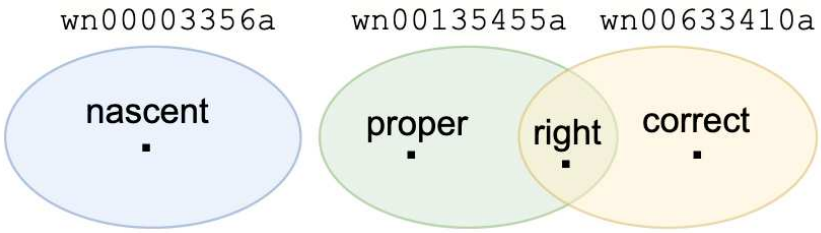

Figure 1.1: Three synsets for the adjectives *nascent* and *right* in the WordNet.

an outline of the rest of this thesis.

1.1 Background

In this section, we describe the background knowledge which contextualizes this thesis. In particular, we discuss the wordnets, synset, synset properties, and multilingual wordnets.

Wordnets, such as Princeton WordNet (Fellbaum, 1998), are lexical resources composed of *synsets*. A synset is a set of synonymous words, and can be used to represent a specific *lexicalized concept*, or simply concept (Miller, 1995). For example, Figure 1.1 shows three synsets where the synset wn00633410a contains {*right*, *correct*} and represents the corresponding concept CORRECT. A word *lexifies* a concept if it can be used to express that concept; that is, if the corresponding synset contains that word. If two words in the same language lexify a single concept, such as *right* and *correct*, the words are synonyms. Each content word in a synset lexifies at least one concept. Each word and concept pair that it lexifies corresponds to a unique sense of that word. A word *sense* is a partition of the word meanings (Kilgarriff, 1997). As shown in Figure 1.1, the word *right* has two senses "appropriate for a condition or purpose or occasion or a person's character, needs" and "correct in opinion or judgment".

Wordnets are central to our work, and synsets are the basic units of its ontology. We list the following synset properties (Hauer and Kondrak, 2020b) which are used in this thesis.

1. A word is monosemous iff it is in a single synset. A word is polysemous

Figure 1.2: A multi-synset representing the concept CORRECT, with words from English, Chinese, French, Russian, and Japanese.

iff it is in multiple synsets. For example, in Figure 1.1, the adjective *nascent* is monosemous because only one WordNet synset contains this word (we treat different part of speech as different words). However, the adjective *right* is polysemous as it is shared by two synsets.

2. Word senses are synonymous iff they are in the same synset. As shown in Figure 1.1, the word senses of *right* and *correct* are synonymous.

3. Every sense of a polysemous word belongs to a different synset. Figure 1.1 demonstrates that the adjective *right* has two senses and each of them is in a seperate synset.

Multilingual wordnets (multi-wordnets) such as BabelNet (Navigli and Ponzetto, 2012) consist of multilingual synsets (multi-synsets), which contain words in many languages (Figure 1.2), each lexicalizing the concept that corresponds to that multi-synset. Multi-wordnets may be constructed by adding translations to the monolingual synsets of a pre-existing wordnet, typically WordNet itself, or by linking the synsets of multiple independently constructed wordnets in different languages. If two words in different languages lexify a single concept, such as the English word *right* and the French word *bon* (Figure 1.2), the words are translational equivalents. Synonymy and translational equivalence are respectively the intra-lingual and inter-lingual components of the relation of *semantic equivalence*, or sameness of meaning (Hauer and Kondrak, 2020b).

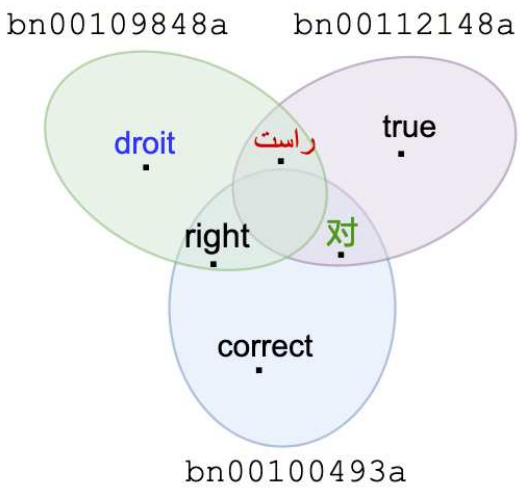

Figure 1.3: Three concepts that are colexified in Persian, English, and Chinese.

1.2 Colexification Hypothesis

Colexification refers to the phenomenon of multiple concepts in the same language being lexified by a single word (François, 2008). If two concepts are referred to by a single word, the concepts are *colexified* by that word. For example, the English word *right* colexifies the concepts of RIGHT (side) and CORRECT (Figure 1.3). A language colexifies two concepts if it contains a word which colexifies them. For example, English colexifies the concepts RIGHT and CORRECT; Chinese does not.

In this thesis, we posit and investigate the hypothesis that *there are no universal colexifications*, or more precisely, that *no two distinct concepts are colexified in every language*. This hypothesis is relevant to the task of word sense disambiguation, because it would imply that any sense distinction in any language could be disambiguated by translation into some language.

We test our hypothesis by analyzing the colexification data from three different lexical resources. The results show that our hypothesis is strongly supported by the colexified concept pairs in those lexical resources.

1.3 Mapping Lexical Resources

Lexical resources are indispensable in many areas of NLP. However, lexical resources vary in how they are constructed, and in how they represent concepts and senses, which makes it difficult to combine information from multiple resources. The task that we address is mapping (or aligning) concepts or senses across lexical resources. Given a concept in one of the resources, such a mapping allows us to identify an equivalent concept in the other resource. For example, we map the direction related concept RIGHT in one resource to its corresponding concept in the other resource.

Aligning concepts between lexical resources has several benefits. First, lexical resources provide complementary knowledge but lack links between them. Aligning concepts between them enables lexical resources connected to each other. Second, combining information from multiple resources increases the total knowledge available about each concept. Third, Inter-resource concept mapping has been shown to yield performance improvements compared to using resources in isolation (Ponzetto and Navigli, 2010).

Most prior methods on concept mapping are based either on similarity measures between pairs of concepts across resources, or graph algorithms that create or exploit the structure of lexical resources. These methods leverage the observation that glosses and semantic relations are the two commonly-used ways to describe word senses in lexical resources.

In this thesis, we propose two novel translation-based methods that leverage sets of lexicalizations from different languages to distinguish and align concepts. Our methods depend exclusively on lexicalization information, without relying on concept glosses, relations between concepts, or other structured information.

We evaluate our approaches on the alignment of WordNet with two other lexical resources: CLICS and OmegaWiki. In both cases, our methods match or exceed the accuracy of the best comparable methods from prior work. To our knowledge, ours is the first time to align the full CLICS concepts with WordNet synsets. Therefore, we release the alignment we produced to facili-

tate further work on this important task.

1.4 Detecting Sense Synonymy

Given a pair of sentences that share a *focus word* in common, the task of sense synonymy detection is to decide whether the *focus word* has the same meaning in both sentences. For example, the word *right* in the two sentences below conveys different meanings:

- It is not *right* to leave the party without saying goodbye.
- Most people write with their *right* hand.

Prior approaches on this task can be roughly divided into two types, which are based on vectorized representation of words, i.e. embedding, or word sense disambiguation (WSD) systems. The embedding based method is first to obtain different representations of the focus words, and then measure the similarity between these embeddings. The WSD-based method is to employ a word sense disambiguation system to first predict the senses of the target words, and then make a decision based on the prediction.

In this thesis, we propose translation-based methods to investigate whether translations can be used to detect semantic equivalence in context. Our methods combine elements of both types. We employ embeddings in our methods. However, we take the embeddings of the translations of the focus words instead of the focus words themselves. Similar to WSD based methods, our methods also analyze the common synsets of the focus words and their translations, with the goal of identifying a probable shared synset.

We evaluate our methods on a standard shared task dataset. The results provide a solid proof-of-concept for the utility of multilingual translation in detecting sense synonymy.

1.5 Outline

The thesis is structured as follows. In the Glossary, we list short definitions of the terms used in this thesis. In Chapter 2, we describe the lexical re-

sources used in this thesis. In Chapter 3, we review the prior work on colexifications, and then formalize our hypothesis and subsequently present our method, results, and analysis. In Chapter 4, we describe our two mapping algorithms, which can be applied to any pair of lexical resources with multilingual lexicalization information. Afterwards, in Chapter 5, we introduce our four translation-based methods on sense synonymy detection, and present our experiment settings and results. Chapter 6 concludes the thesis.

Chapter 2

Resources

In this chapter, we describe the lexical resources used in this thesis. Each resource consists of a set of concepts; each of which is associated with a set of lexicalizations, that is, words that can express the concept. Table 2.1 lists some statistics for each resource. The resources are diverse: CLICS contains data from over 3000 languages, but only about 450 words per language, on average. Contrariwise, BabelNet includes almost 2,848,500 words per language, but covers only 284 languages.

