Augmenting Context with Glosses for Lexical Semantics

by

Talgat Omarov

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 a subfield of natural language processing (NLP) that deals with the study of meaning in language at the level of individual words or phrases using computational models and algorithms. Despite the recent success of large language models and contextualized word embeddings in solving lexical semantic tasks, traditional lexical resources such as WordNet remain critical in providing comprehensive coverage of infrequent word meanings and providing additional information on word definitions, usage examples, and semantic relationships among senses. In this thesis, we explore the idea of leveraging information retrieved from lexical resources to solve lexical semantic tasks. In particular, we demonstrate that augmenting the input context with glosses retrieved from lexical resources improves the performance on two lexical semantic tasks: lexical substitution and idiomaticity detection. The results confirm the utility of additional lexical information and provide empirical evidence supporting our claims.

Preface

The work presented in Chapter 2 is published as B. Hauer, S. Jaura, T. Omarov, and G. Kondrak "UAlberta at SemEval 2022 Task 2: Leveraging Glosses and Translations for Multilingual Idiomaticity Detection" (Hauer et al., 2022). The author of this thesis has implemented the gloss-based method and conducted all experiments described in the chapter.

Chapter 3 is adapted from the research article T. Omarov and G. Kondrak "Grounding the Lexical Substitution Task in Entailment" (Omarov and Kondrak, 2023) in submission. The author of this thesis has implemented all methods and performed all experiments described in the chapter.

Acknowledgements

I would like to thank my supervisor, Professor Grzegorz Kondrak, for his support and guidance throughout the thesis. I also thank Bradley Hauer for his feedback and valuable advice. Finally, I would like to thank my parents and my partner for their constant support throughout my studies.

This thesis was completed with the funding from the Alberta Machine Intelligence Institute (Amii).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Computational lexical semantics is a subfield of natural language processing (NLP) that deals with the study of meaning in language at the level of individual words or phrases using computational models and algorithms. It focuses on the automated extraction of lexical information from context, including word meanings and relationships between them. Studying computational lexical semantics is important because accurately capturing meaning is very important in many downstream NLP applications such as machine translation, semantic role labeling, and question answering (Bevilacqua et al., 2021).

In recent years, there has been a growing trend in the field of NLP to utilize implicit latent representations that are learned from large text corpora to capture the meanings of words. This trend has been driven by significant advances in techniques such as word embeddings (Mikolov et al., 2013) and large language models (Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020). However, despite the success of these approaches, traditional lexical resources such as WordNet (Miller, 1995) remain critical in providing comprehensive coverage of infrequent word meanings and providing additional information on word definitions, usage examples, and semantic relationships with other senses (Navigli, 2018). Therefore, explicit representation of word meaning continues to be relevant to this day.

In this thesis, we explore the idea of leveraging information retrieved from lexical resources to solve lexical semantic tasks. In particular, *we demonstrate that that augmenting the input context with glosses retrieved from lexical re-*

sources improves the performance on two lexical semantic tasks: lexical substitution and idiomaticity detection. Our hypothesis is backed by experiments that provide empirical evidence supporting our claims.

The remainder of the chapter is organized as follows. First, we provide some background information on the fundamental concepts related to our work. Then we briefly describe the main contributions of this thesis. Finally, we provide an outline of the rest of the thesis.

1.1 Background

In this section, we will provide an overview of some key concepts. Specifically, we will discuss lexical resources, word sense disambiguation, and entailment.

1.1.1 Lexical Resources

Lexical resources are important tools for NLP and computational linguistics. They provide a large collection of lexical data, including the meanings, relationships, and semantic properties of words. These resources are typically organized into *wordnets*, which are large semantic networks that link distinct concepts that can be expressed by more than one word (Miller, 1995). A basic semantic unit in a wordnet is a synonym set, or synset. It consists of a set of words that share the same concept and can be used interchangeably in some contexts.

Princeton WordNet (Miller, 1995; Fellbaum, 1998) is one of the first and most influential wordnets. It has been carefully curated by experts to cover synsets for four main open-word classes: nouns, verbs, adjectives, and adverbs. It provides useful lexical information about synsets, such as glosses (short definitions), usage examples, and semantic relationships.

In addition to synonymy, Princeton WordNet covers various other semantic relationships. One of them is hyponymy/hypernymy, which describes the relationship between synsets where the words in a synset (hyponyms) are specific instances of words in another synset (hypernyms) that express a more general concept. Another relationship covered is meronymy/holonymy, which refers

to the relationship between synsets where the words in one synset represent a part (meronyms) of another synset that represents the whole (holonyms). A complete list of semantic relationships can be found on the Princeton WordNet website [1].

BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic network that connects words and concepts across various languages. Like Princeton WordNet, it uses synsets to represent concepts, but BabelNet goes beyond that by incorporating multilingual lexicalizations. Additionally, BabelNet automatically links lexical information from several sources, such as WordNet, Wikipedia, and Wikidata, thereby achieving a more comprehensive coverage of concepts. BabelNet 5.2, for instance, covers 520 languages and has approximately 22 million synsets [2].

1.1.2 Word Sense Disambiguation

Word sense disambiguation (WSD) is a fundamental problem in computational lexical semantics. It refers to the task of automatically determining the most appropriate meaning of a word in context from a predefined sense inventory (Bevilacqua et al., 2021). The correct identification of word senses can be challenging due to the inherent ambiguity of language. For example, the word *crane* can refer to a bird or construction equipment, depending on the surrounding context. Accurate identification of word senses can help in other downstream NLP applications, such as semantic role labeling, machine translation, and question answering, among others.

The approaches for WSD can be broadly categorized into two groups: knowledge-based and supervised (Bevilacqua et al., 2021).

Knowledge-based methods leverage semantic networks, such as WordNet or BabelNet, in which synsets act as nodes and the relationships between them as edges. These approaches then employ various graph algorithms, such as random walks (Agirre et al., 2014; Scozzafava et al., 2020) and game-theoretic algorithms (Tripodi and Navigli, 2019) or contextualized sense embeddings

[1] `https://wordnet.princeton.edu/documentation/wninput5wn`
[2] `https://www.babelnet.org/statistics`

(Wang and Wang, 2020) to disambiguate target words in contexts. One significant advantage of knowledge-based methods is that they do not require training data and are thus useful for low-resource languages.

Supervised models, on the other hand, take advantage of the annotated data to learn the mapping between words in contexts and their senses. Today, supervised systems significantly outperform knowledge-based systems in languages that have adequate training data. Most supervised models take advantage of recent advances in language modeling and use pre-trained transformer models (Vaswani et al., 2017). For example, the highest-performing model, ConSec (Barba et al., 2021), uses BART (Lewis et al., 2020) to jointly encode the context that contains the target word and all its possible definitions and extract a span associated with the most suitable definition.

1.1.3 Entailment

Entailment refers to the relationship between two sentences where the meaning of one sentence can be inferred from the meaning of the other sentence. It has become a very useful framework to reason about semantic relationship between sentences and has found applications in question answering, information extraction, machine translation, and summarization (Poliak, 2020).

More formally, a premise (P) entails a hypothesis (H) if a human reader of P would infer that H is most likely true (Dagan et al., 2005). Entailment is denoted as $P \models H$. For example, the premise "the water is boiling" entails the hypothesis "the water is hot" because if the water is boiling, it must be hot. This definition of entailment assumes common human understanding of language, as well as common background knowledge. Entailment is a directional relation, which means that $P \models H$ does not imply $H \models P$. However, if $P \models H$ *and* $H \models P$ then H and P are semantically equivalent: $P \equiv H$ (MacCartney, 2009).

