Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 42–51
Marseille, 20 June 2022

A Category Theory Framework for Sense Systems

David Strohmaier*, Gladys Tyen*
University of Cambridge
{david.strohmaier, gladys.tyen}@cl.cam.ac.uk

Abstract

Sense repositories are a key component of many NLP applications that require the identification of word senses, a task known as word sense disambiguation. WordNet synsets form the most prominent repository, but many others exist and over the years these repositories have been mapped to each other. However, there have been no attempts (until now) to provide any theoretical grounding for such mappings, causing inconsistencies and unintuitive results. The present paper draws on category theory to formalise assumptions about mapped repositories that are often left implicit, providing formal grounding for this type of language resource. We introduce notation to represent the mappings and repositories as a category, which we call a *sense system*; and we propose and motivate four basic and two guiding criteria for such sense systems.

Keywords: Sense Repositories, Word Sense Disambiguation, Category Theory

1. Introduction

Sense repositories are a key language resource for word sense disambiguation (WSD), semantic inference, specifying lexical relations, and other downstream tasks like question answering. For these purposes, researchers have created many sense repositories with varying levels of granularity, along with mappings between them. In particular, the popular WordNet synsets (Miller et al., 1990; Fellbaum, 1998) have been mapped to many coarser-grained repositories.

The value of systematically mapped repositories has been repeatedly shown (Navigli, 2006; Palmer et al., 2007). However, the particular characteristics of the mappings produced are often the byproduct of practical or engineering decisions, instead of being motivated by theoretical considerations. For example, clustered senses are restricted to one cluster per sense, whereas senses that are mapped to domain labels do not have this restriction and are often associated with multiple labels. Additionally, the lack of constraints on mappings often results in problems during implementation. For example, converting sense labels in a corpus from one type to another (e.g. synsets to domain labels) is not always consistent, because sometimes there are several correct labels.

The present paper provides the theoretical grounding to allow for more systematic understanding of mappings and how they might assist researchers in solving tasks such as WSD. As far as we know, no such theory has been proposed before. Our contributions are twofold:

1. Drawing from category theory, we formalise mapped sense repositories as a category which we call a *sense system*; and

2. Using category theoretic notation, we propose and formally describe criteria for such a sense system.

We hope that future researchers building or adapting sense repositories and mappings will find it useful to consider how their new language resource fits into our framework, and adjust their methodology accordingly.

In the following sections, we first discuss the existing literature on sense repositories and mappings between them. We then introduce sense systems and present the surrounding category-theoretic notation. With these foundations in place, we propose and provide motivation for **basic** and **guiding** criteria for such sense systems.

2. Previous work

2.1. Word Sense Disambiguation

As suggested, word sense disambiguation (WSD), i.e. picking the correct sense of a word in a context, is one of the most prominent uses of sense repositories. Typically, a WSD classifier[1] selects from a pre-determined and enumerative repository of candidate senses (Navigli, 2009).

Different NLP techniques for WSD have been developed over the years, including approaches based on lexical similarity, graphs, and supervised learning. Lesk (1986) offers an influential lexical similarity approach, which uses a) the overlap between context of the word to be disambiguated, and b) the dictionary entry of candidate senses, in order to select a sense. Graph-based approaches make use of the graph structure of some sense repositories such as WordNet and BabelNet to select senses (Moro et al., 2014).

In recent years, machine learning has become the dominant approach. WSD is treated as a supervised classification task, where a trained model selects from a pre-determined list of senses. Earlier methods depend on extracting feature vectors (Zhong and Ng, 2010; Mihalcea and Faruque, 2004), while later methods make

* Both authors contributed equally.

[1] We refrain from using the term *word sense disambiguation system* in this paper to avoid any confusion with *sense systems*.