2.1 WordNet

Princeton WordNet (Fellbaum, 1998), or simply WordNet, is the first large-scale English lexical resource. It was manually created at the Princeton University and contains nouns, adjectives, verbs, and adverbs which are grouped into different synsets (sets of synonymous words), i.e., the set of words that share the same meaning. These synsets are connected with each other through semantic relations, such as hypernymy and hyponymy. WordNet was constructed according to the principle that each synset should consist of words which are interchangeable in some context without altering the meaning of the expression. Each synset is associated with a gloss describing its concept, a part of speech (noun, verb, adjective, or adverb), and, optionally, one or more usage examples. Lexical resources use different conventions to refer to concepts. In WordNet, a synset is typically referred by one of its lexicalizations, along with the part of speech and a number. For example, the synset

	WordNet	BabelNet	OMWN	CLICS	OmegaWiki
Languages	1	284	34	3050	1053
Concepts	117,659	6,113,467	117,659	2919	51,207
Lexicalizations	206,941	808,974,108	1,950,401	1,377,282	248,166

Table 2.1: Statistics of the lexical resources.

$play_n^1$ contains the nouns {*play*, *drama*, *dramatic play*}, and has the gloss "a dramatic work intended for performance by actors on a stage", and the usage example "he wrote several plays but only one was produced on Broadway". We used WordNet version 3.0 and accessed it through NLTK API[1] in Python.

2.2 BabelNet

BabelNet (Navigli and Ponzetto, 2012) is a multilingual wordnet, automatically constructed by adding translations to the monolingual synsets of the Princeton WordNet. BabelNet combines data from Wikipedia, Wikidata, and various other resources, supplemented by machine translation, to cover nearly 300 distinct languages. Each of the multi-synsets in BabelNet corresponds to a unique concept, with a unique eight-digit identifier, and an associated part of speech (noun, verb, adjective, or adverb); each multi-synset contains one or more words which can express the unique concept in various languages. For example, the multi-synset associated with $play_n^1$ is represented by synset `bn:00028604n` which contains the French word *piece de theatre* and *drame* and the Chinese words *xìjù* and *jùběn*. We used BabelNet version 4.0 and accessed it through its Java API.

2.3 Open Multilingual WordNet

Open Multilingual WordNet (OMWN) (Bond and Foster, 2013) is another multilingual wordnet, constructed by linking wordnets in 34 languages to the Princeton WordNet 3.0. Like BabelNet, OMWN consists of multi-synsets, each containing one or more words from one or more languages which lex-

[1] `https://www.nltk.org/`

ify a particular concept. For example, *sign* and *mark* (English), and *signe*, *témoignage*, *preuve*, and *point* (French) all share a multi-synset. Each multi-synset in OMWN corresponds to exactly one WordNet synset. We accessed OMWN through NLTK API in Python.

2.4 CLICS

The Database of Cross-Linguistic Colexifications (CLICS) (Rzymski and Tresoldi, 2019) is an online lexical database constructed by integrating word lists representing thousands of languages. It contains 2919 concepts, each of which is associated with a unique name, a gloss, and a set of lexicalizations. Different from WordNet, each concept is assigned a single unique name, consisting of one or more English words which concisely describe its meaning. Also unlike WordNet, CLICS does not provide relations between concepts. Each concept is also associated with one of the following categories: "Action/Process", "Number", "Person/Thing", "Property", or "Other". As an example, the concept named TREE has the category "Person/Thing", and contains the English word *tree*, the French word *arbre*, and the Italian word *albero*. The gloss of this concept is "any large woody perennial plant with a distinct trunk giving rise to branches or leaves at some distance from the ground." We extracted the CLICS dataset by following the procedure described in List (2018).

2.5 OmegaWiki

OmegaWiki[2] is an online multilingual dictionary which can be freely edited through its website. Each concept in OmegaWiki is represented by a gloss (called *DefinedMeaning* in OmegaWiki), and associated with words from different languages. For example, one OmegaWiki concept has the gloss "the whole of buildings, machines and necessary devices to carry out an activity" and contains the English words *plant*, *industrial plant* and the French word *site*. Like CLICS, each concept in OmegaWiki is identified by one or more English words. Different from CLICS and WordNet, OmegaWiki glosses are

[2] `http://www.omegawiki.org`

translated into different languages, rather than being in English only. For example, the above concept has the Dutch gloss "Geheel van gebouwen, machines en benodigde hulpmiddelen om een handeling te verrichten." We used the database dump from 16 September, 2021.

Chapter 3

On Universal Colexifications

In this chapter, we posit and investigate the hypothesis that there are no universal colexifications, or more precisely, that *no two distinct concepts are colexified in every language*.

The universal colexification hypothesis is relevant for the task of word sense disambiguation because it would imply that any sense distinction in any language could be disambiguated by translation into some language. It is also related to a famous proposal of Resnik (1997) "to restrict a word sense inventory to those distinctions that are typically lexicalized cross-linguistically". If there are no universal colexifications, then a sense inventory based on cross-lingual translation pairs would also include all core concepts in existing lexical resources, which would cast doubt on the commonly expressed opinion that WordNet is too fine-grained (Pasini and Navigli, 2018).

We test our hypothesis by analyzing the colexification data from three different lexical resources: BabelNet, Open Multilingual WordNet, and CLICS. The results show that our hypothesis is supported by over 99.9% of colexified concept pairs in these three lexical resources.

The structure of this chapter is as follows: In Section 3.1, we summarize previous research related to colexification. In Section 3.2, we formalize the concepts of lexification and colexification, and state our hypothesis. Section 3.3 describes how we construct a colexification database from each of these resources. In Section 3.4, we present the empirical verification of the colexification hypothesis and analyze these results further.

3.1 Related Work

Approaches to colexification can be divided into three types, which are based on semantic maps, graphs, and databases, respectively.

The semantic-map approach to colexification is introduced by Haspelmath (2000), who focuses on distinguishing senses in the grammatical domain. Semantic maps are constructed by cross-linguistic comparison, and contain concepts that have distinct colexifications in at least two different languages. Their experiments show that 12 diverse languages are sufficient to build a stable semantic map. Our hypothesis relates this statement to entire lexicons of core concepts. François (2008) also uses colexification data to build a semantic map for studying the world's lexicons across languages. He observes that the more languages are considered, the more distinctions between senses need to be made. This finding is consistent with our hypothesis, and also raises another open question: is a given pair of colexified concepts colexified universally?

The graph-based approach is introduced by List and Terhalle (2013), who analyze cross-linguistic polysemy. They build a weighted colexification graph using data from 195 languages representing 44 language families, and find that clusters of closely-related or similar concepts are often densely connected. Youn et al. (2016) construct colexification graphs in the domain of natural objects to verify if human conceptual structure is universal. Analysis reveals universality of similar patterns in semantic structure, even across different language families.

The database approach is used by Pericliev (2015), who studies colexifications of 100 basic concepts, and introduces heuristics for distinguishing between homonymy and polysemy. Georgakopoulos et al. (2020) use a colexification database to study commonalities between languages in the domain of perception-cognition. They analyze the colexification of four concepts related to perception (SEE, LOOK, HEAR, and LISTEN) to reveal connections between verbs of vision and hearing.

3.2 Colexification

In this section, we begin by providing a formal treatment of the concepts used in this chapter, inspired by the formalization of homonymy and polysemy of Hauer and Kondrak (2020a). Then, we formally state and discuss our hypothesis.

3.2.1 Formalization

Let $\mathcal{C}$ be the set of all concepts. Let $\mathcal{L}$ be the set of all languages. For each language $E \in \mathcal{L}$, let $\mathcal{V}_E$ be the lexicon of E, the set of all words in E. Further, for each concept $c \in \mathcal{C}$, $w_E(c)$ is the set of words in E which lexify c. If $w_E(c) = \emptyset$, c is a lexical gap in E; that is, no word in E lexifies c. Otherwise, if $w_E(c) \neq \emptyset$, c is lexified in E.

Two concepts $c_1, c_2 \in \mathcal{C}$ are colexified by language E if and only if $w_E(c_1) \cap w_E(c_2) \neq \emptyset$. We define $COL(c_1, c_2)$ as the set of languages that colexify c_1 and c_2, and $LEX(c_1, c_2)$ as the set of languages that lexify both c_1 and c_2:

$$COL(c_1, c_2) = \{E \in \mathcal{L} \mid w_E(c_1) \cap w_E(c_2) \neq \emptyset\}$$

$$LEX(c_1, c_2) = \{E \in \mathcal{L} \mid w_E(c_1) \neq \emptyset \neq w_E(c_2)\}$$

Obviously, $COL(c_1, c_2) \subseteq LEX(c_1, c_2)$.

For the purpose of analyzing colexification, we introduce the *colexification ratio*: for any pair of concepts, their colexification ratio is equal to the number of languages which colexify the concepts divided by the number of languages which lexify both concepts. Formally, we define the colexification ratio between two concepts as:

$$r(c_1, c_2) := \frac{|COL(c_1, c_2)|}{|LEX(c_1, c_2)|}$$

$r(c_1, c_2)$ is undefined if $LEX(c_1, c_2) = \emptyset$.