Lexical entailment is a subset of textual entailment that specifically examines the relationship between a premise and a hypothesis where the two differ by a single word or phrase (Kroeger, 2018). It has previously been established that words in context often entail their synonyms, hypernyms, and hyponyms

(Geffet and Dagan, 2005). However, the ability of hypernyms and hyponyms to maintain entailment depends on *monotonicity* of the context. In upward monotone contexts (1), substituting a target word with a more general term maintains the truth of the sentence. In contrast, in downward monotonic contexts (2), a more specific term can replace the target word while preserving the truth (Yanaka et al., 2019). This shows that lexical entailment is a complex phenomenon influenced by the context.

1. I saw a [**penguin** ↑] $\models$ I saw a **bird**
2. All [**birds** ↓] are warm-blooded $\models$ All **penguins** are warm-blooded

1.2 Contributions

In this section, we provide a brief overview of our approaches to two semantic tasks, which represent the primary contributions of our thesis.

1.2.1 Idiomaticity Detection

Identifying idiomatic multiword expressions is a challenging task due to the non-compositional nature of these expressions. To solve this task, we propose a binary sequence classifier that leverages the definitions of the individual words within the target multiword expression, which are obtained from a lexical resource. Furthermore, we devise a type-based heuristic that exploits the fact that some multi-word expressions can be inherently literal or idiomatic. The experimental results show that adding glosses into the input and using our proposed heuristic leads to better performance compared to the sequence classifier model that does not utilize this information, in same settings.

1.2.2 Lexical Substitution

Existing definitions of lexical substitutes are often vague or inconsistent with gold annotations. We propose a new definition which is grounded in the concept of entailment; namely, that the sentence that results from the substitution should be in the relation of mutual entailment with the original sentence. We

argue that the new definition is well-founded and supported by previous work on lexical entailment. We empirically validate our definition by verifying that it covers the majority of gold substitutes in existing datasets. Based on this definition, we create a new dataset from existing semantic resources. Finally, we propose a novel data augmentation strategy motivated by the definition, which relates the substitutes to the sense of the target word by incorporating glosses and synonyms directly into the context. Experimental results demonstrate that our augmentation approach improves the performance of a lexical substitution system on existing benchmarks.

1.3 Outline

The remainder of this thesis is organized as follows: Chapter 2 provides a description of the idiomaticity detection task, as well as an overview of our proposed approaches to address it. Chapter 3 presents our definition of lexical substitution, along with a detailed explanation of our augmentation method for lexical substitution. Finally, Chapter 4 offers concluding remarks on this thesis.

Chapter 2

Idiomaticity Detection

In this chapter, we describe our system for SemEval 2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. (Tayyar Madabushi et al., 2022). We participated in Subtask A which involves classifying multi-word expressions (MWEs) in context as either *idiomatic* or *literal*. Each instance in the data includes an MWE (e.g., closed book), its language, and its context, composed of the three surrounding sentences.

Idiomaticity has long been a topic of great interest in the fields of psycholinguistics, linguistics, developmental psychology, neuropsychology, and computer science due to its widespread usage (Everaert et al., 2014). Although its exact definition is often disputed (Nunberg et al., 1994; Wulff, 2008) and is not explicitly stated in the task description, idiomaticity is widely identified with non-compositionality. We define an idiomatic expression as a multi-word expression (MWE) whose meaning is non-compositional, i.e., cannot be derived from the meaning of its components. For example, the meaning of the phrase "fish story" used in an idiomatic sense "an incredible or far-fetched story" cannot be deduced from the meaning of the individual words "fish" and "story". In addition, the shared task considers all proper noun MWEs (e.g. *Eager Beaver*) as literal.

Idiomatic MWEs are very common in many languages. However, modern computational models are still having a hard time fully capturing the non-compositional nature of idiomatic MWEs (Shwartz and Dagan, 2019; Garcia et al., 2021). Explicit identification of idiomatic phrases seems to be a

promising first step towards solving this issue since it would allow processing idiomatic MWEs differently compared to compositional phrases. Specifically, idiomaticity detection could potentially improve machine translation (to translate non-compositional phrases as a unit), word-sense disambiguation (to avoid assigning senses to individual words in a non-compositional phrase), and semantic parsing (to identify complex predicates and their arguments) systems (Cordeiro et al., 2016).

To solve this task, we propose using a BERT-based (Devlin et al., 2019) binary classifier that takes a context sentence, a target MWE, and glosses of all possible senses of all individual words in the target MWE as input. This method follows from the intuition that the meaning of a given MWE occurrence is related to any of the existing sense glosses of its component words *only if the expression is compositional*. Therefore, the addition of glosses to the context of the expression should help the classifier decide whether the MWE is used in a literal or idiomatic sense. We refer to this method as DEFBERT.

Our results provide evidence that using glosses from existing lexical resources is beneficial for idiomaticity detection. In particular, our method, when combined with a type-based UNATT heuristic, is among the top-scoring submissions in the one-shot setting. The heuristic is based on the observation that some MWEs are inherently idiomatic or literal, regardless of their context, which is confirmed by our analysis of the development set annotations.

The contributions of our work are as follows.

- We build a BERT-based sequence classifier model that takes advantage of glosses of individual words in the target MWE.
- We devise a type-based heuristic that takes into account that some MWEs are inherently idiomatic or literal.
- We evaluate the proposed model on the dataset provided by Tayyar Madabushi et al. (2022) and show that the addition of glosses indeed improves the performance.

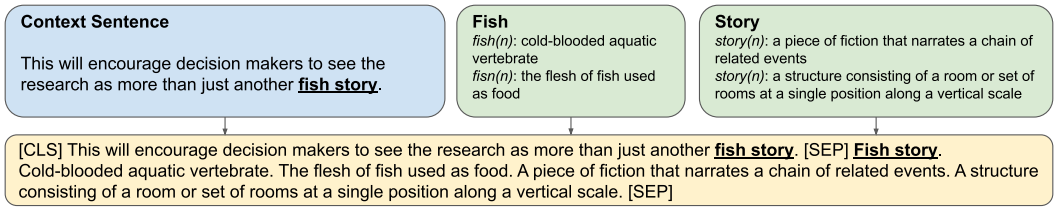

Figure 2.1: An example of DEFBERT input. The input is constructed by appending corresponding target MWEs and the glosses of individual words in these MWE to the context sentence and separating them using "[SEP]" special token. The input is then passed to a binary sequence classifier. For the sake of brevity, only a subset of glosses was shown for each word.

2.1 Related Work

Early attempts to represent idiomatic MWEs involve treating idiomatic phrases as individual tokens and learning the corresponding static embeddings (Mikolov et al., 2013). However, Cordeiro et al. (2016) show that the effectiveness of this method is limited by data sparsity for longer idiomatic expressions. Furthermore, Shwartz and Dagan (2019) and Garcia et al. (2021) conclude that idiomaticity is not yet accurately represented even by contextual embedding models. Tayyar Madabushi et al. (2021) create a new manually labeled dataset containing idiomatic and literal MWEs, and propose a method based on a pre-trained neural language model.