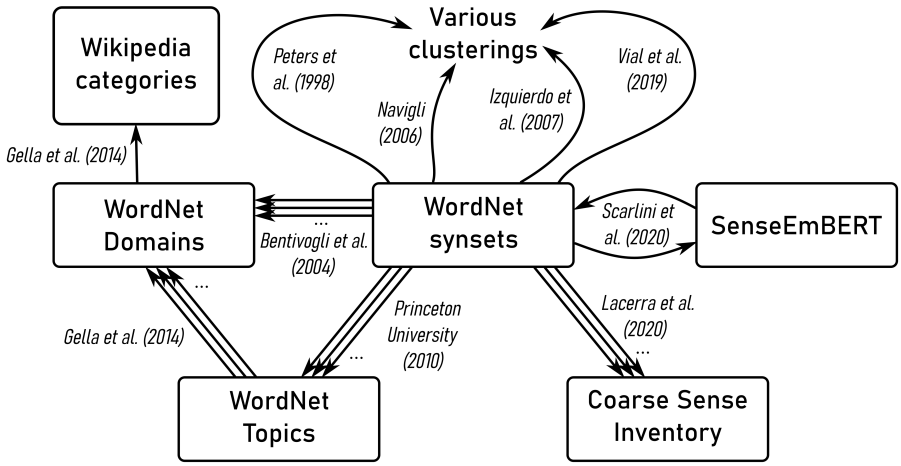

Figure 1: Graph showing mappings between select repositories. ... denotes further possible mappings.

use of word embeddings (Mikolov et al., 2013) and shifted towards neural approaches (Kågebäck and Salomonsson, 2016; Vial et al., 2019; Wiedemann et al., 2019), giving rise to some of the best performing models in WSD. Word embeddings have also been used as features for non-neural machine learning methods (Iacobacci et al., 2016), as well as more traditional lexical similarity approaches (Oele and Noord, 2017).

2.2. Sense representations

Sense repositories are sets of word senses, i.e. representations of lexical meaning. Existing sense repositories range widely in terms of how senses are represented and how fine-grained they are. Sense representations can be roughly divided into 4 types: dictionary definitions, clusters, domain labels, and embedding vectors.

1. **Dictionary definitions** typically consist of a piece of text describing the sense in question. A dictionary is an enumerative listing of such senses, though in practice such a list is unlikely to be exhaustive. WordNet (Miller et al., 1990; Fellbaum, 1998), one of the most widely used sense repository in WSD, is a prime example of a dictionary-like repository: it consists of gloss definitions, each of which is linked to a set of corresponding synonymous words, called a synset.

   Outside of WordNet, there are many repositories where senses are represented as definitions. For example, BabelNet (Navigli and Ponzetto, 2012), MultiWordNet (Pianta et al., 2002), and EuroWordNet (Vossen, 1998) are three multilingual repositories similar to WordNet; and many conventional dictionaries like the *Longman Dictionary of Contemporary English* (LDOCE) and the *Oxford Dictionary of English* (ODE) have also been used for WSD. Due to the popularity of WordNet, much of the WSD work cited in this paper pertains to mappings from WordNet, but many of the techniques can be applied to other repositories as well.

2. **Clusters of senses** are obtained by grouping fine-grained senses by various metrics, which typically approximate semantic similarity. For example, the semantic relations encoded in WordNet have been used to cluster WordNet synsets (Peters et al., 1998; Vial et al., 2019; Izquierdo et al., 2007); similarly, Dolan (1994) clustered definitions from the LDOCE according to semantic information extracted from the dictionary; Agirre and Lacalle (2003), working on clustering WordNet synsets, investigated 4 different sources of information to measure similarity: topic signatures, confusion matrices, translation equivalences, and the context of occurrence.

   Senses within a cluster can be represented as dictionary definitions, embedding vectors, or otherwise — crucially, there is no unified way of determining its semantic content, as it often depends on the clustering technique. For example, clusters that are formed from hypernym/hyponym relations have explicit, shared semantic content, because each cluster member is a hyponym of the highest level hypernym. In other cases, such as WordNet synsets clustered according to confusion matrices, there may not be any semantic content explicitly associated with each cluster.