3.2.2 Hypothesis

We propose the following hypothesis: no pair of concepts is colexified in every language. More precisely, for any pair of concepts that are colexified in some

language, there exists another language that lexifies both concepts but does not colexify them. Formally:

$$\forall c_1, c_2 \in \mathcal{C},\ \exists E \in \mathcal{L}\, s.t.\, w_E(c_1) \cap w_E(c_2) \neq \emptyset$$

$$\Rightarrow \exists F \in \mathcal{L}\, s.t.\, w_F(c_1) \neq \emptyset \neq w_F(c_2)$$

$$\wedge\ w_F(c_1) \cap w_F(c_2) = \emptyset$$

Equivalently, our hypothesis predicts that for every pair of concepts, the colexification ratio is either undefined or less than one:

$$\forall c_1, c_2 \in \mathcal{C},\ |LEX(c_1, c_2)| > 0$$

$$\Rightarrow r(c_1, c_2) < 1$$

This equivalence can be seen by simply substituting r, LEX and COL with the definitions given in Section 3.2.1, and applying some basic principles of set theory.

3.3 Method

We use the following procedure to create a database containing concept pairs and colexification information for each of the three resources: BN, OMWN, and CLICS. As OMWN is evaluated on a set of 5000 *core concepts*, for the purposes of our work, we limit OMWN and BN to their respective 5000 synsets corresponding to these core concepts.

The first step is to extract from each resource the set of concepts it contains, and the set of words lexifying each concept. For CLICS, this is relatively straightforward, as the resource is already structured as a database of concepts and lexifications for each language. Each concept in these resources is represented by a multi-synset, which can be extracted using the corresponding APIs mentioned in Chapter 2.

The second step is to map each of the three sets of concepts to each other, so that identical concepts in distinct resources can be associated with one another for our analysis. This is done by using WordNet 3.0 as a pivot. As described

in Chapter 2, each of the 5000 core concepts in BN and OMWN is already linked to a WordNet 3.0 synset. However, mapping CLICS to WordNet is not trivial because, unlike BN and OMWN multi-synsets, CLICS concepts have no intrinsic connection to WordNet synsets. Therefore, we use a Concepticon mapping created by List et al. (2016) which links a subset of CLICS concepts to WordNet. Unfortunately, the mapping is incomplete, covering only 1372 (47.0%) of CLICS concepts.

The third step is to enumerate all pairs of distinct concepts. There are approximately 4.3 million possible concept pairs in CLICS, and 12.5 million possible concept pairs in BN and OMWN. Although there are millions of concept pairs in each resource, only a subset are lexified by some language (i.e. there exists a language with at least one word for each concept), and only a subset of those are colexified by some language (i.e. there exists a language with a single word for multiple concepts). So, we are working with a subset of a subset of all concept pairs.

The fourth step is to determine which concept pairs are colexified, that is, have words in common. This consists of testing whether the intersection of the corresponding synsets (for BN and OMWN) or the corresponding database entries (for CLICS) are non-empty. We report the number of concept pairs which are colexified in at least one language in Table 3.1. For each pair of concepts, we record the number of languages in our databases that colexify the pair. For example, the CLICS lists 980 languages that lexify both RIGHT (side) and CORRECT. Taking the intersection of the words lexifying each concept, we find that 41 languages have a word which lexifies both concepts, that is, 41 languages colexify these concepts in the CLICS resource. Therefore, the colexification ratio for this concept pair, in CLICS, is $41/980 \approx 0.042$.

Our hypothesis states that the colexification ratio for any concept pair, for any of our databases, is always less than 1, given that it is defined. That is, there is always some language that lexifies both concepts, but does not colexify them.

Resource	Languages	Concepts	Lexifications	Colexifications	Exceptions	Support
CLICS	3050	2919	1,377,282	75,089	64	99.9%
BN	284	5000	1,441,990	88,907	3	99.9%
OMWN	34	5000	267,503	54,615	4	99.9%

Table 3.1: The statistics on the lexical resources, and the empirical validation of our hypothesis.

3.4 Results

In this section, we describe the empirical validation of our hypothesis on the colexification data from CLICS, BN, and OMWN. Our results are summarized in Table 3.1, which shows that all three resources provide very strong evidence for our hypothesis. Namely, 99.9% of all colexified concept pairs have a colexification ratio less than 1 in all three resources. We find only 71 apparent exceptions in the individual resources.

The three most frequently colexified concept pairs in each resource are shown in Table 3.2. For example, the concepts LEG and FOOT are both lexified in 1038 languages (i.e. CLICS contains words for them in those languages) but only 336 languages colexify both concepts (i.e. have a single word that can express both of them). So, the colexification of LEG and FOOT is far from universal. In fact, approximately 76% of the 75,089 colexified concept pairs in CLICS are colexified in only a single language.

3.4.1 Analysis

The 71 apparent exceptions to our hypothesis must be qualified by the fact that none of the three resources makes any claim of completeness. For each seemingly universal colexification, it may be the case that there exists a language that lexifies both concepts, and does not colexify them, but this fact is not recorded in the corresponding resource. In this section, we perform a cross-database analysis, to investigate how many, if any, of these apparent exceptions are actual counterexamples to our hypothesis, and how many are simply the result of resource incompleteness.

For example, there are only six languages[1] which lexify both of the con-

[1]Indonesian, Klon, Lavukaleve, Mbaniata, Mbilua, Savosavo

Resource	Colexified Concept Pair	COL	LEX	Ratio
CLICS	LEG - FOOT	336	1038	0.324
	WOOD - TREE	335	1036	0.323
	MOON - MONTH	313	538	0.582
BN	$town^1_n$ - $city^1_n$	100	121	0.826
	$painting^1_n$ - $image^1_n$	89	93	0.957
	$house^1_n$ - $dwelling^1_n$	88	117	0.752
OMWN	$book^2_n$ (work) - $book^1_n$ (object)	23	25	0.920
	$wing^2_n$ (airplane) - $wing^1_n$ (animal)	22	22	1.000
	$shout^2_n$ (cry) - $shout^1_n$ (with loud voice)	22	24	0.917

Table 3.2: The concept pairs colexified by the most languages in each of the three databases.

cepts DULL and BLUNT in CLICS. This is surprising, as English words lexifying these concepts are, in fact, used to name them. However, the concept DULL does not have the English word *dull* listed in CLICS. All six of the languages which do lexify both of these concepts have a single word which lexifies both; based on our criteria, this would represent a universal colexification, if CLICS was fully complete and correct. However, by cross-checking this example against the information in the other two resources, we find several languages that do not colexify the two concepts (Table 3.3).

The 64 apparent exceptions in CLICS involve 113 distinct concepts. Unfortunately, in all 64 cases, at least one of the concepts is not mapped any of the WordNet core synsets. To remedy this, we manually map a subset of the 64 exceptions to OMWN and BabelNet. We choose all four instances that are colexified in more than two languages, plus ten more instances that are selected at random. We find that none of these 14 pairs are exceptions in OMWN or BN (Table 3.3). In other words, there is at least one language in each of OMWN and BN that lexifies the pairs but does not colexify them. Based on this analysis, we conclude that the 14 exceptions are caused by data sparsity.

In BabelNet, there are only three apparent exceptions to our hypothesis (Table 3.3). Considering BabelNet alone, they appear to be counterexamples to our hypothesis. Unfortunately, the corresponding WordNet concepts are not mapped to CLICS concepts. However, we find that none of these three

Colexified Concept Pair	CLICS Ratio	BN Ratio	OMWN Ratio
RUN_AWAY - FLEE	**10/10**	24/36	13/17
DULL - BLUNT	**6/6**	34/37	6/8
RIVER - FLOWING_BODY_OF_WATER	**4/4**	2/69	0/20
FISHING - CASSOWARY	**3/3**	0/45	0/12
SKIN (human) - SKIN (animal)	**3/3**	10/13	7/10
SAME_SEX_OLDER_SIBLING -BROTHER	**2/2**	44/96	9/16
PIMPLE - BOIL (of skin)	**2/2**	37/63	5/15
MALE - BRASS_INSTRUMENT	**1/1**	0/41	0/13
GAZELLE - DEER	**1/1**	4/79	0/17
WRAPPER - DRESS	**1/1**	1/51	0/12
HYENA - CART	**1/1**	0/55	0/16
ECHIDNA - ANTEATER	**1/1**	6/58	4/10
STRIKE - CAST	**1/1**	0/20	0/14
WRAPPER - CLOTH	**1/1**	0/53	0/12
$intention_n^3$ - $purpose_n^1$	n/a	**19/19**	14/15
$reserve_v^3$-$reserve_v^4(book)$	n/a	**20/20**	14/16
$increase_n^4$ - $increase_n^3(increment)$	n/a	**26/26**	20/22
$wing_n^2(airplane)$ - $wing_n^1(animal)$	n/a	31/47	**22/22**
$short_a^1(time)$ - $short_a^2(length)$	n/a	36/37	**20/20**
$probability_n^1$ - $probability_n^2(event)$	n/a	32/33	**18/18**
$new_a^1(time)$ - $new_a^{11}(unfamiliar)$	n/a	18/19	**16/16**

Table 3.3: The concept pairs with the ratio of 1 represent possible exceptions to our hypothesis. The fact that the corresponding ratio is less than 1 in another resource provides evidence against the exception.

pairs are exceptions in OMWN; for all three, the OMWN colexification ratio is less than 1. For example, Chinese lexifies $reserve_v^3$ as *liu* and $reserve_v^4$ as *ding*. Based on this analysis, we conclude that the three apparent exceptions in BabelNet are artifacts of data sparsity.

