Taxonomy of Problems in Lexical Semantics

Bradley Hauer and **Grzegorz Kondrak**
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,gkondrak}@ualberta.ca

Abstract

Semantic tasks are rarely formally defined, and the exact relationship between them is an open question. We introduce a taxonomy that elucidates the connection between several problems in lexical semantics, including monolingual and cross-lingual variants. Our theoretical framework is based on the hypothesis of the equivalence of concept and meaning distinctions. Using algorithmic problem reductions, we demonstrate that all problems in the taxonomy can be reduced to word sense disambiguation (WSD), and that WSD itself can be reduced to some problems, making them theoretically equivalent. In addition, we carry out experiments that strongly support the soundness of the concept-meaning hypothesis, and the correctness of our reductions.

1 Introduction

This paper proposes a taxonomy of several problems in lexical semantics, consisting of a clear definition of each task, and a theory-driven analysis establishing the relationships between them (Figure 1). The taxonomy includes word sense disambiguation (WSD), word-in-context (WiC), lexical substitution (LexSub), and word synonymy (Syn). We consider their monolingual, cross-lingual, and multilingual variants. With the exception of WSD, they are all defined as binary decision problems.

Our theoretical problem formulations correspond to well-studied semantic tasks. In practice, these tasks are rarely precisely defined, and instead depend on annotated datasets. For example, the definitions of lexical substitution differ between papers, and involve imprecise terms, such as "the overall meaning of the context" or "suitable substitute." The exact relationships between these tasks have not been rigorously demonstrated. Altogether, the recent literature suggests that a more detailed taxonomy is very much needed.

We start by formally defining the problems in terms of concepts and contexts, and proceed to determine their relative hardness by specifying reduction algorithms which produce a solution for one problem by applying an algorithm for another. In particular, we demonstrate that all problems in the taxonomy can be reduced to WSD, which confirms the principal role of this problem in lexical semantics. Furthermore, we show by mutual reductions that WSD and multilingual variants of WiC and LexSub are theoretically equivalent. Finally, we shed light on how they relate to lexical translation and wordnets.

The soundness of the problems in the taxonomy hinges on the consistency of judgments of sameness of word meaning. Hauer and Kondrak (2022) demonstrate the theoretical equivalence of the monolingual WiC and WSD via mutual reduction. We posit the following generalization of their sense-meaning hypothesis to multilingual concepts: *different word instances have the same* **meaning** *if and only if they express the same* **concept**. This empirically falsifiable proposition, which we refer to as the *concept-meaning hypothesis*, allows us to incorporate multilingual tasks, including lexical synonymy and substitution, into our theoretical framework.

In addition to showing that our theoretical propositions follow directly from our definitions and assumptions, we perform a series of experiments for the purpose of testing their empirical applicability and soundness. In particular, we test three problem reductions on standard benchmark datasets using independently developed systems based on pre-trained language models. Manual error analysis reveals no counter-examples to our concept-meaning hypothesis.

Our main contribution is a novel taxonomy of formally-defined problems, which establishes the reducibility or equivalence relations between the principal tasks in lexical semantics. In addition, we carry out a series of experiments that support the correctness of our theoretical findings.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9833–9844