3. **Domain labels** are very coarse-grained senses represented by a word or short phrase that denotes a topic domain, such as *biology*, *economics*, etc. Domain label repositories aim to cover the largest semantic space with the fewest possible domain labels (Lacerra et al., 2020; Izquierdo et al., 2007).

   Mappings to domain labels can be determined manually, automatically, or both. For example, Magnini and Cavaglia (2000) began with a small set of manual annotations, then extended

them automatically based on a semantic hierarchy; Camacho-Collados and Navigli (2017) produced their mappings according to similarity metrics and other heuristics, then evaluated a subset according to manual annotations. Many dictionary repositories like WordNet and the LDOCE also comes with manually annotated domain labels.

Unlike clusters, there is no way to ensure that all fine-grained senses can be mapped to a substantive domain, so a miscellaneous or "catch-all" label is sometimes used for uncategorised senses. For example, the WordNet Domains Hierarchy (Bentivogli et al., 2004) contains the label "factotum" for when no better label is available. Additionally, it is possible for fine-grained senses to be mapped to multiple domain labels.

4. **Embedding vectors** represent senses as a dense vector. Early word embedding techniques like Word2Vec (Mikolov et al., 2013) produce one embedding per word type, but later techniques such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) can be used to produce contextualised embeddings, which are effectively very fine-grained senses. Scarlini et al. (2020a; Scarlini et al. (2020b) have also created embeddings for WordNet synsets.

2.3. Mapping sense repositories

Most work on mapping sense repositories is motivated by a common concern: that WordNet synsets are too fine-grained to achieve reasonable results on the WSD task (Ide and Wilks, 2007; Lacerra et al., 2020). Some researchers advocate for multiple levels of grain, so that downstream applications are free to select the level as appropriate. For example, Palmer et al. (2004) employ WordNet synsets, synset groupings, and framesets as three repositories at different levels of grain. It has been argued that there is no single correct repository of senses that is independent of the use case (Kilgarriff, 2003).

It has been established that using multiple mapped repositories can improve the performance on the WSD task, demonstrating the practical value of mappings. Navigli (2006) clustered WordNet synsets based on partial mappings to the *Oxford Dictionary of English*, and showed that this mapping-based clustering improved the performance on the WSD task. Similarly, Palmer et al. (2007) showed that the possibility of backing off to coarse-grained sense groups improves WSD, further supporting the usefulness of mapping sense repositories of different grain.

None of this work, however, provides general theoretical grounding and restrictions for the mappings between multiple sense repositories. Formal features such as the transitivity of mappings are more often the result of practical exigencies and methodological choices rather than theoretical motivations. For example, some WordNet synsets were mapped to the Coarse Sense Inventory (CSI) indirectly via BabelDomains (Lacerra et al., 2020), suggesting that sense mappings are transitive. The present paper will make such implicit assumptions explicit using category theory.

3. Formal notation for a sense system

We introduce the term *sense system* to denote an interconnected system of sense repositories and mappings. We represent a sense system as a small category S, where the object set of S, denoted by $\mathbf{Ob}(S)$, is a set of sense repositories; and the homomorphism set or hom-set of S, denoted by $\mathbf{Hom}(S)$, is a set of mappings between these repositories. The set of mappings from repository R to repository R' in S is denoted by the hom-set $\mathbf{Hom}_S(R, R')$. The general hom-set $\mathbf{Hom}(S)$ is the union of all these repository-specific hom-sets.

Note that each R in $\mathbf{Ob}(S)$ only contains senses – other information such as word type exists separately (see Section 4.1.2) and we make no assumptions about the form or content of the senses themselves. Our sense system representation will be applicable regardless of whether the senses are dictionary definitions, embeddings, domain labels, or otherwise.

As a category, S has the following two properties:

1. **$\mathbf{Hom}(S)$ is closed under function composition.** If, in $\mathbf{Hom}(S)$, R is mapped to R' and R' is mapped to R'', then there must be some composite mapping that maps R to R'' in $\mathbf{Hom}(S)$.