Another line of research used lexical translations for the detection of idiomaticity. For example, Moirón and Tiedemann (2006) measures semantic entropy in bitext alignment statistics, while Salehi et al. (2014) predict compositionality by presenting an unsupervised method that uses Wiktionary translation, synonyms, and definition information.

2.2 Methods

In this section, we describe the baseline and our method for idiomaticity detection.

2.2.1 Baseline

We re-implemented the mBERT (Devlin et al., 2019) classifier baseline following the methodology of Tayyar Madabushi et al. (2021). The model takes the context sentence and the relevant MWE as an input and outputs a label indicating the idiomaticity of the target MWE. The input sequence is constructed by concatenating the corresponding MWE at the end of the context sentence after the special [SEP] token.

It is important to note the differences between our re-implementation and the official baseline provided by the task organizers. In the official baseline, the organizers add the target MWE as an additional feature in the one-shot setting, but not in the zero-shot setting. Furthermore, the organizers include the sentences preceding and succeeding the target sentence only in the zero-shot setting. In our re-implementation, we add the target MWE and exclude the preceding and succeeding sentences in both zero-shot and one-shot settings.

2.2.2 defBERT

Our proposed method, which we refer to as DEFBERT, extends the baseline model by adding glosses of all possible senses of each individual word in the target MWE to the input of the classifier. The intuition behind this method is that the addition of glosses to the input should help the classifier decide if the meaning of the target MWE can be deduced from the definitions of the individual words,i.e. if it is compositional. In the example in Figure 2.1, the disparity between the context in which *fish story* appears, and the glosses of the various senses of the words *fish* and *story* indicates that the MWE is idiomatic in this context.

The intuition for this method is in line with the way non-native speakers can identify idiomatic expressions, provided they understand the standard meanings of the words which comprise them. Suppose that the vocabulary of a non-native speaker covers most of the essential words necessary to understand a language but not idiomatic expressions. Even if the speaker cannot deduce the meaning of an idiomatic expression in context, they can guess that the

expression was used in an idiomatic sense because individual words of this expression do not make sense in the given context.

2.2.3 Unattested Heuristics

In the one-shot setting, we also use a type-based heuristic that we refer to as UNATT. The intuition behind this heuristic is that certain MWEs are inherently idiomatic or literal, regardless of the context in which they appear. If the training data has no example of an MWE in a particular class, the heuristic exploits this fact as evidence that the MWE should always be classified as the opposite attested class. For example, this heuristic always classifies *life vest* as idiomatic and *economic aid* as literal, as these are the only classes in which these MWEs appear in the training data. In practice, since UNATT does not return a classification output if the training set contains instances that belong to either class, this heuristic must be used in combination with another method.

2.3 Dataset

We evaluate the models on the dataset provided by the SemEval 2022 Task 2 organizers (Tayyar Madabushi et al., 2022). The dataset contains context sentences, target MWEs, and labels that indicate the idiomaticity of the MWE in the English and Portuguese languages (the test split additionally includes examples in the Galician language). Label 0 indicates an idiomatic MWE, and label 1 indicates a non-idiomatic MWE including proper nouns. The dataset also provides previous and next sentences for context. The dataset was split into training, development, and test sets. The training set contains 4491 zero-shot examples, 140 one-shot examples associated with the development set, and 209 one-shot examples associated with the test set. The development set contains 739 examples, and the test set contains 2342 examples. In the zero-shot setting, the MWEs in the training set do not overlap with the MWEs in the development and test sets. In the one-shot setting, they overlap. Here is a sample entry from the dataset: *"If the UK were to leave, it could trigger*

a chain reaction" is the context sentence labeled idiomatic where the target MWE is *"chain reaction"*.

2.4 Experiments

We now describe our experiments, including the tools and resources, the experimental setup, the results, and a discussion of our findings.

2.4.1 Baseline and defBERT

We fine-tune the mBERT-based (Devlin et al., 2019) models using the binary classification objective on the labeled training dataset. In the zero-shot setting, we train the models only on the zero-shot training set. In the one-shot setting, we train the models on both the zero-shot and one-shot training sets. In particular, we fine-tune the models for 20 epochs with a maximum sequence length of 256, a learning rate of 2e-5, and a per device batch size of 16 using the HuggingFace Transformers library.[1]

2.4.2 Lexical Resources

We use BabelNet (BN; Navigli and Ponzetto, 2010, 2012), and Open Multilingual WordNet (OMW; Bond and Foster, 2013) as lexical resources to retrieve glosses. We access OMW through the NLTK interface Bird et al. (2009)[2] and BN 4.0 through the Java API.

Both BN and OMW contain English glosses for most concepts, but the availability of glosses in other languages varies. In particular, OMW does not contain Portuguese or Galician glosses. With BabelNet, we experiment with two techniques: using English glosses for all languages, and using glosses from the language of the instance, i.e. the source language, when available. I refer to these variants as "BN-EN" and "BN-SRC", respectively. Since DEFBERT uses a multilingual pre-trained language model, it can seamlessly handle input from multiple languages. Furthermore, because of the relatively poor coverage

[1] `https://huggingface.co`

[2] `https://www.nltk.org/api/nltk.tag.html`

of Galician in OMW (only 54% glosses are available in this language) and its close relationship with Portuguese, we experiment with processing Galician instances as if they were Portuguese.

2.4.3 Development Experiments

	Method	Zero-Shot EN	Zero-Shot PT	One-Shot EN	One-Shot PT
0	Baseline	66.2	63.9	87.0	86.7
1	mBERT	74.6	62.5	85.7	85.9
2	defBERT BN-SRC	75.5	64.8	85.4	86.7
3	defBERT BN-EN	75.3	66.4	87.6	86.6
4	defBERT	74.8	64.5	87.1	84.5
5	UNATT + defBERT	-	-	**92.0**	**87.7**

Table 2.1: Macro F1 scores calculated on the development dataset in the zero-shot and one-shot settings. Where not otherwise specified, defBERT is in the OMW-EN configuration.

Our experiments with defBERT explore the impact of adding glosses to the mBERT model, including the source and language of the glosses. The results show that the addition of glosses can improve performance. defBERT with English glosses retrieved from BabelNet (row 3) improves the overall score on both EN and PT over the baseline (row 1) in all languages in the zero-shot setting. The effect is particularly pronounced for Portuguese in the zero-shot setting, where an improvement of nearly 4% is observed. The results suggest that English glosses may be preferable to glosses in the source language, a finding that would greatly simplify work on lower-resourced languages, where glosses simply may not be available.

Combining the predictions of the mBERT-based models with the UNATT heuristic improves the overall F1 score by approximately 4 points in the one-shot setting (cf. row 5 vs. row 4).

2.4.4 Test Set Results

Table 2.2 contains the experimental results on the test set. The combination of defBERT method with the UNATT heuristic (row 5) continues to per-

	Method	Zero-Shot EN	PT	GL	ALL	One-Shot EN	PT	GL	ALL
1	mBERT	**75.1**	63.3	**61.1**	68.2	90.0	83.6	86.6	87.7
2	defBERT BN-SRC	72.0	66.4	57.8	67.2	**95.7**	88.5	88.9	92.2
3	defBERT BN-EN	73.4	**68.4**	59.7	**69.5**	95.0	**89.3**	87.9	91.8
4	defBERT	71.0	65.6	56.5	66.5	92.4	86.7	88.5	90.1
5	unatt + defBERT	-	-	-	-	94.5	89.2	**91.2**	**92.4**

Table 2.2: The macro F1 scores on the test dataset obtained from the CodaLab system.

form well, achieving the best overall result, and the best result in Galician, demonstrating the applicability of the method to low-resource languages. Even without the unatt heuristic, defBERT achieves competitive results, though BN glosses then give better results. In fact, the results without unatt are just 0.2% below the best overall result.