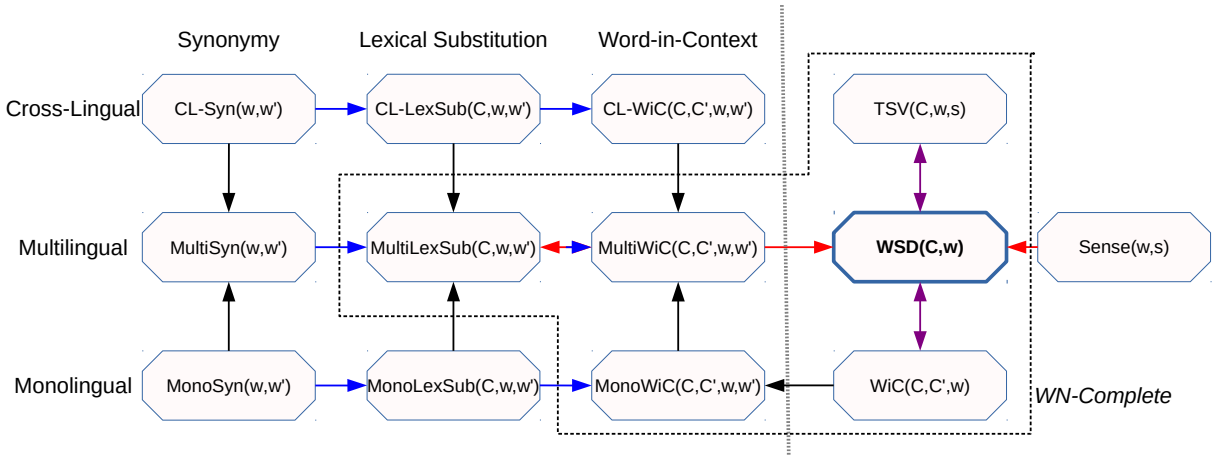

Figure 1: Taxonomy of problems in lexical semantics. Arrows indicate reducibility. The six wordnet-complete problems within the dotted area are equivalent, and all other problems in the taxonomy are reducible to them.

2 Theoretical Formalization

In this section, we formally define the problems in our proposed taxonomy, and discuss the relationship between these theoretical problems and the computational tasks addressed in prior work.

2.1 Words

All semantic problems in Figure 1 take at least one word as a parameter. In our definitions, a *word* is not necessarily an orthographic word, but rather a triple consisting of a lemma, a part of speech, and a language. The problems are divided into three categories based solely on the language of the words (rather than contexts): *monolingual* (same language), *cross-lingual* (different languages), and *multilingual* (same or different languages). Thus, a multilingual problem can be seen as the union of the corresponding monolingual and cross-lingual problems. While this categorization theoretically admits "monolingual" problem instances consisting of a word in one language and a context in a different language, such instances are rare in practice.

2.2 Contexts and Concepts

Alternatively, we can categorize semantic problems according to the number of contexts which must be considered in each instance: zero, one, or two, respectively, in the leftmost three columns of Figure 1. Contexts are denoted by the variable names starting with C. We broadly define a *context* as a discourse (not necessarily a sentence) with a *focus*, which is a word or sequence of words that express a specific concept. Contexts that consist of the same discourse but differ in focus are considered distinct. The expression "a word expresses a concept given a context" signifies that the word can be used to refer to the concept that corresponds to the focus of that context. Note that the word itself is not required to occur in the context, or even match the language of the context.

For example, consider the context "bats live in caves" which disambiguates the word *bat* to its animal sense. The underlined word represents the focus of the context, which can be expressed by the words *bat* or its synonym *chiropteran*. The languages of the word and the context need not be the same. For example, the Spanish context "un murciélago entro en mi casa" disambiguates the English word *bat* as an animal rather than an instrument.

A lexical concept, or simply *concept*, refers to a discrete word meaning. A *concept gloss*, such as "flying nocturnal rodent," is a special type of a context, in which the entire definition is the focus, and which uniquely determines the concept. We assume that the concept gloss C_s which defines the meaning of the concept s can be expressed in any language.

We assume the availability of complete sets of words (i.e., *lexicons*) and lexical concepts. The methods for creating such resources are beyond the scope of this paper.

2.3 Monolexical Problems

We first define three problems that take a single word argument. We refer to these theoretical problems by the same acronyms as their corresponding computational tasks: WSD, TSV, and WiC.

Word sense disambiguation (WSD) is the task

of classifying a word in context according to its sense, given an inventory of possible senses for each word. For each word, there is a one-to-one mapping between its senses and the concepts that it can express. We can therefore define the WSD problem more generally, to return a concept rather than a sense. This avoids the need for a predefined sense inventory for each word.

WSD(C, w) := "the concept which is expressed by the word w given the context C"

Note that this formulation does not require the word to occur in the context. By convention, the return value of the WSD predicate is *undefined* if the word is not meaningful given the context; for example, the English word *metre* does not express any concept given the Italian context "la metro di Roma è efficiente." In contrast, any binary predicate is assumed to return FALSE in such cases.

Target sense verification (Breit et al., 2021, TSV) is the binary classification task of deciding whether a given word in a given context expresses a given sense. As with WSD, we define the TSV problem on concepts rather than senses. We assume that the concept s is represented by its gloss C_s.

TSV(C, w, s) := "the word w expresses the concept s given the context C"

The TSV problem can be viewed as a binary analogue of the WSD problem, such that the following equivalence holds:

$$\text{TSV}(C, w, s) \Leftrightarrow \text{WSD}(C, w) = s$$

The word-in-context task (WiC) is a binary classification task proposed by Pilehvar and Camacho-Collados (2019): given a pair of sentences, decide whether or not a word has the same meaning in both sentences. We define the corresponding WiC problem using concepts, on the basis of the concept-meaning hypothesis.