2. **Each repository in $\mathbf{Ob}(S)$ has an identity function id in $\mathbf{Hom}(R, R)$ mapping R to itself.**

Both of these properties are trivially fulfilled by the common understanding of sense mappings.

We conceptualise each mapping as a way of converting a label from one repository to another label from another repository. For example, if WordNet synsets are mapped to WordNet Domains, one could take a corpus like SemCor (Landes et al., 1998), which is labelled with WordNet synsets, and convert the synset labels to Domain labels.

Since there can be multiple ways of converting, in principle multiple mappings from one repository to another can coexist. For example, the WordNet 2.0 synset for *amethyst* is linked to three WordNet Domain labels, as seen in Figure 2. When encountering the word *amethyst* in SemCor, one could select a label randomly, or according to some arbitrary order, or by frequency, etc. Each of these methods would correspond to a different mapping between the two repositories.

Mappings in $\mathbf{Hom}(S)$ have the following properties:

1. **Mappings are unidirectional.** A mapping from R to R' does not entail a mapping from R' to R.

   While this property is often assumed, it is not always made explicit. For example, WordNet

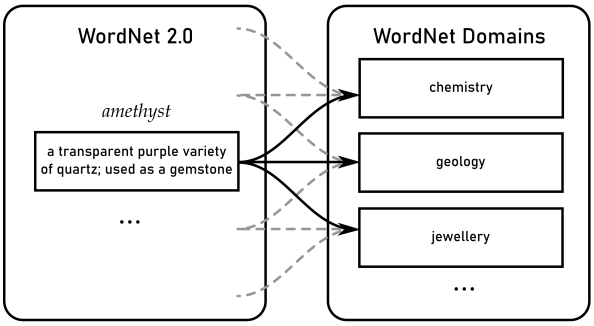

Figure 2: Example mapping from the WordNet 2.0 synset for *amethyst* to WordNet Domains.

synsets are often mapped to domain labels or clusters that are coarser-grained, making it impossible to reverse the mapping.[2] Therefore, repositories are typically mapped from finer-grained ones to coarser-grained ones, not vice versa. Bidirectional mappings would only be possible between repositories that are of equal grain and mapped one-to-one to each other, e.g. when embeddings are created specifically for WordNet synsets (Scarlini et al., 2020a).

2. **Mappings are not multivalued.** That is, each mapping in $\mathbf{Hom}_S(R, R')$ maps each sense in R to at most one sense in R', though multiple senses in R can be mapped to the same sense in R'.

   This is consistent with the idea that mappings represent a way of converting labels (as suggested above), because each conversion method takes one input and gives only one output.

3. **Mappings are total functions.** A mapping from R to R' ensures that all senses in R are mapped to at least one sense in R'.

   In practice, there are some cases where mappings are not total. For example, Navigli (2006) partially mapped WordNet synsets to definitions in the Oxford Dictionary of English, leaving synsets that are not mapped to any ODE senses. There may also be repositories that were built for a reduced vocabulary, such as dictionaries for learners, or repositories that only contain certain types of words, such as English verbs (Green et al., 2001).

   For the purposes of this theory, we follow Navigli (2006), Navigli and Ponzetto (2012), etc. and use ϵ as a null value, so senses that are not mapped to anything are instead mapped to ϵ.

The category theoretic properties described in this section will be assumed throughout this paper. Formalising a sense system as a category posits very minimal assumptions about sense repositories and their mappings, and should therefore be applicable to most existing sense systems.

However, such a flexible representation of sense systems is not very informative. Previous work on mapping repositories often impose further assumptions, resulting in sense systems that are more useful and informative. In the following sections, we formally describe these assumptions and formulate them as **basic** and **guiding** criteria for sense systems.