2.4.5 Shared Task Results

Tayyar Madabushi et al. (2022) present the results of the shared task, including the performance of 20 teams in Subtask A in the zero-shot setting and the results of 16 teams in the one-shot setting. In addition, they reported the results of the official baseline model. Our best-performing model shows promising results, surpassing the baseline model in both settings. In fact, our model's performance ranks at 14th place in the zero-shot setting and an impressive 3rd place in the one-shot setting, demonstrating the competitiveness of our approach.

2.5 Error Analysis

We found that the defBERT method performs slightly better by approximately 1% F1, on literal instances compared to idiomatic instances in the one-shot setting. In other words, the method is less likely to make an error when given a literal instance. We speculate that this is explained by the model's consistent classification of proper nouns as literal expressions. Indeed, a proper noun is incorrectly identified in only one instance. The fraction of idiomatic vs. literal instances is 39% in English and 56% in Portuguese.

Manual analysis performed on the development set corroborates our hypothesis that most multi-word expressions are inherently idiomatic (e.g., *home run*) or literal (e.g., *insurance company*). Only about one-third of the expressions are ambiguous in the sense that they can be classified as either class depending on the context (e.g. *closed book*). Our judgments are generally corroborated by the gold labels, with the exception of proper nouns, which are consistently marked as literal. The UNATT heuristic, which is based on this observation, obtains a remarkable 98.3% precision and 55.8% recall on the set of 739 instances in the development set.

Chapter 3

Lexical Substitution

Lexical substitution is the task of finding appropriate substitutes for a target word in a given context sentence. This task was first introduced as an application-oriented alternative to the word sense disambiguation (WSD) task which does not depend on a predefined sense inventory (McCarthy, 2002). Lexical substitution has been applied in various tasks, such as word sense induction (Amrami and Goldberg, 2018), lexical relation extraction (Schick and Schütze, 2020), and text simplification (Al-Thanyyan and Azmi, 2021).

Numerous definitions have been used in the literature to describe lexical substitution. The existing formulations tend to be vague and/or are inconsistent with the evaluation datasets. For example, Hassan et al. (2007) and Roller and Erk (2016) leave the criteria for lexical substitution to the discretion of human annotators. Studies such as Sinha and Mihalcea (2009, 2014) and Hintz and Biemann (2016) require substitutes to be synonyms, which creates a discrepancy with established lexical substitution benchmarks that allow annotators to provide slightly more general terms (hypernyms) (McCarthy, 2002; Kremer et al., 2014). Most prior work requires substitutes to preserve the meaning of the original sentence (McCarthy and Navigli, 2007; Giuliano et al., 2007; Szarvas et al., 2013a,b; Kremer et al., 2014; Melamud et al., 2015; Garí Soler et al., 2019; Zhou et al., 2019; Lacerra et al., 2021; Michalopoulos et al., 2022; Seneviratne et al., 2022; Wada et al., 2022). However, as we show in this work, not all gold substitutes necessarily preserve the meaning of the sentence taken in isolation.

We propose a definition of lexical substitution that is more precise and well-founded. Our aim is not only to address the inconsistency in the literature, but also to align the task definition with established evaluation datasets. We draw on insights from natural language inference (NLI), which provides a framework for understanding the semantic relationship between sentences and words. According to our definition, the sentence that results from a lexical substitution must be in the relation of *mutual entailment* with the original sentence. For example, *position* is a suitable substitute for *post* in the sentence "I occupied a *post* in the treasury" because occupying a post in this context entails occupying a position, *and vice versa*. The entailment criterion takes into account the implicit background knowledge (Dagan et al., 2005), which allows lexical substitution to generalize over simple synonym replacement, encompassing a wider range of semantic relations, such as hypernymy and meronymy (Geffet and Dagan, 2005).

The classification of the entailment relation between two sentences requires the identification of the target word's sense. For example, *position* is a proper substitute for *post* only if it is used in the sense corresponding to "job in an organization". Based on this observation, we develop an augmentation method that helps to ground the substitutes by incorporating glosses and synonyms of the target word's sense directly into the context. Since the word sense is latent, the method leverages a WSD system to account for the probabilities of each candidate sense.

We show the effectiveness of the proposed definition and our augmentation method through experiments on existing lexical substitution datasets. Our analysis indicates that the proposed definition encompasses gold substitutes that could not previously be explained by existing definitions. Furthermore, our empirical evaluation shows that our augmentation method improves the performance on the lexical substitution benchmarks by up to 4.9 F1 points, surpassing the previous state-of-the-art models in certain settings.

The main contributions of this paper are as follows. (1) We propose a task formulation for lexical substitution that is based on entailment and is more suitable for use with existing datasets. (2) We construct a new automatically

constructed dataset for lexical substitution induced by our definition. (3) We demonstrate empirically that augmenting the context with glosses and synonyms can improve the performance of lexical substitution systems.

3.1 Related Work on Lexical Substitution

In this section, we review the available datasets and provide a brief overview of the methods.

3.1.1 Datasets

The first English lexical substitution dataset, proposed by McCarthy and Navigli (2007) in SemEval-2007 Task 10 (SE07), consists of 2003 context sentences with one target word per sentence. The authors ask the annotators to provide substitutes for the target word in context that preserve the original meaning of the sentence.

Biemann (2012) construct Turk Bootstrap Word Sense Inventory (TWSI) that encompasses a sense inventory induced by lexical substitutes for 1,012 common English nouns. It is created by annotating 25,851 sentences with lexical substitutes using Amazon Mechanical Turk platform, which can be used as a dataset for lexical substitution.

Kremer et al. (2014) present CoInCo, an "all-word" lexical substitution dataset where all content words of a corpus are annotated with substitutions. The authors argue that the all-word setting provides a more realistic distribution of target words and their senses. It is important to note that (McCarthy and Navigli, 2007) and (Kremer et al., 2014) explicitly allow annotators to provide phrases or more general words when they could not think of a good substitute.

The SWORDS dataset (Lee et al., 2021) is based on the CoInCo dataset but uses a slightly different annotation approach. Instead of relying on annotators to come up with substitutes from their memory, they are asked to provide only a binary judgement if they would use a given candidate substitute in the place of the target word. The data set contains 1,250 context sentences, each

of which contains a target word.

The task of lexical substitution is not limited to the English language, and datasets have also been created for other languages. For example, Toral (2009) create a lexical substitution dataset in Italian, Cholakov et al. (2014) build a dataset in German, and Miller et al. (2016) provide sense annotations for the German dataset. Additionally, a cross-lingual English-Spanish dataset is introduced in the SemEval-2010 Task 2 (Mihalcea et al., 2010). In this dataset, the target words and sentences are in English, while the gold substitutes are in Spanish.