WiC(C_x, C_y, w) := "the word w expresses the same concept given the contexts C_x and C_y"

Hauer and Kondrak (2022) demonstrate the equivalence of WiC, TSV, and WSD by pairwise reductions, which are denoted by purple arrows in Figure 1. In particular, the following formula specifies the reduction of WiC to WSD:

$$\text{WiC}(C_x, C_y, w) \Leftrightarrow \text{WSD}(C_x, w) = \text{WSD}(C_y, w)$$

2.4 Word-in-Context Problems

We now introduce a set of binary predicates which include WiC and its variants. We start with the most general problem of the set, MultiWiC, and then define MonoWiC, and CL-WiC as its special cases, in which the two words w_x and w_y are constrained to be either in the same or different languages, respectively.

MultiWiC(C_x, C_y, w_x, w_y) := "the words w_x and w_y express the same concept given the contexts C_x and C_y, respectively"

The WiC problem defined in Section 2.3 is a special case of MonoWiC, in which $w_x = w_y$.

MonoWiC(C_x, C_y, w_x, w_y):= "the words w_x and w_y from the same language express the same concept given the contexts C_x and C_y, respectively"

Martelli et al. (2021) extend the WiC task to include cross-lingual instances, which consist of a pair of contexts in different languages, in which the two focus words have the same meaning.[1] Our definition of the corresponding theoretical problem is similar:

CL-WiC(C_x, C_y, w_x, w_y): "the words w_x and w_y from different languages express the same concept given the contexts C_x and C_y, respectively"

Clearly, any instance of MultiWiC is either an instance of MonoWiC or CL-WiC.

2.5 Lexical Substitution Problems

The next set of problems each involve a pair of words in a single context. These problems formalize the semantic task of lexical substitution (McCarthy and Navigli, 2007), and its different variants and settings, such as cross-lingual substitution (Mihalcea et al., 2010). Our definitions are more precise than conventional ones, as we define substitutes on the basis of identity of expressed concepts. By virtue of our concept-meaning hypothesis, the definitions formalize the notions of "meaning-preserving substitutions" and "correct translations" present in previous work. However, they are restricted to lexical substitutions, excluding compositional compounds and phrases.

MonoLexSub(C, w_x, w_y) := "the words w_x and w_y from the same language express the same concept given the context C"

In other words, w_x and w_y are mutually substitutable given the context C. For example, MonoLexSub returns TRUE given C = "the gist of the prosecutor's argument", w_x = *core*, and w_y = *heart*.

[1] An instance was annotated as positive "if and only if the two target word occurrences were used with exactly the same **meaning** or, in other words, if, using a dictionary, the **definition** of the two target words was the same" (Martelli et al., 2021).

The CL-LexSub problem is a cross-lingual counterpart of MonoLexSub. The definition of CL-LexSub is the same as that of MonoLexSub, except that the two words are required to be in different languages. For example, MonoLexSub("she batted the ball", *bat*, *murciélago*) returns FALSE.

CL-LexSub(C, w_x, w_y) := "the words w_x and w_y from different languages express the same concept given the context C"

Finally, we define a multilingual lexical substitution problem which generalizes MonoLexSub and CL-LexSub by removing their respective language constraints:

MultiLexSub(C, w_x, w_y) := "the words w_x and w_y from any language(s) express the same concept given the context C"

While the goal of many conventional lexical substitution datasets is to produce sets of substitutes, these generative problems are reducible to the corresponding binary classification problems by iterating over the set of substitution candidates. More formally, the problem of generating lexical substitutes reduces to MultiLexSub by returning the set: $\{w \mid \text{MultiLexSub}(C, w_x, w)\}$.

2.6 Word Synonymy Problems

Our final set of semantic problems are defined on a pair of word lemmas, without any context parameters.

The MonoSyn predicate formalizes the relation of word synonymy in the monolingual setting. Given two words in the same language, it returns TRUE iff they are mutually substitutable in some context.

MonoSyn(w_x, w_y) := "the words w_x and w_y from the same language express the same concept in some context"

For example, MonoSyn(*core*, *heart*) is TRUE because there exist a contexts in which the two words express the same concept (c.f., Section 2.5). The MonoSyn problem formalizes the linguistic Substitution Test for synonymy: *w_x and w_y are synonyms if the meaning of a sentence that contains w_x does not change when w_y is substituted for w_x* (Murphy and Koskela, 2010).