The situation in OMWN is similar: we find only four apparent exceptions, and none of them are exceptions in BabelNet. For example, according to BabelNet, Icelandic lexifies $new_a^1(time)$ as *nýr*, and $new_a^{11}(unfamiliar)$ as *óþekktur*, but no Icelandic word lexified both concepts.

Chapter 4

Lexical Resource Mapping via Translations

A lexical resource links words in one or more languages with concepts which they can express. Each pair of a word and a concept that it lexicalizes corresponds to a unique sense of that word. For example, the word "plant" would have distinct senses corresponding to the industrial and vegetation concepts it can lexicalize. The task that we address is mapping (or aligning) concepts or senses across lexical resources. Given a concept in one of the resources, such a mapping allows us to identify an equivalent concept in the other resource.

Aligning concepts between lexical resources facilitates several tasks. First, lexical resources provide complementary knowledge but lack links between them. Aligning concepts between them enables lexical resources connected to each other. Second, combining information from multiple resources increases coverage of words and languages. For example, Open Multilingual WordNet (Bond and Foster, 2013) contains the translations of a large number of senses in all the languages covered by the aligned multilingual resources from which it was constructed. Third, in addition to word senses, lexical resources may contain other types of information, such as relations between concepts, glosses, and usage examples. By mapping the senses in one resource to their equivalents in another, information about a given concept can be retrieved from both resources, increasing the total knowledge available about each concept. Finally, lexical resources are important for various tasks in natural language processing (NLP). Inter-resource concept mapping has been shown to yield performance

improvements compared to using resources in isolation (Ponzetto and Navigli, 2010).

In this chapter, we introduce two methods for estimating semantic similarity between concepts: WordVote considers the sets of shared multilingual lexicalizations, while LangVote is based on the sets of languages that colexify the two concepts. Our methods depend exclusively on lexicalization information, without relying on concept glosses, relations between concepts, or other structured information. We do not attempt to combine different types of information in order to assess how far translations can take us towards solving the alignment task.

We evaluate our approach on the alignment of WordNet with two other lexical resources: CLICS and OmegaWiki. In both cases, our methods outperform three previous gloss-based methods. For the WordNet-OmegaWiki mapping, our best method matches the performance of a strong graph-based method. We release the WordNet-CLICS alignment that we produce in order to facilitate further work on this important task. Our work is the first to produce such an alignment with this level of coverage. Therefore, the release of this data constitutes a truly novel resource, allowing information from two distinct knowledge sources to be combined, providing potential benefit to downstream applications.

The chapter is structured as follows: Section 4.1 provides an overview of related work. Section 4.2 outlines our methods, and Section 4.3 describes our experiments and results.

4.1 Related Work

In this section we review prior work on cross-resource concept mapping. These papers vary in the source and target resources. In some cases, these differences preclude comparison to our own work.

A major line of prior work is concerned with linking WordNet synsets to Wikipedia articles. There are various motivations for doing so: improving word sense disambiguation, obtaining multi-lingual lexicalizations, and evalu-

ating mapping algorithms on a pair of highly dissimilar resources. Ponzetto and Navigli (2010) calculate the word similarity between "disambiguation contexts" constructed from the two resources. They compute English lexicalization overlap, but unlike ours, their approach is exclusively monolingual. Navigli and Ponzetto (2012) extend this approach by leveraging the graph of WordNet semantic relations to calculate the similarity between concepts. Finally, Pilehvar and Navigli (2014) propose a method for constructing graphs representing different lexical resources. The PageRank algorithm is then applied to these graphs to compute a similarity measure between pairs of concepts. They report accuracy values of 0.960, 0.930, and 0.893, on the mapping WordNet to Wikipedia, Wiktionary, and OmegaWiki, respectively. However, each of the papers considers a different subset of WordNet concepts, which complicates comparison between the methods.

Another sequence of papers focuses on aligning senses in various resources, including WordNet, GermaNet, Wiktionary, and OmegaWiki. The method of Meyer and Gurevych (2011) is based on the similarity between sense definitions. Gurevych et al. (2012) extend this approach to align WordNet and German OmegaWiki. In particular, they use machine translation to translate the lemmas and glosses of one resource into the language of the other resource, and then compute the similarity between sense definitions. Matuschek and Gurevych (2013) propose a graph-based method which, different from Navigli and Ponzetto (2012), considers relations between all senses. A similar approach is applied by Matuschek et al. (2018) to align Wiktionary and OmegaWiki. Their method is again based on the similarity between sense definitions, and the application of the personalized PageRank algorithm.

Resource alignment continues to attract attention of researchers. Annika Tjuka (2021) propose a frequency-based method of mapping words from psychology and linguistics in 40 languages to concepts in the Concepticon dataset. For a given word, the algorithm first finds the concepts containing that word in their names, and returns the most frequent such concept. McCrae and Cillessen (2021) apply several similarity techniques to map WordNet synsets to entities in Wikidata. Yao et al. (2021) frame the task of map-

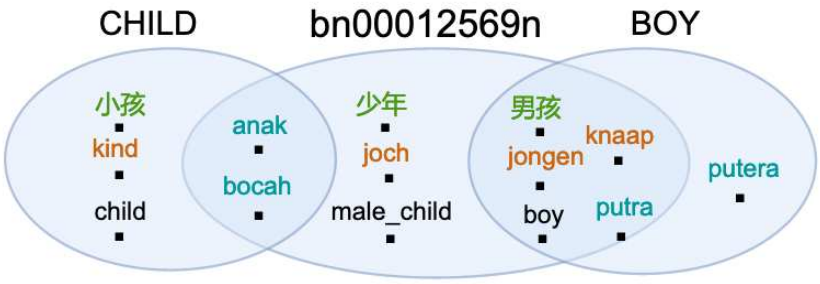

Figure 4.1: An example of concept lexicalization overlaps, with words from Chinese, Indonesian, Dutch, and English.

ping WordNet senses to dictionaries as the maximum-weight bipartite graph matching problem. They use a gloss similarity measure to weight the edges of a bipartite graph, and then find an exact solution to the maximum bipartite matching problem. We compare this method to our own in our experiments.

4.2 Methods

In this section, we introduce two translation-based methods for mapping concepts across lexical resources. Different from prior work, our methods depend exclusively on measures of concept similarity based on multilingual lexicalizations, that is, various translations of the concepts to be mapped. Both methods attempt to map each source concept in one resource to a target concept in another resource by maximizing a similarity measure which is based on the shared concept lexicalizations.

We hypothesize that the number of shared lexicalizations is correlated with the semantic similarity between concepts. For example, if two concepts both have the English lexicalization *plant*, that is a relatively weak evidence for similarity than if the concepts also share the English lexicalizations *industrial plant* and *factory*, as well as the French lexicalization *site*. The main difference between the two methods is that WordVote considers the total number of shared lexicalizations, whereas LangVote considers only the number of languages that exhibit shared lexicalizations.

Our first method, WordVote, maps a given source concept to the target concept by maximizing a similarity measure based on the size of the lexical

intersection between the concepts. Formally, for this method, we define the similarity measure between two concepts as:

$$s_W(c_s, c_t) = |\{lex(c_s, \mathcal{L}) \cap lex(c_t, \mathcal{L})\}|$$

where $lex(c, \mathcal{L})$ is a function that returns the set of lexicalizations of the concept c in a set of languages $\mathcal{L}$.

Figure 4.1 shows an example: the WordNet synset $male\_child_n^1$ shares all five of its lexicalizations (in $\mathcal{L}$ = {English, Chinese, Indonesian, Dutch}) with the CLICS concept BOY, but only two of its five lexicalizations with CHILD. So, the WordVote method maps bn00012569n to BOY.

Our second method, LangVote, maps a given source concept to the target concept by maximizing the number of languages in which the two concepts share at least one lexicalization. This can be viewed as a variant of the WordVote method, in which at most one lexicalization from each language may be included in the intersection. Formally, we define this similarity measure between two concepts as:

$$s_L(c_s, c_t) = |\{L \in \mathcal{L} : lex(c_s, L) \cap lex(c_t, L) \neq \emptyset\}|$$

Returning to the example in Figure 4.1, $male\_child_n^1$ and BOY share at least one lexicalization in all four of the languages under consideration. However, this WordNet synset and CHILD have a common lexicalization only in one language, Indonesian (the fact that there are two common Indonesian lexicalizations is not relevant for this method). So the LangVote method maps the synset bn00012569n to the CLICS concept BOY.

In both the WordVote and LangVote methods, a given source concept is mapped to the target concept which maximizes a similarity measure, based on their lexicalizations in a set of languages. These similarity measures both return an integer, which can result in ties between candidate concepts. To break ties between multiple target concepts, we select the concept which has the highest number of lexicalizations in set of languages. The intuition is that this strategy should favor more frequently used concepts, which in turn should have more reliable information. In our development experiments, we found

that this tie-breaking strategy works better than normalizing the similarity value by the number of lexicalizations, or breaking ties randomly.

4.3 Experiments

In this section, we describe a series of concept mapping experiments. We describe the systems we compare to and our evaluation metrics in Sections 4.3.1 and 4.3.2, respectively. In Section 4.3.3, we describe how we chose a set of languages $\mathcal{L}$. In Section 4.3.4 and 4.3.5, we present experiments mapping WordNet and CLICS, and WordNet and OmegaWiki, respectively.