3.1.2 Methods

Early methods retrieved candidate substitutes from lexical resources such as WordNet (Miller, 1995). Various approaches to rank candidate substitutes used web queries (Zhao et al., 2007; Martinez et al., 2007; Hassan et al., 2007), ngram models (Giuliano et al., 2007; Yuret, 2007; Dahl et al., 2007; Hawker, 2007; Hassan et al., 2007), latent semantic analysis (Giuliano et al., 2007; Hassan et al., 2007), delexicalized features (Szarvas et al., 2013a), and word embeddings (Melamud et al., 2015, 2016; Roller and Erk, 2016)

Pre-trained neural language models (NLMs) have greatly advanced the field, and their contextualized embedding representation has become a standard for many tasks, including lexical substitution. Garí Soler et al. (2019) used contextual embeddings from ELMo (Peters et al., 2018) to calculate similarity between the target and candidate substitutes. To fix the bias toward the target word, Zhou et al. (2019) applied a dropout embedding policy that partially masks the target word's BERT embedding. Arefyev et al. (2020) proposed combining masked language model probability score with a contextual embedding-based proximity score. Lacerra et al. (2021) proposed training a supervised sequence-to-sequence model that takes a context sentence containing a target word as input and outputs a comma-separated list of substitutes. Wada et al. (2022) made use of contextualized and decontextualized embeddings (the average contextual representation of a word in multiple contexts). Yang et al. (2022) injected information about the target word in context and

used BERT to generate initial candidates. Furthermore, they proposed using RoBERTa trained on MNLI dataset to calculate semantic similarity score to further refine the ranking.

Recent proposals of Michalopoulos et al. (2022) and Seneviratne et al. (2022) leverage knowledge from WordNet to improve the quality of substitutes retrieved from pretrained neural language models. Although these two approaches are similar to our method, there are some significant differences. Michalopoulos et al. (2022) injected synonyms by linearly interpolating their contextual embeddings, while we insert synonyms and glosses directly into the context. Seneviratne et al. (2022) and the remaining approach of Michalopoulos et al. (2022) use knowledge from WordNet only in the ranking stage after candidates had been generated from an NLM. In contrast, our approach injects WordNet information into the NLM's input from the beginning, which may produce more relevant candidates initially.

3.2 Entailment-Based Lexical Substitution

In this section, we present the theoretical formulation of the proposed definition, and demonstrate its suitability through empirical validation.

3.2.1 Lexical Substitution Definition

We anchor our definition of lexical substitution in textual entailment. Let C_t be a context sentence that contains a target word t, and let C_w be the same context sentence where t is replaced with a word or phrase w. We define w as a lexical substitute for t in C_t if and only if C_t entails C_w and C_w entails C_t:

$$\text{LexSub}(C_t, w) \Leftrightarrow C_t \models C_w \text{ and } C_w \models C_t$$

This binary definition can be adapted to the task of substitute generation by considering a finite set of all words and short phrases. Specifically, the output of the generation task would consist of all candidate substitutions that satisfy the above condition.

Although entailment is recognized as an important substitutability criterion within the NLI community (Geffet and Dagan, 2004, 2005; Zhitomirsky-Geffet and Dagan, 2009), it has been largely overlooked in lexical substitution. A notable exception is Giuliano et al. (2007), who recognize the significance of the relationship between lexical substitution and entailment. Although their mutual textual entailment criterion is similar to ours, we disagree with their conclusion that the mutual equivalence requirement restricts substitutes to synonyms only. Next, we show that this criterion not only extends beyond word synonymy, but also naturally allows for the integration of common-sense reasoning and knowledge about the world.

3.2.2 Semantic Equivalence

In this section, we explicitly spell out our assumptions about the relationship between lexical substitution and the meaning preservation criteria.

The first proposition states that all contextual synonyms are good substitutes.

Proposition 1. *If t and w express the same concept in C then w is a lexical substitute for t in C.*

Proof. When we replace a target word with another word that expresses the same concept in a given context, the truth conditions of the sentence do not change. This is because the truth conditions are determined by the relationships between concepts that are expressed in the sentence. Therefore, the mutual entailment between C_w and C_t must hold, which by our definition implies that w is a lexical substitute for t in the context C. □

If words express the same concept in some context, they must belong to the same wordnet synset. A wordnet is a lexical ontology in which words are grouped into sets of synonyms (synsets), each representing a distinct concept (Miller, 1995). The suitability of contextual synonyms with lexical substitution provides a theoretical basis for the use of wordnets to generate substitutes (McCarthy and Navigli, 2007).

The implication in Proposition 1 is unidirectional; that is, not all substitutes must be synonyms.

Proposition 2. *If w is a lexical substitute for t in C then t and w do not necessarily represent the same concept in C .*

To prove this point, it suffices to show one counter-example. Consider the following sentence from the SWORDS dataset: "Those hospitals were not for us. They were for an *expected* invasion of Japan." where the word *planned* is among gold substitutes for the target word *expected*. While the verbs *expect* and *plan* are not synonyms, this particular substitution is correct considering the broader historical context of World War II. From the point of view of the US military, the invasion was both planned and expected. Thus, although the two words do not express the same concept, the corresponding sentences entail each other.

Taken together, these two propositions imply that synonymy within a narrow context is a sufficient but not a necessary condition for mutual entailment between the sentences. Thus, mutual entailment provides a more flexible criterion for substitution, which extends beyond mere synonymy and meaning preservation. The mutual entailment criterion captures the nuances of lexical substitution better than the existing definitions which are based on strict meaning preservation, because it takes into account both context *and* background knowledge. This is essential to identify a wider range of substitutions in scenarios such as the ones described above. Furthermore, this definition may facilitate the job of annotators by breaking down lexical substitution into two concrete entailment conditions, which are easier to reason about.

3.2.3 Empirical Validation

To validate our proposed definition for lexical substitution, we perform a manual analysis of a random sample of 20 gold substitutes labeled "acceptable" by humans and their corresponding contexts from the SWORDS dataset. We aim to assess whether these substitutes are adequately covered by our definition.

Our manual analysis, presented in Table 3.1, shows that 19 substitutes are successfully covered by our definition. One substitute (row 10), which is not covered by our definition, is also not covered by the existing definition of meaning preservation. Furthermore, it seems that the proposed gold substitu-

Row	Context C_t	Substitute w	$Cw \models Ct$	$Ct \models Cw$
1	I **swear**. They all thought I was Steve Martin .	vow	Yes	Yes
2	"Excuse me," I **said**, ignoring Nepthys' warning look,	mention	Yes	Yes
3	...many clinical psychologists already receive **inadequate** training	insufficient	Yes	Yes
4	Now, will you tell me how you **know** my family?	have knowledge of	Yes	Yes
5	Please, **walk** this way.	proceed	Yes	Yes
6	It's **okay**, you can trust him.	alright	Yes	Yes
7	They were for (an **expected** invasion of Japan)	planned	Yes*	Yes
8	...you **know** some way to locate the undead, don't you ?	have	Yes	Yes
9	But in some **areas**, the seabass are being overfished.	location	Yes	Yes
10	I am glad to be out of the favor-trading scene for half a **minute**	moment	No	No
11	The Persian Gulf War **destroyed** much of the country's medical infrastructure	devastate	Yes	Yes
12	That was **very** kind of her.	exceedingly	Yes	Yes
13	...considers prescriptive authority a **logical** extension of psychologists' role as health-care providers	rational	Yes	Yes
14	...we simply want to **discover** whether this individual is in fact, a vampire.	find	Yes	Yes
15	Energy Secretary Bill Richardson went to Baghdad in 1995 while a **representative** for New Mexico.	elected official	Yes	Yes*
16	But they **liked** the way (Jose) has played and they're giving him a chance.	enjoy	Yes	Yes
17	Karnes had his own Jeep, and **went** to the beach	head	Yes	Yes
18	Ochoa has played in the majors for five different teams **starting** in 1995	commence	Yes	Yes
19	The new plant is part of IBM 's push to gain a **strong** lead in chip-making beyond the personal computer business	formidable	Yes	Yes
20	He ran down a hallway and slipped behind one of the **doors**	doorway	Yes	Yes

Table 3.1: The table contains a random sample of 20 substitutes from the SWORDS dataset. The target words are in bold. * denotes that the specified entailment holds if we assume relevant background knowledge. Instances not covered by existing definitions are highlighted red.

tion might not be of good quality in the first place, so we can treat it as an

annotation error rather than as an exception to our definition.