We define the cross-lingual synonymy problem CL-Syn in a similar manner. The only difference with MonoSyn is that the two words are required to be from different languages.

CL-Syn(w_x, w_y) := "the words w_x and w_y from different languages express the same concept in some context"

The CL-Syn predicate corresponds to the relation of translational equivalence between words. Two words in different languages are translationally equivalent if there exists a context in which they are literal translations. For example, CL-Syn(*heart/EN*, *cœur/FR*) is TRUE because the two words are mutual translations given the context "the heart of the matter."

As with the other problem families, we unify MonoSyn and CL-Syn into a single predicate MultiSyn, which places no constraints on the language of the given words.

MultiSyn(w_x, w_y) := "the words w_x and w_y from any language(s) express the same concept in some context"

MultiSyn is not only a generalization but also the union of the relations of synonymy and translational equivalence, which are represented by MonoLexSub and CL-LexSub, as postulated by Hauer and Kondrak (2020).

3 Problem Reductions

Given an algorithm for a problem **Q**, a **P**-to-**Q** reduction solves an instance of a problem **P** by combining the solutions of one or more instances of **Q**. The reducibility of **P** to **Q** is denoted $\mathbf{P} \leq \mathbf{Q}$. Mutual reductions of two problems to one another, i.e. $\mathbf{P} \leq \mathbf{Q}$ and $\mathbf{Q} \leq \mathbf{P}$, demonstrate their equivalence.

In this section, we present several problem reductions, which constitute the main contribution of this paper. The reductions are shown in Figure 1 by the directed arrows from **P** to **Q**. The black arrows denote the special cases, which immediately reduce to the more general problems. Taken together, the reductions establish the equivalence of six problems: WSD, TSV, WiC, MonoWiC, MultiWiC, and MultiLexSub. A method which solves any of these problems can be used to construct methods which solve the other problems by applying a sequence of reductions. As well, a method for one of those six problems can be used to solve any of the other problems in Figure 1, again via reductions.

3.1 *Syn $\leq$ *LexSub $\leq$ *WiC

We first present a set of six reductions, which are denoted by blue arrows in Figure 1. Each of the corresponding nine problems involves comparing the meanings of a pair of words, given some contexts.

The three lexical substitution problems defined in Section 2.5 can be viewed as special cases of the

corresponding word-in-context problems, in which both contexts are identical. Succinctly:

$$\text{*LexSub}(C, w_x, w_y) \Leftrightarrow \text{*WiC}(C, C, w_x, w_y)$$

The asterisk in these and the following reductions can be replaced on both sides by "Mono", "CL-", or "Multi". To reiterate, a cross-lingual problem explicitly assumes that the input words are in different languages, while a multilingual problem can accept inputs in the same or different languages.

The three word synonymy problems defined in Section 2.6 are reducible to the corresponding lexical substitution problems. In particular, to reduce MultiSyn to MultiLexSub, we search for a concept gloss C_s in which both words express the same concept. Succinctly:

$$\text{*Syn}(w_x, w_y) \Leftrightarrow \exists s : \text{*LexSub}(C_s, w_x, w_y)$$

The correctness of these six reductions follows from the fact that the (infinite) set of all contexts is partitioned into equivalence classes, each of which corresponds to a single concept.

3.2 Reductions to WSD

The reductions in the preceding section demonstrates that all theoretical problems defined in Section 2 can be reduced to MultiWiC. We next demonstrate that all those problems, including MultiWiC itself, can also be reduced to WSD. Thus, an algorithm that solves WSD would be sufficient to solve all other problems. For clarity, the nine reductions in this section are not shown explicitly in Figure 1, with the exception of the crucial MultiWiC-to-WSD reduction, denoted by a red arrow.

Given a method for solving WSD, we can solve any *WiC instance by checking whether the concepts expressed by the two words in the corresponding contexts are the same. This set of reductions generalize the WiC-to-WSD reduction (Section 2.3) to MonoWiC, CL-WiC, and MultiWiC:

$$\text{*WiC}(C_x, C_y, w_x, w_y) \Leftrightarrow \text{WSD}(C_x, w_x) = \text{WSD}(C_y, w_y)$$

Similarly, to solve any *LexSub instance, it is sufficient to check the identity of the concepts expressed by the two words in the given context:

$$\text{*LexSub}(C, w_x, w_y) \Leftrightarrow \text{WSD}(C, w_x) = \text{WSD}(C, w_y)$$

Finally, the word synonymy problems can be solved by searching for a concept which can be expressed by both words.

$$\text{*Syn}(w_x, w_y) \Leftrightarrow \exists s : \text{WSD}(C_s, w_x) = \text{WSD}(C_s, w_y)$$

The correctness of the reductions in this section follows directly from the concept-meaning hypothesis which underlies our theory.