In order to ensure that the mapping we produce is one-to-one, we align concepts starting with the most similar pairs. Once a pair of concepts have been aligned, those concepts are removed from consideration.

4.3.1 Comparison Methods

In this section, we describe three gloss-based methods and a graph-based method from prior work that we compare with our translation-based methods.

The method of Meyer and Gurevych (2011), which we refer to as MG11, creates a sparse, interpretable embedding for each concept, based on its gloss. Each dimension in these embeddings corresponds to a word, and the value of each dimension is the frequency of that word in the gloss. A similarity function on gloss embeddings is then used to map each source concept to its most suitable target concept. We use our own re-implementation of the method, and the method produces one-to-one mapping.

Reimers and Gurevych (2019) introduce SBERT, a method for generating dense sentence embeddings, which we apply to the concept mapping task. The intuition is that semantically similar sentences should have similar embeddings. We first generate an embedding for each concept based on its gloss using the provided code.[1] We use the cosine similarity between the embeddings of two glosses as a concept similarity measure. Using this similarity, we then perform

[1] `https://huggingface.co/sentence-transformers/stsb-mpnet-base-v2`

concept alignment between resources as with our own methods. This method produces one-to-one mapping.

For both SBERT and MG11, we expand the glosses with additional information from the resources. For WordNet, we follow Meyer and Gurevych (2011) in combining the gloss of the synset with its synonyms and hypernyms. For CLICS, we expand the gloss with its concept name.

Yao et al. (2021) introduce SemEq, a bipartite graph mapping method. The algorithm first builds a bipartite graph using glosses from two resources, where nodes represent concepts and edges indicate whether two nodes are candidates to each other. Then, the algorithm weights the edges with a gloss cosine similarity measure. Finally, the method finds the correct alignment between senses by maximizing the sum of the edge weights in the alignment. We use the implementation of this method made available by the authors[2] , and the method produces one-to-one mapping.

Pilehvar and Navigli (2014) introduce SemAlign, a state-of-the-art mapping algorithm. The method first constructs graphs using glosses and structural information from two resources, respectively. Then, the PageRank algorithm is applied to these graphs to compute a similarity measure between pairs of concepts. The method was evaluated on mapping different resource pairs. We only use the result of mapping WordNet and OmegaWiki reported by the authors.

4.3.2 Evaluation Measures

As measures of the quality of concept mapping approaches, we report accuracy and mean reciprocal rank (MRR). Accuracy is the proportion of source concepts that are mapped to the correct target concept. MRR is calculated as follows: One resource is designated the "source", while the other is designated the "target"; in our experiments, these designations correspond to the direction of the gold standard mapping which we use. For each source concept, the target concepts are ranked in order of their similarity values. The rank of the correct target concept in this ordering is identified, and its reciprocal is

[2] `https://github.com/tencent-ailab/EMNLP21_SemEq`

computed. The maximum reciprocal rank is therefore 1. If the correct target concept is not in the ranking, due to not being among the candidates for that source concept (because, due to our method design, only a subset of concepts constitutes the candidates, and that most concepts are not among the candidates), the reciprocal rank is zero. The average of these reciprocal ranks over all source concepts gives the MRR. There is no gold standard that ranks the target concepts, but it specifies a single correct target concept; we use MRR as an alternative to measure the quality of our mapping methods.

4.3.3 Language Selection

The only tunable parameter in our methods is the set of lexicalization languages. As languages differ greatly in their BabelNet coverage, simply using all languages is suboptimal in terms of both running time and mapping accuracy. We establish the set of languages on our WordNet-CLICS development set.

Our language selection procedure is as follows: In addition to English (EN), we considered the languages with the highest lexicalization overlap between CLICS and BabelNet: Dutch (NL), Romanian (RO), Spanish (ES), Portuguese (PT), Italian (IT), Indonesian(ID), Irish (GA), French (FR), and German (DE). We also included Chinese (ZH) and Russian (RU), as are typologically and orthographically different from the above languages.

In the first step, we aimed to establish the ranking of languages within these 12 languages. We performed experiments on the development set with each individual language coupled with English. The ranking is shown in Table 4.1, with the languages ordered by the accuracy of our LangVote method, with ties broken randomly. In the second step, we constructed the final set of languages by adding languages one by one, following the ranking in Table 4.1. We continued to add languages from the list until a decrease in accuracy was observed. In short, we applied a greedy strategy of adding languages according to the accuracy they produced on our development set, with English always being included by default. This process yielded the set of seven languages (EN, ID, NL, DE, RO, IT, and GA) which we use in all experiments that follow.

Lang.	Overlap	ACC
EN	1881	0.799
EN & ID	3562	0.892
EN & NL	4069	0.869
EN & DE	3466	0.869
EN & RO	3899	0.860
EN & IT	3692	0.854
EN & GA	3546	0.851
EN & ES	3782	0.848
EN & PT	3693	0.843
EN & ZH	2705	0.843
EN & RU	2628	0.840
EN & FR	3538	0.837

Table 4.1: Results of the LangVote method on the development set with different language pairs, including the size of the word overlap between CLICS and BabelNet; the first row represents the lexical overlap for English, and the accuracy of mapping the two resources using English only.

4.3.4 Aligning WordNet and CLICS

Our principal concept-mapping dataset comes from Concepticon (List et al., 2016), which includes hand-crafted mapping between a subset of CLICS concepts and WordNet. We extracted the dataset by following the procedure described in List (2018). The mapping contains 1372 one-to-one pairings of CLICS concepts and WordNet synsets. As our development set, we used 343 concept pairs that include usage examples. The remaining 1029 concept pairs constitute our test set.

Both of our methods, WordVote and LangVote, require translation information. As described in Chapter 2, this information is readily available for CLICS, while the source of translation information for WordNet is BabelNet.

As mentioned in Chapter 2, CLICS concepts are associated with categories, whereas WordNet concepts are marked with a part of speech. Since our methods use part-of-speech information to map concepts, based on our analysis of the development data, we mapped CLICS categories to parts of speech as follows: Action/Process: Verb; Number or Person/Thing: Noun; Property: Adjective or Adverb; Other: Adjective, Noun, or Adverb.

While our evaluation is limited to the 1029 WordNet synsets and the cor-

Method	ACC	MRR
MG11	0.517	0.654
SBERT	0.591	0.713
SemEq	0.667	0.657
LangVote	0.706	0.818
WordVote	0.711	0.823

Table 4.2: Accuracy (ACC) and MRR for the alignment of WordNet and CLICS, on the test set.

responding 1029 CLICS concepts which comprise our test set, the experiment involves *all* the synsets and concepts in these resources, that is, we map all synsets/concepts between WordNet and CLICS. Since CLICS has fewer concepts than WordNet has synsets, this means that each method that we apply attempts to align each CLICS concept with a single WordNet synset. However, in some cases, no alignment is found, due either to the lack of any overlap in the set of languages we consider, or due to the one-to-one constraint removing all viable alignments; 317 CLICS concepts are not mapped by our WordVote method because of these issues. Our work is the first to construct CLICS-WordNet alignment with such extensive coverage; we will release the CLICS-WordNet alignment produced by our methods to facilitate further work with these resources.

The results of our experiment on the test set are presented in Table 4.2. Our two translation-based methods perform well above the three gloss-based comparison methods. WordVote achieves slightly better accuracy than LangVote, which suggests that the total number of shared lexicalizations provides useful information to the number of shared languages.

Our error analysis revealed two main sources of error. First, some CLICS concepts are duplicate and/or combine multiple concepts, which complicates the identification of a correct one-to-one mapping. For example, CLICS contains separate concepts named "STONE OR ROCK" and "STONE". Second, many translations are missing from the resources. For example, the CLICS concept "TO DRIP" has no lexicalizations in Indonesian, Dutch, German, or Romanian, while the BabelNet synset $drop_n^1$ contains no Dutch words. Our

analysis is based on four random-selected instances, and all the errors are due to the resources. So, we conclude that most of the apparent errors are due to issues with the resources rather than flaws in our methods. More principled methods of defining concepts, and improvements in the multilingual coverage of lexical resources, would likely improve the resource alignment results, in addition to yielding other benefits.

4.3.5 Aligning WordNet and OmegaWiki

To validate the generality of our translation-based approach, we carry out an experiment on WordNet and OmegaWiki. The gold data for this experiment was originally developed on the German part of OmegaWiki, which consists of German lexicalizations and concept glosses (Gurevych et al., 2012). Building upon this, Matuschek and Gurevych (2013) evaluate their mapping algorithm on this dataset directly, as each German OmegaWiki concept has at least one English lexicalization associated with it. For our evaluation, we use the version of the dataset provided by Pilehvar and Navigli (2014), who added additional English OmegaWiki candidates.

The data contains 315 WordNet synsets, but only 215 of them are aligned with OmegaWiki concepts. Since the dataset contains no lexicalizations of OmegaWiki concepts, we must refer to OmegaWiki itself to apply our methods. Unfortunately, the version of OmegaWiki that served as the basis for this dataset is no longer available; we therefore use a more recent version (from 16 September 2021). Because of the dynamic nature of OmegaWiki, some concepts in the gold data are missing from the current version. We therefore restrict our evaluation to those OmegaWiki concepts that still have identical glosses as the current version. This yields a test set consisting of 276 WordNet synsets, of which 148 are aligned to OmegaWiki. No part of this data was used in development.