We also observe that the two substitutes (rows 7 and 15) that are not covered by existing definition of meaning preservation are covered by our definition. For example, consider the context in row 15. Generally speaking, the word *representative* and the phrase *elected official* do not have the same meaning. If someone is an elected official, it does not necessarily mean that the person is a congress representative. However, the sentence provides enough historical context to allow for substitution. On the basis of the context, we could infer that the elected official position refers to a congress representative. This observation is significant because it shows that our proposed definition covers substitutes that previously did not match the existing definitions.

3.2.4 Dataset Induced by Entailment

Based on Proposition 1, we propose to use synonyms from lexical resources such as WordNet to construct a new lexical substitution dataset, which we refer to as WNSub. This is because replacing the target word with synonyms is guaranteed to generate sentences that are mutually inferrable from the original sentences.

To generate the WNSub dataset, we use SemCor (Miller et al., 1994), the largest corpus manually annotated with WordNet senses. The sense annotations are crucial for our dataset as synonyms are defined in relation to word senses rather than word lemmas. For example, for the sentence "can your insurance company aid you in reducing administrative costs?" we retrieve substitutes *help* and *assist* from the synset that corresponds to the annotated sense of the target word *aid*. In total, we obtain 146,303 sentences with 376,486 substitutes.

Although synonyms do not necessarily capture all aspects of lexical substitution, WNSub can be used for pre-training supervised systems, in combination with other datasets. We verify this claim experimentally in Section 3.4.3.

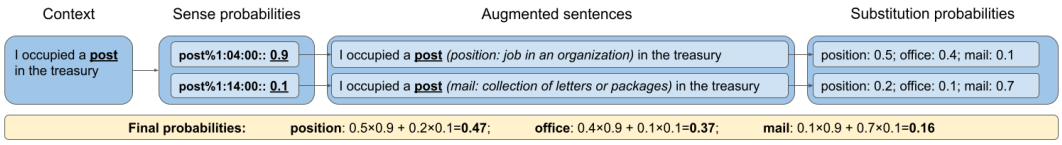

Figure 3.1: An example of augmenting a context with target word definitions, and calculating substitute scores. For brevity, not all candidate senses and substitutes are shown.

3.3 Sense-based Augmentation Method

In this section, we describe our sense-based augmentation method for lexical substitution. Our approach is based on the observation that knowing the sense of the target word is key to deciding whether a substitution induces an entailment relation between the two sentences. For example, *position* is a proper substitute for *post* in some context only if the latter is used in the sense corresponding to "job in an organization". We posit that inserting sense glosses directly into the context will help lexical substitution systems identify substitutes that are mutually entailed by the original context. Our hypothesis is supported by prior findings that this technique works well for semantic tasks such as WSD (Huang et al., 2019) and idiomaticity detection (Hauer et al., 2022).

Our method is based on two stand-alone modules: a WSD system, and a lexical substitution generation system. The method is sufficiently flexible to incorporate new systems as the state of the art on those two tasks continues to improve. The only requirement is that these systems output probabilities for each candidate sense or substitute.

The formula below is used to combine the probabilities from the two systems. Figure 3.1 shows an example of soft constraint augmentation. Let C_t be a context sentence containing the target word t, w be a candidate substitute, and $s \in senses(t)$ be a candidate sense for t in C_t. Under the assumption that the substitutes depend on the sense of the target word, the conditional probability $P(w|C_t)$ can be derived by marginalizing the senses out:

$$P(w|C_t) = \sum_{s \in senses(t)} P(w|C_t, s) \times P(s|C_t)$$

In the equation above, we model $P(s|C_t)$ using a WSD system, and obtain $P(w|C_t, s)$ from a lexical substitution system that operates on the context augmented with sense information.

We experiment with two types of constraints: hard and soft. In the hard-constraint approach, a WSD system is used to identify the most likely sense of the target word, which is effectively assigned the probability of 1.0. Next, the glosses and synonyms corresponding to this sense are retrieved from a lexical resource and inserted in parentheses after the target word. This augmented context is then passed to a lexical substitution system, which generates substitutes along with their substitute probabilities. In the soft-constraint approach, for each possible sense of the target word, a WSD system first computes its probability, the context is augmented with glosses and synonyms of that sense, and finally a lexical substitution system generates and assigns final probabilities to candidate substitutes using the formula above.

Soft constraint allows grounding of lexical substitutes in the target word senses, while taking into account the probability of each candidate sense. We posit that considering all candidate senses and their probabilities should work better than committing to a single most likely sense, by improving robustness against WSD errors. In addition, in some cases, the context itself may not provide enough information to reliably disambiguate the sense of the target word. We verify this hypothesis experimentally in the next section.

3.4 Experiments

In this section, we investigate the effectiveness of our dataset and augmentation method in improving the performance of lexical substitution systems. The experiments were conducted on a machine with two NVIDIA GeForce RTX 3090 video cards.

3.4.1 Evaluation Datasets and Metrics

We evaluate our methods using test splits from two benchmarks: the SemEval 2007 Task 10 (SE07) (McCarthy and Navigli, 2007) and SWORDS (Lee et al.,

2021). Each benchmark has its own set of evaluation metrics, which we will outline in this subsection.

The SE07 benchmark uses *best* and *oot* metrics, which measure the quality of the system's top-1 and top-10 predictions, respectively. These metrics assign weights to gold substitutes based on how frequently annotators selected them. The benchmarks also use *mode* variations of *best* and *oot*, which evaluate performance against a single gold substitute chosen by the majority of annotators, provided such a majority exists. We consider the *mode* metrics theoretically problematic because they disregard instances without an annotation majority, and because many instances could involve multiple equally valid substitutes,

The SWORDS benchmark uses F^{10} scores, the harmonic mean of precision and recall, calculated with respect to the system's *top 10 predictions* and *acceptable* (F_a^{10}) or *conceivable* (F_c^{10}) gold substitutes. A candidate is labeled as *conceivable* if it was selected by at least one annotator, and *acceptable* if selected by at least half of the annotators. Furthermore, the benchmark includes two evaluation settings: lenient and strict. In the lenient setting, any system-generated substitutes that are not in SWORDS are removed. In the strict setting, all system-generated substitutions are considered. The lenient settings were originally proposed to compare against "oracle" baselines whose predictions are guaranteed to be in SWORDS. We posit that the lenient setting provides an unreliable basis for measuring lexical substitution performance in real-world scenarios because systems are not provided with a predefined vocabulary of possible words that can occur during testing.

All existing evaluation metrics require a ranking mechanism to select top-k system predictions, which could be problematic for two reasons. First, there is a lack of clarity on objective criteria for ranking substitute words. For example, in the sentence *"the FBI <u>said</u> that explicit conversations about the scheme had been recorded"*, it is debatable whether *disclosed* is a better substitute for *said* than *declared*. Second, the existing metrics reward systems for generating a specific number of candidates, regardless of how many substitutes actually exist. This can lead to an inaccurate representation of the system's ability to

generate correct substitutes.