3.3 MultiWiC $\leq$ MultiLexSub

We close this section by demonstrating that MultiWiC is reducible to MultiLexSub, which is denoted by a red arrow in Figure 1. This reduction, along with the reverse reduction presented in Section 3.1, establishes the equivalence between the two problems. Formally:

$$\begin{aligned}&\text{MultiWiC}(C_x, C_y, w_x, w_y) \Leftrightarrow \\ &\text{MultiLexSub}(C_x, w_x, w_y) \wedge \text{MultiLexSub}(C_y, w_y, w_x) \wedge \\ &\forall w : \text{MultiLexSub}(C_x, w_x, w) \Leftrightarrow \text{MultiLexSub}(C_y, w_y, w)\end{aligned}$$

The first two terms on the right-hand side of the reduction formula test whether the two words are mutually substitutable in their respective contexts. The universal quantifier ensures that every substitute in one of the contexts is also an appropriate substitute in the other context, and vice versa.

The correctness of this reduction hinges on the assumption that there are no universal colexifications (Bao et al., 2021), which states that *for any pair of concepts, there exists some language which lexifies but does not colexify them*. In other words, there exists a language in which no word can express both concepts. Therefore, if the sets of contextual synonyms of w_x in C_x and w_y in C_y are identical, the concept expressed by the two word tokens must be the same.

In theory, the universal quantifier in the reduction formula is defined over all words in all languages. In practice, only the synonyms and translations of the two words need to be checked, and a smaller set of diverse languages may be sufficient to obtain good accuracy.

4 Relationship to Synsets

A wordnet is a theoretical construct which is composed of synonym sets, or *synsets*, such that each synset corresponds to a unique concept, and each sense of a given word corresponds to a different synset. Actual wordnets, such as Princeton WordNet (Miller, 1995), are considered to be imperfect implementations of the theoretical construct.

We define the following monolexical problem, which decides whether a given word can express a given concept:

Sense(w, s) := "the word w expresses the concept s in some context"

An algorithm for the Sense problem could be used to decide whether a given word belongs to the synset that corresponds to a given concept.

4.1 *Syn $\leq$ Sense $\leq$ WSD

The word synonymy problems defined in Section 2.6 are reducible to the Sense problem. Two words are synonyms if they both express the same concept in some context. In particular, to reduce MultiSyn to Sense, we search for a concept which can be expressed by both words.

$$\text{MultiSyn}(w_x, w_y) \Leftrightarrow \exists s : \text{Sense}(w_x, s) \wedge \text{Sense}(w_y, s)$$

A monolingual wordnet can be converted into a thesaurus, in which the entry for a given word consists of all of its synonyms. A bilingual wordnet can be converted into a translation dictionary, in which the entry for a given word consists of all its cross-lingual synonyms possibly grouped by sense, and accompanied by glosses.

Given a method for solving WSD, we can solve a Sense instance by checking whether the word expresses the concept given the context of its gloss. Formally:

$$\text{Sense}(w, s) \Leftrightarrow \text{WSD}(C_s, w) = s$$

The correctness of this reduction follows from the assumption that a concept gloss uniquely determines the concept. Under our definitions, given a concept gloss, the WSD predicate can only return the corresponding concept, and does so if and only if the given word can express that concept; otherwise the return value is undefined.

The reducibility of Sense to WSD implies that implementing the WSD predicate as it is defined in Section 2.3 would make it possible to construct synsets from nothing more than a list of concept glosses, as well as correct and expand existing wordnets to new domains and languages. In fact, any of the set of six WSD-equivalent problems (Figure 1) could be used for these tasks; we therefore refer to them as wordnet-complete (WN-complete).