We made no changes to our methods to adapt them to this dataset, and attempted to keep our experimental setup as close as possible to the experiment described in Section 4.3.4. The gold data is not one-to-one mapping, and contains positive examples (concepts should be mapped to each other) and

Method	ACC
MG11	0.840
SBERT	0.854
SemEq	0.853
SemAlign	0.893
LangVote	0.879
WordVote	0.894

Table 4.3: Results for aligning WordNet and OmegaWiki, evaluated on the test set in terms of accuracy.

negative examples (concepts should not be mapped to each other). In order to compare with the results of the SemAlign system as reported by Pilehvar and Navigli (2014), accuracy in this experiment is defined as follows: Accuracy = (TP+TN) / (TP+TN+FP+FN), where true positives (TP) is the number of examples correctly detected as positive by our method; true negatives (TN) is the number of examples correctly detected as negative by our method; false positives (TP) is the number of examples which are aligned by our method but should not be; false negatives (FN) is the number of examples which are not aligned by our method but should be.

Table 4.3 shows the mapping accuracy of several methods. Following Pilehvar and Navigli (2014) we do not report MRR, as the gold mapping is not one-to-one, and so MRR is not well defined, as there need not be a single correct mapping. The three comparison systems, MG11, SBERT, and SemEq, all achieve similar results to one another, below our methods or SemAlign. Our WordVote method performs comparably to SemAlign. This is remarkable considering that our approach is based exclusively on translation information, whereas SemAlign depends on glosses as well on semantic relations between concepts. We conclude that this result provides evidence for the generality of our approach.

In our error analysis, we found two principal causes of errors. First, while our approach is designed to produce a one-to-one alignment, the data contains both one-to-many alignments and unaligned concepts. Second, due to the volatile nature of OmegaWiki, some concepts in the gold data are not in the current version. For example, our approach has no chance to find the correct

mapping for the WordNet synset $terminology_n^1$ ("A system of words used to name things in a particular discipline.") because three out of six candidates, including the correct one, are no longer in the current OmegaWiki. Such errors are thus not due to a flaw in our method; rather, it highlights the risk in using volatile online resources as a source of gold-standard data.

Chapter 5

Determining Sense Synonymy via Translations

This chapter describes our systems for SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (Martelli et al., 2021). We focus on the monolingual (English) variant of the task, which is the same as the original Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2018). WiC dataset is proposed for the evaluation of contextualized word embeddings, because very few benchmarks exist for evaluating context-sensitive embeddings that are used to capture the "polysemous nature of words" (Martelli et al., 2021). The dataset includes training, development, and testing splits. An instance of the WiC task consists of two sentences that share a *focus word* in common; the word may be inflected differently in each sentence (e.g. "they had searched his flat a few days before" and "the production of lithium from salt flats") but will share the same lemma and part of speech. A WiC task system must decide, given such a pair of sentences, whether the *focus word* has the same meaning in both sentences. Systems are compared in terms of their accuracy (Martelli et al., 2021), the percentage of test instances correctly identified as TRUE (same meaning) or FALSE (different meaning).

The top three systems (Yuan and Strohmaier, 2021; Zhestiankin and Ponomareva, 2021; Gupta et al., 2021) of this task mentioned in Martelli et al. (2021) achieve an accuracy of 0.933, 0.927, and 0.925, respectively. All the three systems are based on supervised methods. They first augment the training dataset using external resources, such as examples from the Cambridge

Advanced Learner's Dictionary. Then, they use pre-trained models to obtain contextual representations of the target words. Finally, the systems are fine-tuned on augmented training data. Compared with these three systems, our results are much lower because our methods are unsupervised, we do not use the training data.

In this task, we investigate whether translation can be used to detect semantic equivalence in context. The intuition underlying our work is that distinctions in meaning tend to be reflected in distinctions in translation. Our focus is on developing principled theoretical approaches which are grounded in linguistic phenomena, leading to more explainable models.

Our methods depend upon a mapping between word senses and translations, as different senses of a word often translate differently. We obtain such a mapping from BabelNet (Navigli and Ponzetto, 2012), and treat BabelNet as an imperfect implementation of a universal multi-wordnet with the theoretical properties described by Hauer and Kondrak (2020b).

Our results can be interpreted as a proof-of-concept for the use of contextual translations as indicators of semantic similarity. We show that the methods that we develop for the WiC task can leverage translations to improve over baselines, especially when multiple target languages are considered.

This chapter is structured as follows: Section 5.1 provides an overview of relevant prior literature. Section 5.2 discusses the theoretical model underlying our work. Section 5.3 outlines our methods. Section 5.4 describes our experiments and results.

5.1 Related Work

Methods for WiC task can be roughly divided into two paradigms: embedding-based systems and word sense disambiguation-based systems. Pilehvar and Camacho-Collados (2018) introduce the word-in-context dataset as a benchmark for evaluating context sensitive word representations. Soler et al. (2019) achieve improvements by combining similarity scores from different types of contextual word and sentence embeddings. Liu et al. (2020) propose a method

to enhance contextual representations by leveraging other pre-trained contextual or static embeddings. Some other embedding-based systems fine-tune pre-trained language models on augmented training data, then use logistic regression to perform a binary classification (Yuan and Strohmaier, 2021; Zhestiankin and Ponomareva, 2021; Gupta et al., 2021).

Another approach to WiC task is to employ a word sense disambiguation (WSD) system to tag the target words with senses from a pre-defined sense inventory and subsequently make a decision based on the predicted synsets of the target words. Loureiro and Jorge (2019b) use the LMMS sense embeddings (Loureiro and Jorge, 2019a) to disambiguate the target words. A simple approach of checking if the disambiguated senses are equal lead to competitive performance in this task (Espinosa-Anke et al., 2019). SENSEMBERT (Scarlini et al., 2020a) and ARES (Scarlini et al., 2020b) embeddings, when used as features in a BERT-based model, also achieve competitive results on the WiC task.

Our methods combine elements of the two paradigms. We employ contextual embeddings in our proposed translation-based methods. However, we take the embeddings of the translations of the target words instead of the target words themselves. Similarly to WSD based approaches, our methods also analyze the common synsets of the focus tokens and their translations, with the goal of identifying a probable shared synset. The most similar prior work to our approach is that of Pessutto et al. (2020) at the graded word similarity task (Armendariz et al., 2020), who propose a translation-based approach to evaluate the contextual similarity of a pair of words. They hypothesize that leveraging similarity information from more languages would allow greater accuracy. We follow a similar intuition in our work.

5.2 Theoretical Solution

We first present a theoretical solution, which provides the foundation for the development of our actual methods described in Section 5.3. We assume that the two source sentences S_1 and S_2 can be translated into any natural language

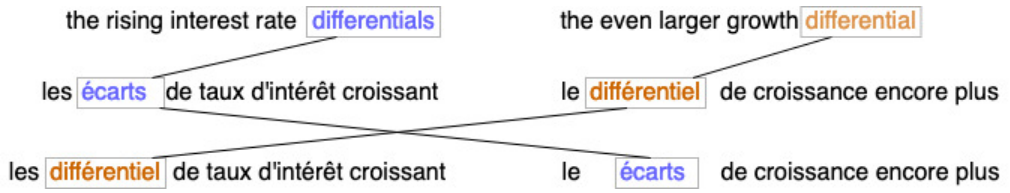

Figure 5.1: An example of the "translation criss-cross" described in Section 5.2.2.

as sentences T_1 and T_2. Furthermore, we assume that the literal lexical translations t_1 and t_2 of the focus word s can be identified in T_1 and T_2, respectively. For example, in Figure 5.1, the focus word s in the English sentences S_1 and S_2 is the noun *differential*, and word alignment identifies *écart* and *différentiel* as t_1 and t_2. Note that the two translations may have the same POS and lemma, a scenario we denote as $t_1 = t_2$.

5.2.1 Substitution Test

Our theoretical solution is based on the notion of the linguistic *substitution test* for verifying the synonymy of senses (Hauer and Kondrak, 2020b), which takes as input two sentences *which differ only in a single word*, and returns TRUE if and only if the two sentences have the same meaning. In other words, it decides whether the substitution of one word with another changes the meaning of the sentence. Note that this substitution test is not sufficient to decide the sense synonym in this task, as the input sentences for this task *share* a single word, rather than *differ* in a single word. The substitution test can be implemented by consulting a native speaker, or approximated by a computer program. In Section 5.3, we discuss an implementation based on contextual embeddings.

An example of a valid input to the substitution test would be the sentences *I work at the plant* and *I work at the factory*. For this input, the substitution test would return TRUE, since the word substitution does not change the meaning of the sentence. The sentences *I work at the plant* and *I work at the flower* would likewise constitute a valid input; however, given these sentences, the substitution test would return FALSE, since the sentences differ semantically.