Despite these limitations, our method builds upon existing systems that have been optimized using these metrics, so we have no choice but to use them for the evaluation. However, we posit that it would be beneficial for future lexical substitution systems to consider metrics, which do not depend on substitution ranking, such as the standard F1 score calculated with respect to all predicted substitutes, .

3.4.2 Comparison Systems

On the SE07 dataset, we compare against KU (Yuret, 2007), supervised learning (Szarvas et al., 2013a), BERT for lexical substitution (Zhou et al., 2019), GeneSis (Lacerra et al., 2021), LexSubCon (Michalopoulos et al., 2022), and CILex (Seneviratne et al., 2022). The reported results are sourced from Michalopoulos et al. (2022) and Seneviratne et al. (2022).

On the SWORDS dataset, we compare against GPT-3 (Brown et al., 2020) that uses "in-context" learning, a commercial lexical substitution system Word-Tune[1], and BERT baseline (Devlin et al., 2019) that produces substitutes according to the masked language modeling head. The results of these models are reported by Lee et al. (2021). We also include the results of Yang et al. (2022).

3.4.3 WNSub Experiments

The objective of the experiments with WNSub (Section 3.2.4) is to determine whether the dataset could enhance the performance of two supervised sequence-to-sequence lexical substitution models when used as a pre-training dataset.

The first model is our own implementation of a simple supervised sequence-to-sequence (seq2seq) model. It takes a context where the target word is tagged with two brace tokens, and generates a substitute word or phrase as a prediction. We use beam search to generate multiple likely substitutes. Our

[1] `https://www.wordtune.com`

underlying seq2seq model is *bart-large* (Lewis et al., 2020). We utilize the same set of hyperparameters for both pre-training and fine-tuning. Specifically, we train our model for 19,000 steps with a batch size of 64 and a learning rate of 4e-5.

The second model is GeneSis (Lacerra et al., 2021), also a sequence-to-sequence model. Unlike our model, GeneSis produces a comma-separated list of substitutes instead of single substitutes, and filters out words that are not in WordNet. It incorporates a fallback strategy in *oot* settings. When the model generates fewer than ten substitutes, additional words are retrieved from WordNet, and ranked using neural language model embeddings. To assess the model's performance based solely on annotated data, we disable both lexicon filtering and fallback strategy. We use their default settings for both pre-training and fine-tuning.

We evaluate the models using two different training approaches. In the baseline approach, we train the systems on existing datasets, specifically the CoInCo and TWSI datasets, following the methodology of Lacerra et al. (2021). In the pre-training approach (+ wns), we first pre-train the systems on the WNSub dataset, and then fine-tune on the union of the CoInCo and TWSI datasets. Our evaluation is on the SE07 test set only, as the SWORDS dataset includes instances from the CoInCo dataset.

The results on the SE07 test set in Table 3.2 indicate that pre-training on the WNSub dataset does improve the performance of both supervised models. The only exception is GeneSis evaluated in the *oot* setting, in which the credit for each correct guess is not divided by the number of guesses. Thus, there is no penalty for attempting to fill all 10 candidate substitutes, even if some of them can be incorrect. However, when evaluated using the standard F1 score that considers all predictions, pre-training does improve GeneSis' performance from 26.8 to 27.7 points. This suggests that the F1 metric may better reflect the quality of the system when they are not forced to produce a fixed number of substitutes.

Models	best	oot
Yuret (2007)	12.9	46.2
Szarvas et al. (2013a)	15.9	48.8
Zhou et al. (2019)	20.3	55.4
GeneSis (2021)	21.6	52.4
Michalopoulos et al. (2022)	21.1	51.3
Seneviratne et al. (2022)	**23.3**	56.3
WNSub experiments		
seq2seq baseline	9.7	44.0
+ WNSub	**10.7**	**44.8**
GeneSis*	19.2	**34.3**
+ WNSub	**19.6**	34.1
Augmentation experiments		
LexSubGen (2020)	21.7	55.1
+ soft constraint	**21.9**	**57.9**
Wada et al. (2022)	21.8	58.0
+ soft constraint	**22.0**	**58.4**

Table 3.2: Results on the SE07 test set. *To compare the supervised aspect of GeneSis, we disable vocabulary filtering and fallback strategy.

3.4.4 Augmentation Experiments

We use WordNet 3.0 (Miller, 1995) available via NLTK (Bird et al., 2009) interface to retrieve synonyms and glosses for the target word for our augmentation approach.

As our WSD system, we use ConSec[2] (Barba et al., 2021). The model jointly encodes the context containing the target word and all possible sense definitions and extracts the span of the definition that suits the target word the most. ConSec also leverages the senses assigned to nearby words to improve the performance. Since the original implementation outputs only predicted senses, we changed the source code to capture the probability scores for all candidate senses.

As our lexical substitution base system, we use LexSubGen[3] (Arefyev et al., 2020). The best-performing model in this paper uses XLNet (Yang et al., 2019) that contains 340M parameters and injects the target word information by combining the substitute probability from XLNet with the contextual

[2] `https://github.com/SapienzaNLP/consec`

[3] `https://github.com/Samsung/LexSubGen`

Models	F_a^{10}	F_c^{10}
GPT-3	22.7	36.3
WordTune	22.8	33.6
BERT	19.2	30.3
Yang et al. (2022)	18.3	28.7
LexSubGen (2020)	19.4	29.9
+ soft constraint	**21.5**	**34.8**
Wada et al. (2022)	24.5	39.9
+ soft constraint	**24.7**	**42.5**

Table 3.3: Results on the SWORDS test set.

embedding similarity of the substitute to the target word.

To test the generalizability of our approach, we also apply our data augmentation method to the model of Wada et al. (2022). Their model is based on the similarity of contextualized and decontextualized embedding that are calculated by taking the average contextual representation of a word in multiple contexts.

The results on the SE07 dataset in Table 3.2 show that our augmentation method leads to an improvement over the base model in both *best* and oot settings. In addition, our result of 58.4 establishes a new state of the art in the *oot* setting.

The results in Table 3.3 demonstrate that our augmentation approach produces an improvement over both LexSubGen of Arefyev et al. (2020) and the system of Wada et al. (2022) in both strict settings. Our results in the last row of Table 3.3 represent the new state of the art on the SWORDS dataset.

Overall, the experimental results demonstrate that both our augmentation approach and WNSub dataset improve the performance on the lexical substitution task.

3.4.5 Ablation Study

To assess the impact of incorporating our augmentation method, we conduct an ablation study on the SWORDS test dataset. We evaluate the contribution of synonyms and glosses by removing them individually in separate experiments. This allows us to investigate which type of information is most beneficial for

Models	F_a^{10}	F_c^{10}
LexSubGen	19.4	29.9
+ hard constraint	21.2	34.2
+ soft constraint	**21.5**	**34.8**
- gloss	20.6	32.7
- synonyms	21.1	33.6

Table 3.4: Ablation study on the SWORDS test set.

the task of lexical substitution.