4.2 Substitution Lemma

The final proposition formalizes the relationship between synsets, senses, and lexical translations. It follows directly from the previously stated definitions, reductions, and assumptions.

$$\text{MultiLexSub}(C_x, w_x, w_y) \Leftrightarrow \text{Sense}(w_y, \text{WSD}(C_x, w_x))$$

The lemma provides a theoretical justification for methods that associate contextual lexical translations and synonyms with the synset identified by a WSD model. For example, BabelNet synsets are populated by translations of word instances that correspond to a given concept (Navigli and Ponzetto, 2010). Specifically, the existence of a translation pair (w_x, w_y) in a context C_x implies that w_y lexicalizes the concept expressed by w_x in C_x. Another example is the method of Luan et al. (2020), which leverages contextual translations to improve the accuracy of WSD.

5 Empirical Validation

In this section, we implement and test three principal reductions: MultiWiC to WiC, MultiWiC to WSD, and MultiLexSub to WSD. For each reduction, we reiterate its theoretical basis, describe our implementation, and discuss the results. We emphasize that the goal of our experiments is not challenging the state of the art, but rather empirically testing the reductions, and, by extension, the hypothesis they are based on. Since the resources used for the implementations are necessarily imperfect, and the systems are each designed and optimized for a different target task, the reductions are expected to produce much less accurate predictions on the existing benchmark datasets compared to state-of-the-art methods.

Our primary interest is in identifying any possible counter-examples to our concept-meaning hypothesis. However, it must be noted that the presence of a small number of such exceptions in the existing datasets does not invalidate the theory. On the other hand, the scarcity of counter-examples should not be interpreted as a *proof,* but rather as supporting evidence for the correctness of our theoretical claims.

5.1 Solving MultiWiC with WiC

We first empirically test the counter-intuitive proposition that a multilingual semantic task can be reduced to a set of monolingual instances. In particular, given a method for solving WiC, we can solve any MultiWiC instance by deciding whether there exists a concept such that both given words express the concept given their corresponding contexts and the concept gloss. Formally:

$$\text{MultiWiC}(C_x, C_y, w_x, w_y) \Leftrightarrow$$
$$\exists s : \text{WiC}(C_x, C_s, w_x) \wedge \text{WiC}(C_y, C_s, w_y)$$

The correctness of this reduction follows from the assumption that a concept gloss uniquely disambiguates every word that can express the concept.

5.1.1 Implementation of the Reduction

In practice, instead of checking all possible concepts, we limit our search to concepts that can be expressed by either of the two words. For each

such concept, we create two WiC instances, one in each language, using a gloss retrieved from a lexical resource, and translated, as needed, into the language of each instance. We then solve each of the created WiC instances using a model trained exclusively on WiC data in that language. The reduction returns TRUE iff both WiC instances are classified as positive.

We test the reduction on the English-French test set of the MCL-WiC shared task (Martelli et al., 2021), which contains 1000 MultiWiC instances. The dataset is agnostic toward WordNet sense distinctions and annotations. We train the English WiC model on the English training and development sets (8k and 1k instances, respectively), and the French WiC model on the French development set (1k instances). The latter set is quite small, but we are not aware of any larger dedicated French WiC training data.

We create each WiC instance by prepending the input word, followed by a separator token, to each input context, including concept glosses. We retrieve concept glosses from BabelNet (Navigli and Ponzetto, 2010), using the Python API.[2] While English lemmas are provided in the dataset, French lemmas are not. We therefore lemmatize French words using the SpaCy FR_CORE_NEWS_MD model. Since BabelNet does not contain French glosses for all concepts, we generate them by translating the first English gloss in BabelNet using the OPUS-MT-EN-FR model from Helsinki NLP.[3]

We train our English and French WiC models using LIORI (Davletov et al., 2021). All training was completed in under eight hours on two NVIDIA GeForce RTX 3090 GPUs. With the default hyper-parameter settings, the models obtain the accuracy of 87.0% and 73.3% on the English and French monolingual test sets, respectively. This is lower than the 91.1% and 86.4% results reported by Davletov et al. (2021). We attribute this to our use of smaller, purely monolingual training data, which is in line with our theoretical reduction. Based on these numbers, we estimate the probability of a pair of WiC instances being both correctly classified as $0.870 * 0.733 = 0.638$.

5.1.2 Results and Discussion

Our implementation correctly classifies 631 out of the 1000 instances in the test set. This is very close to the estimate computed in the previous section, which suggests that our reduction is approximately as reliable as our imperfect resources and systems allow.