5.2.2 Translation Criss-Cross

In order to apply the substitution test to an instance of the WiC task, we first translate the two source input sentences S_1 and S_2 into a target language, producing two target sentences T_1 and T_2. We identify the two lexical translations t_1 and t_2 of the focus word s in T_1 and T_2. Assuming that the translations are correct and literal, the senses of s in S_1 and t_1 in T_1 will be synonymous, as well as the senses of s in S_2 and t_2 in T_2. If t_1 and t_2 have the same POS but different lemmas, we can replace t_1 with t_2 in T_1 to produce a sentence T_1' which differs from T_1 in a single word. The application of the substitution test to (T_1, T_1') returns TRUE if and only if the sense of t_2 in T_1' is synonymous with the sense of s in S_1, which implies that, in addition to s and t_1, the multi-synset containing the sense of s in S_1 must also include t_2.

Using our running example in Figure 5.1, T_1' would be created by replacing *écarts* with *différentiel* in T_1. This produces *les différentiel de taux d'intérêt croissant*, which, while not necessarily grammatical, can still be evaluated by the substitution test to decide whether the substitution alters the semantic content of the sentence. (Or, equivalently, whether *écart* and *différentiel* are synonymous in this particular context.)

We repeat the process with the roles of T_1 and T_2 reversed. That is, we construct T_2' by replacing t_2 with t_1 in T_2 in order to verify whether the sense of t_1 in T_2' is synonymous with the sense of s in S_2. If the substitution test returns FALSE for either of the two target sentence pairs, we can conclude that the two multi-synsets that correspond to the senses of s in S_1 and S_2 must be different. Therefore, this instance of the WiC task is resolved as FALSE. However, if the substitution test returns TRUE for both pairs of sentences, we cannot immediately resolve the instance of the WiC task, because there could exist two (or more) multi-synsets that all contain s, t_1, and t_2.

A complete theoretical solution can be obtained by considering translations in multiple languages. If the focus word s is not used in the same sense in S_1 and S_2, we would expect that in *some* language, the translations t_1 and t_2 will be different *and* not mutually replaceable in both sentences. This expectation

is consistent with the speculation of Palmer et al. (2007) that translation into a sufficiently large set of language will eventually lexicalize every sense distinction. It is also supported by the findings of Bao et al. (2021) who found no evidence for the existence of universal colexifications, that is, pairs of concepts that are expressed by the same word in every natural language.

5.2.3 Multi-Synset Intersection

For each language F_i in the set of all natural languages $\mathcal{L}$, let t_1^i and t_2^i be the lexical translations of the focus word s in the first and second input sentences, respectively. Let T be the set consisting of the focus word, and all its lexical translations; that is $T = \{s\} \cup_{F_i} \{t_1^i, t_2^i\}$. Assuming access to a perfect universal multi-wordnet, we define the set C to be the set of multi-synsets that contain all words in T.

The size of C provides clues to the resolution of the WiC task. We need to consider three cases: $|C| = 0$, $|C| = 1$, and $|C| \geq 2$. With some caveats, these three cases roughly imply the following answers to the WiC task: False, True, and Unknown, respectively. We discuss these three cases in turn.

If $|C| = 0$, then no single concept can be expressed by s and all its translations in T, according to the multi-wordnet. That is, there exist two translations of the focus word which cannot express the same concept, assuming the completeness of the multi-wordnet. Therefore, the two focus tokens must correspond to distinct multi-synsets, implying False.

If $|C| = 1$, there exists exactly one multi-synset that contains the focus word and all its translations. Therefore, it is possible, albeit not guaranteed, that the focus word in both source sentences is used in the sense that corresponds to that unique multi-synset. In order to be sure, we could apply the criss-cross method described in Section 5.2.2.

$|C| \geq 2$ would imply that there exist two concepts which are colexified (expressed by a single word) in all languages. Following Bao et al. (2021), we assume that universal colexifications are at best extremely rare. Even if they exist at all, we could still apply the solution described in Section 5.2.2 to decide the WiC task. Of course, if we are considering translations into only a small

number of languages, the possibility of $|C| \geq 2$ is much more likely. In fact, we observe $|C| = 3$ in our running example, because three different BabelNet multi-synsets contain the English focus word and its two French translations.

5.3 Methods

In this section we describe four methods based on the theoretical ideas in Section 5.2. All four methods rely on identifying lexical translations of the focus word in both source sentences. If the lexical translations cannot be recovered from the translated sentences for any of the target languages, all methods use the same backoff approach, which is to return FALSE for that test instance.

5.3.1 Ident and CVal

Our two simplest methods are IDENT and CVAL. IDENT is a baseline method which returns TRUE **iff** the lexical translations t_1 and t_2 have the same lemma and POS in all applicable target languages in which we can identify the lexical translations.

CVAL is a method directly based on the cardinality of the set C as defined in Section 5.2.3. CVAL returns TRUE **iff** the translations of the focus word are identical in each language **and** $|C| > 0$.

5.3.2 Synonymy Check

We implement the substitution test as a heuristic *synonymy check* using dense contextualized embeddings. Such embeddings allow us to construct, for any word token in a given sentence, a vector in a continuous semantic space. The objective in designing such embeddings is that semantically similar tokens should have similar vectors, commonly measured by cosine similarity. Additional technical details of the embeddings are provided in Section 5.4.

Given a pair of sentences which differ only in the substitution of single word, we obtain dense contextualized embeddings of the distinguishing word in each sentence. We then calculate the cosine similarity between the two em-

beddings. If the similarity is greater than a threshold tuned on a development set, this is taken as an indication that replacing one of the distinguishing words with the other does not alter the meaning of the sentence, as the replacement word has the same meaning as the original word. This implementation of the substitution test is used as a subroutine by our remaining two methods.

5.3.3 Sub and CSub

The SUB method attempts to apply the synonymy check to each pair of translated sentences T_1 and T_2 in each target language, without referring to the $|C|$ value. If the translations of the focus word in T_1 and T_2 differ, we create the sentences T_1' and T_2', as described in Section 5.2.2, and apply the synonymy check to (T_1, T_1') and (T_2, T_2'). SUB returns TRUE if the synonymy check succeeds for all target languages for which the translations t_1 and t_2 can be identified. The synonymy check trivially succeeds if t_1 and t_2 have the same POS and lemma; intuitively, tokens which translate the same way are likely to have similar meanings. If either application of the synonymy check fails, SUB returns FALSE. In summary, this method is similar to the IDENT method, except that the synonymy check is applied if the translations differ.

CSUB combines CVAL with SUB. The only difference with the SUB method is that the synonymy check is not applied when $|C| = 0$. This is because the lack of any common multi-synset in a complete perfect multi-wordnet is theoretically sufficient to exclude the possibility of the two source focus tokens having the same sense.

5.4 Experiments

In this section, we describe the application of our methods to the development and test sets. We begin by specifying various implementation details. Next, we describe our development experiments, including results and error analysis. Finally, we present our results on the test set.

5.4.1 Translation and Lemmatization

We use BabelNet (Navigli and Ponzetto, 2010, 2012) as our multi-wordnet; in particular, we make use of the BabelNet multi-synsets which are linked to Princeton Wordnet synsets. This allows us to exclude synsets that refer to named entities, rather than lexicalized concepts, to limit the impact of noise in BabelNet.

For translation, we use Google Translate, as it is fast and publicly available. In our analysis, we found the lexical translations obtained using Google Translate to be of generally high quality, which is important given our method's dependence on machine translation. We use French, Italian, and Russian as our languages of translation. The choice of the translation languages is based on the languages selected for the shared task, and also on the BabelNet coverage. French and Russian are two of the languages covered by the shared task. On the other hand, Italian seems to have the best BabelNet coverage among the non-English languages.

For lemmatization, we use TreeTagger (Schmid, 1999, 2013), with pretrained lemmatization models for the source and all target languages. We lemmatize the bitexts to improve the quality of the word alignment.

5.4.2 Word Alignment

Following lemmatization, we first align each input sentence with its translation in each target language. Then, we apply an unsupervised knowledge-based alignment algorithm to identify the word or phrase in the translated sentence corresponding to the source focus word. Finally, we extract the lemmas aligned with each focus word token. These lemmas are the lexical translations of the focus word. To carry out the alignment, we use BabAlign (Luan et al., 2020), a state-of-the-art knowledge-based aligner. BabAlign leverages translation information from BabelNet to create synthetic training data and post-process the alignment produced using a base unsupervised alignment method.

Lang.	FR	IT	RU	ALL
IDENT	**59.6**	**58.1**	**57.1**	59.7
CVAL	58.9	57.6	54.3	55.5
SUB	59.3	58.0	55.6	**60.8**
CSUB	59.2	57.8	54.3	54.1

Table 5.1: Accuracy on the development set with different methods and languages of translation; the source language is English, and the columns are the target languages.

5.4.3 Contextual Embeddings

To obtain contextual representations for the purposes of deciding the substitution check, we use BERT (Devlin et al., 2019), a deep neural architecture trained with the masked language model. We chose BERT because it has been proven to capture the semantics of a word in context (Coenen et al., 2019). The context is the sentence containing the focus word. Specifically, we use cased multilingual BERT embeddings with 768 dimensions, 12 layers, 12 attention heads, and 179M parameters. To implement the substitution check, we generate contextualized embeddings of the translations of the focus tokens, and their substitutes, by summing the last four hidden layers of the BERT model. This choice was based on the results achieved by Devlin et al. (2019) in the named entity recognition task, and by Soler et al. (2019) in the SemDeep-5 WiC shared task.[1] Since BERT uses sub-tokens to generate embeddings, we analyzed the impact of two different sub-token selection techniques for predicting word similarity: using only the first sub-token, and using the mean over all the sub-tokens. In our development experiments, we found that the former yielded better results. Therefore, only the first sub-token is used to create contextualized embeddings for the substitution method.