4. Basic criteria for sense systems

In this section, we formalise and motivate 4 **basic** criteria for sense systems. These criteria capture linguistic intuitions that are often implicitly assumed, while simultaneously accounting for downstream application concerns.

1. **Correctness preservation: Mappings should preserve the correctness of sense labels in all contexts.**

   Intuitively, if the correct sense for a word token is mapped to another sense, this sense should also be correct. To formalise this criterion, we postulate the existence of a WSD oracle Ω, which evaluates to 0 or 1 depending on whether a given word token in a usage context has a given sense. Note that Ω makes no assumption about the number of correct senses.

   We formalise the preservation of correctness as follows:

$$\begin{aligned} &\forall R, R' \in \mathbf{Ob}(S) \\ &\quad \forall m \in \mathbf{Hom}_S(R, R') \\ &\quad\quad \forall s \in R \\ &\quad\quad\quad \forall t \in T \\ &\Omega(t, s) = 1 \;\Rightarrow\; \Omega(t, m(s)) = 1 \end{aligned} \tag{1}$$

   where t denotes any given word **t**oken from the set of tokens T covered by both R and R'.

2. **Candidacy preservation: Mappings should preserve the lexical candidacy of sense labels.**

[2] One notable exception to this is the sense compression technique developed by Vial et al. (2019), which allows for mappings from coarse to fine senses in virtue of the way they were produced.

To introduce the concept of candidacy, we distinguish word types from word tokens: word tokens are words in a usage context; word types, also known as a lemma, refer to the abstract notion of a word, and is independent of morphological variants.

We postulate that word types exist separately for each repository R as the set W_R, which are mapped to senses in R like in a dictionary, i.e. each word type is associated with a set of candidate senses. We formalise this **d**ictionary function as $d_R : W_R \rightarrow \mathcal{P}(R)$, where $\mathcal{P}(R)$ denotes the power set of R.

For a sense s in R to be a candidate for a word type w, the dictionary function d_R must map w to a set that contains s. For example, in WordNet 3.1, the word *manuscript* is mapped to the set of two synsets: "the form of a literary work submitted for publication", and "handwritten book or document". Both of these senses are candidates of *manuscript*.

Having introduced the dictionary function, candidacy preservation can then be formulated as follows: if a sense s that is a candidate for a word type w is mapped to another sense, that sense must also be a candidate for w. Formally,

$$\begin{aligned} &\forall R, R' \in \mathbf{Ob}(S) \\ &\quad \forall w \in (W_R \cap W_{R'}) \\ &\quad \forall m \in \mathbf{Hom}_S(R, R') \\ &s \in d_R(w) \;\Rightarrow\; m(s) \in d_{R'}(w) \end{aligned} \tag{2}$$

3. **Uniqueness criterion: There should be at most one mapping from one repository to another.**

   The uniqueness criterion states that for each pair of repositories R and R', there is at most one mapping from R to R', and at most one mapping from R' to R, making S a *posetal* or *thin* category. Note that this criterion is direction-sensitive, so for each pair of repositories, there can be at most two mappings, one in each direction. For example, SensEmBert embeddings are mapped one-to-one to WordNet synsets, and vice versa. This criterion prevents WordNet embeddings from being mapped to a different WordNet synset, or vice versa.

   Formally:

$$\forall R, R' \in \mathbf{Ob}(S) \quad |\mathbf{Hom}_S(R, R')| = 1 \tag{3}$$

4. **Connectivity: A sense system should be a connected category.**

   The connectivity criterion states that S is a connected category, i.e. all repositories in $\mathbf{Ob}(S)$ and their mappings in $\mathbf{Hom}(S)$ must form a single connected graph. For example, WordNet synsets are mapped to CSI labels, but neither are mapped to or from, say, the *Macmillan English Dictionary*. This means that the sense system formed by these three repositories does not fulfil the connectivity criterion.