We manually analyzed a random sample of 50 MultiWiC classification errors. For each of the 25 false negatives, LIORI returned FALSE for *all* sentence pairs in either English (12 instances), French (8 instances), or both languages (5 instances). Each instance could be explained by either a LIORI error, or a missing sense in BabelNet. For the 25 false positives, we identified one or more incorrect positive WiC classifications. The final false positive was caused by an incorrect tokenization of the target word in the MCL-WiC dataset: *disordered* instead of *mentally disordered*.

Since all errors can be attributed to the systems and resources, they constitute no evidence against the correctness of our reduction. On the other hand, these results support our theoretical finding that multilingual problems can be reduced to monolingual problems. This in turn supports our methodology of grounding lexical semantics in the expression of language-independent concepts.

5.2 Solving MultiWiC with WSD

In this section, we test our MultiWiC-to-WSD reduction. In doing so, we generalize the WiC-to-WSD reduction of Hauer and Kondrak (2022) to multiple words and languages. Given a MultiWiC instance, we apply a WSD system to each context-word pair, and classify it as positive iff both words are tagged with the same BabelNet synset:

$$\text{MultiWiC}(C, C', w, w') \Leftrightarrow \text{WSD}(C, w) = \text{WSD}(C', w')$$

5.2.1 Implementation of the Reduction

Our system of choice is AMuSE-WSD (Orlando et al., 2021). It provides pre-trained WSD models for a diverse set of languages, and handles all stages of the WSD pipeline, including tokenization, lemmatization, and part-of-speech tagging. We apply the provided AMUSE-LARGE-MULTILINGUAL-CPU model, with all other parameters left at their default values.

Following Hauer and Kondrak (2022), we estimate an upper-bound on the performance of our reduction, using analogous notation and formulae. For the expected accuracy of English and non-English WSD, we use the English-ALL and XL-WSD accuracy results reported by the AMuSE-WSD authors, 0.739 and 0.673. This estimation method also depends on the average number of

[2] https://babelnet.org/guide#python
[3] https://huggingface.co/Helsinki-NLP

senses per target word. Per the BabelNet API[4], an average MCL-WiC target word has 14 senses. The resulting overall accuracy estimate is 0.752, which is the average of 0.539 and 0.965 for the positive and negative MultiWiC instances, respectively.

5.2.2 Results and Discussion

The results on the MCL-WiC test sets range from 51.8% on English-Arabic to 55.1% on English-French. While the estimate in the previous section is substantially higher, it does not take into account tokenization errors and missing senses in BabelNet. On the English-French dataset, we found that false negatives outnumber false positives by a factor of six; the accuracy is 22.8% and 87.4% on the positive and negative MultiWiC instances, respectively.

For our manual analysis, we randomly selected 25 false positives and 25 false negatives produced by our implementation on the English-French test set. In 41 of the 50 cases, we determined the cause of the incorrect MultiWiC classification to be an incorrect sense returned by AMuSE-WSD for one or both target words. In addition, 7 of the 50 cases represent tokenization errors. One MultiWiC instance, which involves English *reflected* and French *consignée*, is most likely a MCL-WiC annotation error. The final error is attributable to a sense missing from BabelNet, which prevents AMuSE-WSD from considering it as a candidate. Specifically, it is the "administer" sense of the verb *dispense* (as in "dispense justice"), which can be found in the Merriam-Webster Online Dictionary.[5]

Since manual analysis yields no counterexamples to our theory, we interpret the results as empirical support for this reduction, and, by extension, our taxonomy of semantic tasks, and the hypothesis on which it is based.

5.3 Solving MultiLexSub with WSD

In the final experiment, we test the MultiLexSub-to-WSD reduction derived in Section 3.2:

$$\text{MultiLexSub}(C, w, w') \Leftrightarrow \text{WSD}(C, w) = \text{WSD}(C, w')$$

The overall method is similar to that of Guo and Diab (2010), but using our precise binary formulation of lexical substitution.

5.3.1 Implementation of the Reduction

We use the dataset from the SemEval 2010 shared task on cross-lingual lexical substitution (Mihalcea et al., 2010), which consists of a trial set of 300 instances, and a test set of 1000 instances. Each instance consists of an English sentence which includes a single target word and a list of Spanish gold substitutes provided by annotators.

Since our formulation of lexical substitution is binary rather than generative or ranking-based, we convert each of the SemEval instances into a pair of binary instances: one positive and one negative. For the positive instance, we take the first Spanish substitute, the one that was most frequently suggested by the annotators. For the negative instance, we randomly select a Spanish word from the set of all substitutes in the dataset for that English target word, provided that it is not among the gold substitutes for that specific instance. If there is no such substitute, we instead choose a random Spanish word from the dataset.