5.4.4 Development Results

Table 5.1 shows the results of our development experiments. The baseline translation identity method IDENT does relatively well, outperforming both methods based on intersecting sets of multi-synsets, CVAL and CSUB. In-

[1]https://www.dfki.de/ declerck/semdeep-5/challenge.html

deed, these methods tend to suffer accuracy degradation as more languages of translation are added. We speculate that this is due to these methods being more vulnerable to noise (errors or omissions) in the multi-wordnet and in the extraction of lexical translations. However, the best performing method is SUB, which also shows improvement when combining all three languages of translation. Thus, it also shows the most promise for further improvement by adding additional languages.

Our error analysis suggests that there are three principal causes of errors. First, translation may be non-literal. For example, in one instance, the adverb "unevenly" is translated into French as the adjective "inégale" ("unequal"), leading to a false negative. Second, distinct but synonymous translations may lead to false positives. In one instance, the focus word "stain" is translated as "souillé" in one sentence and "tachée" in the other. The focus tokens have distinct meanings, reflected in their distinct translations, "stain on a reputation" versus "stain on a surface". However, the translations pass the BERT-based synonymy check, since they can be synonymous in some contexts. Finally, in some cases, distinct senses of a word may nevertheless translate the same way. For example, in one instance, the focus word "superior" was used in two distinct meanings. Both these meanings can be expressed by the French word "supérieur", and indeed, "superior" was translated as "supérieur" in both sentences, resulting in a false positive.

5.4.5 Test Results and Discussion

Table 5.2 shows our results on the test data. Consistent with our development experiments, the SUB method achieves the best performance with the combination of all three languages. The IDENT method once again performs relatively well despite its simplicity, outperforming the more complex CVAL and CSUB methods. Different from the development experiments, when only one language of translation is used, Russian yields substantially better performance compared to French or Italian across all four methods, and Italian likewise yields better performance than French.

Table 5.3 gives additional details for the results of the SUB method. For

Lang.	FR	IT	RU	ALL
IDENT	55.8	**58.9**	**61.0**	61.1
CVAL	54.8	55.6	56.0	55.2
SUB	**56.1**	57.6	60.6	**63.2**
CSUB	55.2	55.2	55.8	55.7

Table 5.2: Accuracy on the test set with different methods and languages of translation.

each of the three languages, and the combination of all three, we provide the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as well as the accuracy. We observe that using multiple languages of translation results in a substantial reduction in false positives, at the possible expense of an increase in false negatives, while maintaining an overall higher accuracy.

Lang.	TP	TN	FP	FN	Accuracy
FR	369	192	308	131	56.1
IT	376	200	300	124	57.6
RU	327	279	221	173	60.6
ALL	339	293	207	161	63.2

Table 5.3: Detailed breakdown of the results of our best performing method, SUB.

Chapter 6

Conclusion

In this thesis, we have explored the idea of leveraging translations in lexical semantics. We have proposed a novel hypothesis which states that *there are no universal colexifications*. We provided evidence that the few apparent exceptions to the hypothesis we found in three multilingual resources are attributed to omission errors in the resources. In the future, we plan to leverage our hypothesis to improve the accuracy of multi-lingual word sense disambiguation.

The validation of our hypothesis provides novel insights into several open issues in lexical semantics. It implies that every sense distinction in every language can be disambiguated by translation into some language. It also provides support for the informal conjecture of Palmer et al. (2007) that every possible sense distinction can be identified by translation into multiple languages. Finally, it furnishes evidence that the fine-granularity of wordnets and multi-wordnets is necessary for distinguishing between lexical translations of concepts.

Next, we present two novel methods of leveraging translations for aligning concepts across lexical resources. Our work is the first to explicitly use multilingual lexicalization information for this task; moreover, our method depends exclusively on such translation information, without any dependence on lexical relations, glosses, embeddings, or other sources of semantic or lexical knowledge. This demonstrates the utility of multilingual translation for resource mapping, while giving us a method which is highly explainable. We test our methods on two pairs of resources, and find that our methods match or exceed

the accuracy of the best comparable methods from prior work. We will make available the WordNet-CLICS concept mapping produced by our methods to facilitate comparison and further research.

Furthermore, based on translational equivalence, we propose four translation-based methods for determining sense synonymy. Our results provide a solid proof-of-concept for the utility of multilingual translation for determining sense synonymy. Our results also empirically verify the hypothesis that translations convey semantic information, and that this phenomenon has applications in lexical semantics.

References

Johann-Mattis List Annika Tjuka, Robert Forkel. 2021. Linking norms, ratings, and relations of words and concepts across multiple language varieties. In *Behavior research methods*.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference*, pages 1–7, University of South Africa (UNISA). Global Wordnet Association.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2022. Lexical resource mapping via translations. In *In Submission*.

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

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *arXiv preprint arXiv:1906.02715*.

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.

Luis Espinosa-Anke, Thierry Declerck, Dagmar Gromann, Jose Camacho-Collados, and Mohammad Taher Pilehvar, editors. 2019. *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*. Association for Computational Linguistics, Macau, China.

Christiane Fellbaum. 1998. WordNet: An on-line lexical database and some of its applications. *MIT Press*.

Alexandre François. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. *From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations*, 163-215.

A. Georgakopoulos, E. Grossman, D. Nikolaev, and S. Polis. 2020. Universal and macro-areal patterns in the lexicon. *Linguistic Typology*.

Rohan Gupta, Jay Mundra, Deepak Mahajan, and Ashutosh Modi. 2021. Mcl@ iitk at semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation using augmented data, signals, and transformers. *arXiv preprint arXiv:2104.01567*.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France. Association for Computational Linguistics.

Martin Haspelmath. 2000. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *The new psychology of language*.

Bradley Hauer, Hongchang Bao, Arnob Mallik, and Grzegorz Kondrak. 2021. UAlberta at SemEval-2021 task 2: Determining sense synonymy via translations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 763–770, Online. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2020a. One homonym per translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7895–7902.

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

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Johann-Mattis List. 2018. Cooking with CLICS. *Computer-assisted language comparison in practice*, 14-18.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association (ELRA).

Johann-Mattis List and Anselm Terhalle. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. *Proceedings of the 10th International Conference on Computational Semantics*.

Qianchu Liu, Diana McCarthy, and Anna Korhonen. 2020. Towards better context-aware lexical semantics: Adjusting contextualized representations through static anchors. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4066–4075.

Daniel Loureiro and Alipio Jorge. 2019a. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Daniel Loureiro and Alipio Jorge. 2019b. Liaad at semdeep-5 challenge: Word-in-Context (WiC). *arXiv preprint arXiv:1906.10002*.

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.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164.

Michael Matuschek, Christian M. Meyer, and Iryna Gurevych. 2018. Multilingual knowledge in aligned Wiktionary and Omegawiki for translation applications. In *Language technologies for a multilingual Europe*.

John P. McCrae and David Cillessen. 2021. Towards a linking between WordNet and Wikidata. In *Proceedings of the 11th Global Wordnet Conference*, pages 252–257, University of South Africa (UNISA). Global Wordnet Association.

Christian M. Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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

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

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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.

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.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Tommaso Pasini and Roberto Navigli. 2018. Two knowledge-based methods for high-performance sense distribution learning. In *Proc. of the 32th AAAI Conference on Artificial Intelligence*.

Vladimir Pericliev. 2015. On colexification among basic vocabulary. *Journal of Universal Language*, 63-93.

Lucas RC Pessutto, Tiago de Melo, Viviane P Moreira, and Altigran da Silva. 2020. Babelenconding at semeval-2020 task 3: Contextual similarity as a combination of multilingualism and language models. *arXiv preprint arXiv:2008.08439*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–478, Baltimore, Maryland. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts. *Artif. Intell.*, 228(C):95–128.

Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden. 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*. Association for Computational Linguistics.

Philip Resnik. 1997. Selectional preference and sense disambiguation. *Tagging Text with Lexical Semantics: Why, What, and How?*

Christoph Rzymski and Tiago et al. Tresoldi. 2019. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*.

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-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC challenge: Context representations for word usage similarity estimation. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 6–11.

Wenlin Yao, Xiaoman Pan, Lifeng Jin, Jianshu Chen, Dian Yu, and Dong Yu. 2021. Connect-the-Dots: Bridging semantics between words and definitions via aligning word sense inventories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7741–7751, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

Zheng Yuan and David Strohmaier. 2021. Cambridge at SemEval-2021 task 2: Neural WiC-model with data augmentation and exploration of representation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 730–737, Online. Association for Computational Linguistics.

Boris Zhestiankin and Maria Ponomareva. 2021. Zhestyatsky at semeval-2021 task 2: Relu over cosine similarity for BERT fine-tuning. *ArXiv*, abs/2104.06439.