The results of the ablation experiment are presented in Table 3.4. Removal of both synonyms and glosses simultaneously is equivalent to the baseline LexSubGen showed in the first row. The hard constraint approach that depends on a single most likely sense yields lower performance than the soft constraint, which is more robust to WSD errors, Furthermore, the results indicate that incorporating glosses in the context is more important than adding synonyms, which suggests that the former provide more information than the latter. Overall, the results of the ablation study provide further evidence that augmentation improves lexical substitution systems.

3.4.6 Error Analysis

We perform a manual error analysis of LexSubGen on a randomly selected sample of 20 instances from the SWORDS test split. We did not find any instance where the augmentation resulted in missed predictions compared to the base model without augmentation. However, we found one instance where the augmentation helped to identify two gold substitutes correctly. Specifically, the augmentation method aided in correctly identifying *overlook* and *neglect* as substitutes for *miss* in certain context.

Chapter 4

Conclusion

In this thesis, we have successfully leveraged the information retrieved from lexical resources to solve lexical semantic tasks. In particular, we have demonstrated that augmenting the input with glosses of the target word enhances performance on two lexical semantic tasks: lexical substitution and idiomaticity detection. Our findings provide strong empirical evidence to support our hypothesis that incorporating lexical information leads to superior performance on these tasks.

For idiomaticity detection, we have proposed DEFBERT, a binary sequence classifier that leverages glosses of individual words in the target MWE, and UNATT, a type-based heuristic that takes into account that some MWEs are inherently idiomatic or literal. The proposed method and its combination with type-based heuristics outperforms the baseline model, showcasing the utility of using glosses for the task of idiomaticity detection. Our top result ranks third overall in the one-shot setting in SemEval 2022 Task 2. The corresponding method is applicable to a wide variety of languages. It takes advantage of the ability of neural language models to seamlessly incorporate textual information such as glosses, even if it is expressed in a different language. These results strongly support our hypothesis that glosses of individual words can improve idiomaticity detection.

We have also provided a novel analysis of a definition of the lexical substitution task based on the concept of entailment, addressing the inconsistencies that exist between the definition of the task and the evaluation. We have

experimentally validated our definition and compared it with existing ones. Additionally, we have constructed a new training dataset, which is induced by our definition, from existing semantic resources. To improve the performance of lexical substitution systems, we proposed an augmentation approach that directly inserts glosses and synonyms into the context and demonstrated its effectiveness on both existing and newly constructed benchmarks. Further research could explore the generalizability of our approach to other lexical substitution systems and languages.

References

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

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels, Belgium. Association for Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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

Chris Biemann. 2012. Turk bootstrap word sense inventory 2.0: A large-scale resource for lexical substitution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4038–4042, Istanbul, Turkey. European Language Resources Association (ELRA).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Lexical substitution dataset for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1406–1411, Reykjavik, Iceland. European Language Resources Association (ELRA).

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

George Dahl, Anne-Marie Frassica, and Richard Wicentowski. 2007. SW-AG: Local context matching for English lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 304–307, Prague, Czech Republic. Association for Computational Linguistics.

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.

Martin Everaert, Erik-Jan Van der Linden, Rob Schreuder, Robert Schreuder, et al. 2014. *Idioms: Structural and psychological perspectives*. Psychology Press.

Christiane Fellbaum. 1998. Wordnet: An on-line lexical database and some of its applications.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564. Association for Computational Linguistics.

Aina Garí Soler, Anne Cocos, Marianna Apidianaki, and Chris Callison-Burch. 2019. A comparison of context-sensitive models for lexical substitution. In

Proceedings of the 13th International Conference on Computational Semantics - Long Papers, pages 271–282, Gothenburg, Sweden. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2004. Feature vector quality and distributional similarity. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 247–253, Geneva, Switzerland. COLING.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114, Ann Arbor, Michigan. Association for Computational Linguistics.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. 2007. FBK-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 145–148, Prague, Czech Republic. Association for Computational Linguistics.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413, Prague, Czech Republic. Association for Computational Linguistics.

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. UAlberta at SemEval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 145–150, Seattle, United States. Association for Computational Linguistics.

Tobias Hawker. 2007. USYD: WSD and lexical substitution using the Web1T corpus. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 446–453, Prague, Czech Republic. Association for Computational Linguistics.

Gerold Hintz and Chris Biemann. 2016. Language transfer learning for supervised lexical substitution. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 118–129, Berlin, Germany. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Paul Kroeger. 2018. *Analyzing meaning: An introduction to semantics and pragmatics*. Language Science Press.

Caterina Lacerra, Rocco Tripodi, and Roberto Navigli. 2021. GeneSis: A Generative Approach to Substitutes in Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10810–10823, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 2021. Swords: A benchmark for lexical substitution with improved data coverage and quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4362–4379, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.

David Martinez, Su Nam Kim, and Timothy Baldwin. 2007. MELB-MKB: Lexical substitution system based on relatives in context. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 237–240, Prague, Czech Republic. Association for Computational Linguistics.

Diana McCarthy. 2002. Lexical substitution as a task for WSD evaluation. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 089–115. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, Colorado. Association for Computational Linguistics.

George Michalopoulos, Ian McKillop, Alexander Wong, and Helen Chen. 2022. LexSubCon: Integrating knowledge from lexical resources into contextual embeddings for lexical substitution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1226–1236, Dublin, Ireland. Association for Computational Linguistics.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Tristan Miller, Mohamed Khemakhem, Richard Eckart de Castilho, and Iryna Gurevych. 2016. Sense-annotating a lexical substitution data set with ubyline. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 828–835, Portorož, Slovenia. European Language Resources Association (ELRA).

Begona Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

Roberto Navigli. 2018. Natural language understanding: Instructions for (present and future) use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5697–5702. International Joint Conferences on Artificial Intelligence Organization.

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

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.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Talgat Omarov and Grzegorz Kondrak. 2023. Grounding the lexical substitution task in entailment. In *submission*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Stephen Roller and Katrin Erk. 2016. PIC a different word: A simple model for lexical substitution in context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1121–1126, San Diego, California. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, Doha, Qatar. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8766–8774.

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

Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy, and Hanna Suominen. 2022. CILex: An investigation of context information for lexical substitution methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4124–4135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of the International Conference RANLP-2009*, pages 404–410, Borovets, Bulgaria. Association for Computational Linguistics.

Ravi Sinha and Rada Mihalcea. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering*, 20(1):99–129.

György Szarvas, Chris Biemann, and Iryna Gurevych. 2013a. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1131–1141, Atlanta, Georgia. Association for Computational Linguistics.

György Szarvas, Róbert Busa-Fekete, and Eyke Hüllermeier. 2013b. Learning to rank lexical substitutions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1932, Seattle, Washington, USA. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477. Association for Computational Linguistics.

Antonio Toral. 2009. The lexical substitution task at evalita 2009. In *Proceedings of EVALITA Workshop, 11th Congress of Italian Association for Artificial Intelligence, Reggio Emilia, Italy*.

Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 88–99, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Takashi Wada, Timothy Baldwin, Yuji Matsumoto, and Jey Han Lau. 2022. Unsupervised lexical substitution with decontextualised embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4172–4185, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Stefanie Wulff. 2008. *Rethinking idiomaticity: A usage-based approach*. A&C Black.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11613–11621.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Deniz Yuret. 2007. KU: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, Prague, Czech Republic. Association for Computational Linguistics.

Shiqi Zhao, Lin Zhao, Yu Zhang, Ting Liu, and Sheng Li. 2007. HIT: Web based scoring method for English lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 173–176, Prague, Czech Republic. Association for Computational Linguistics.

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

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.