For each binary instance created in this way, we create two WSD instances using a simple template: *'w' as in 'C'*, where w is the target word, and C is the context. We obtain the context for the Spanish word by translating the English context via Helsinki NLP's OPUS-MT-EN-ES model. We return a positive MultiLexSub classification iff AMuSE-WSD assigns the same BabelNet synset ID to both English and Spanish target words.

Our procedure for estimating the expected accuracy of our reduction is the same as in Section 5.2.1. The only difference is the average number of senses per word, which in this case is 23, yielding an estimated accuracy of 75.8%.

5.3.2 Results and Discussion

The binary classification accuracy of our implementation on 2000 MultiLexSub instances created from the SemEval test set is 63.2%, which is substantially below the estimate in the previous section. This can be partially explained by a relatively high number of tokenization errors in the test set. We again observe a strong bias toward negative classification: the results on the positive and negative instances are 27.1% and 99.3% accuracy, respectively. Because of this, we selected only positive instances for our error analysis.

We manually analyzed a sample of 50 randomly-selected false negatives from the test set. In 44 of the 50 cases, the cause of the misclassification was an AMuSE-WSD error (on English in 30 cases, on Spanish in 14). Some of those errors may be caused by an imperfect translation of the English context, or a missing BabelNet sense of the Spanish

[4] https://babelnet.org/guide#python

[5] https://www.merriam-webster.com/dictionary/dispense

gold substitute. In 5 cases, the English input was incorrectly tokenized; for example, the compound noun *key ring* was split into two word tokens, with one instance having *ring* as its focus. The final case likely involves an annotation error in the SemEval dataset: *campo* as a translation of *field* given the context of "effective law enforcement in the field."

We conclude that all incorrect classifications can be attributed to a resource or system used by our implementation, and thus none of them represents a counter-example to our hypothesis.

6 Conclusion

Starting from basic assumptions about the expression of concepts by words in context, we have developed consistent formulations of thirteen different problems in lexical semantics. We have shown that a "wordnet-complete" subset of these tasks can each be used to solve any of the others via reduction. These problems can be used to construct, correct, or expand multilingual synonym sets, the building blocks of important linguistic resources such as WordNet and BabelNet. We believe that this work will lead to a greater understanding of lexical semantics and its underlying linguistic phenomena, as well as new applications and better interpretation of empirical results. Based on our theory, we intend to develop methods for constructing fully explainable and interpretable linguistic resources.

Limitations

While we do include multilingual datasets in our experiments, our error analysis is limited to languages of the Indo-European family, specifically English, French, and Spanish, as these are the languages covered by our datasets which we can confidently analyze. In addition, it is possible to question some of the assumptions made in our theory, which should be kept in mind when considering our work. For example, we assume that, for each content word token in a discourse, there exists a single concept which that word is intended by the sender to express, regardless of whether it appears unambiguous to the receiver. However, unlike in mathematics, theoretical assumptions may not always hold in practice; for example, puns often exploit multiple meanings of a word for humorous effect. While such cases are not frequently considered in lexical semantics, we can expect exceptions to almost any assumption or conclusion regarding human languages.

Acknowledgements

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

References

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference (GWC2021)*, pages 1–7.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645.

Adis Davletov, Nikolay Arefyev, Denis Gordeev, and Alexey Rey. 2021. LIORI at SemEval-2021 task 2: Span prediction and binary classification approaches to word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 780–786.

Weiwei Guo and Mona Diab. 2010. COLEPL and COLSLM: An unsupervised WSD approach to multilingual lexical substitution, tasks 2 and 3 SemEval 2010. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 129–133, Uppsala, Sweden. Association for Computational Linguistics.

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

Bradley Hauer and Grzegorz Kondrak. 2022. WiC = TSV = WSD: On the equivalence of three semantic tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2478–2486, Seattle, United States. Association for Computational Linguistics.

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.

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 Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

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.

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.

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

M. Lynne Murphy and Anu Koskela. 2010. *Key terms in semantics*. London: Continuum.

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.

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

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Discussed, with citations, throughout Section 5.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All resources we used are freely available for research purposes.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Discussed, with citations, throughout Section 5.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Discussed, with citations, throughout Section 5.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Discussed, with citations, throughout Section 5.

C ☑ Did you run computational experiments?

Left blank.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The models we used were previously presented in other papers. We have cited those papers.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.