   Formally, for any two repositories R and R' in $\mathbf{Ob}(S)$, there is a sequence $R = R_0,\ R_1,\ R_2,\ ...\ R_n = R'$ where $(R_0, ..., R_n) \in \mathbf{Ob}(S)$, and for each i up to (but not including) n, there is at least one mapping in either $\mathbf{Hom}_S(R_i, R_i + 1)$ or $\mathbf{Hom}_S(R_i + 1, R_i)$.

4.1. Motivation

4.1.1. Correctness preservation

This criterion is endorsed by virtually all existing mappings. Without this assumption, existing mappings would be unusable. Nonetheless, repositories occasionally contain errors, particularly ones which are automatically mapped. Because of this, manual annotations are more highly valued (Pradhan and Xue, 2009), while automatically mapped repositories are often evaluated afterwards to reveal errors. For example, Seppälä et al. (2016) checked their automatically generated mappings against their manually identified mappings for medicine-related words, and discovered that only 85% were correctly identified automatically. They also found two "obvious mistakes" made during manual annotation, which were promptly corrected.

Since mappings are not multivalued (section 3), preserving correctness allows us to cross-check labelled data for any inconsistencies. Using the word *mouse* as an example, one annotator or classifier might select the WordNet synset referring to the rodent, and another might select the WordNet Domain label of "computer science". Since the rodent synset is not mapped to "computer science", we know (by *modus tollens*) that there was a disagreement between the two annotators/classifiers, even though they make use of different sense repositories.

Note that the correctness preservation is only defined with respect to the selection of the correct sense, but does not place any restrictions on candidacy and word type.

4.1.2. Candidacy preservation

Candidacy preservation is intuitive from a semantic perspective. If a word sense s is mapped to a semantically more encompassing word sense s', it must be the case that this broader sense is also a candidate. This criterion is trivially fulfilled by clustering-based approaches, but is not typically explicitly stated for repositories.

A violation would only occur if an instance of a word type could carry the sense s without also being able to carry s' in any context. Such a violation would suggest that s' has some semantic specificity that s lacks. For example, the WordNet synset `mind.n.01` (with the gloss definition "that which is responsible for one's

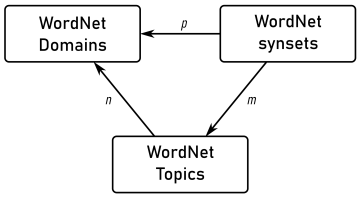

Figure 3: Mappings for WordNet synsets, WordNet Domains, and WordNet Topics have been created. By the compositionality of morphisms and the uniqueness criterion, $n \circ m = p$.

thoughts and feelings; the seat of the faculty of reason") is a candidate sense for the word types *brain*, *head*, *psyche*, and *nous*. If this synset is mapped to a domain label called "anatomy", it would be a violation of candidacy preservation, because "anatomy" is not a candidate sense for "psyche" or "nous".

Relatedly, candidacy preservation is required for a straightforward way of comparing granularity levels for each word type: by counting the number of senses. For example, WordNet 3.1 contains 42 senses for *head*, while the online *Oxford Learner's Dictionary* contains 20. If we map all of WordNet's synsets to OLD entries and preserve candidacy, we can postulate that the 20 senses are coarser-grained than the 42 in WordNet. On the other hand, if we do not preserve candidacy, it may be the case that semantic content was lost in applying the mapping, and hence the fewer senses of the *Oxford Learner's Dictionary* might not be more coarse-grained, but just leave semantic gaps.

4.1.3. Uniqueness

For many existing mappings that were produced through clustering (Dolan, 1994; Vial et al., 2019), the uniqueness criterion is assumed implicitly, because each sense can belong to at most one cluster. The same is true for embedding-based senses that are mapped one-to-one to a dictionary-based repository.

However, there are other types of mappings that do not fulfil this criterion. As mentioned in Section 3, WordNet Domains maps the synset for *amethyst* to the domains of "chemistry", "geology", and "jewellery". Similarly, the Coarse Sense Inventory (CSI) (Lacerra et al., 2020) maps the synset for *abbatoir* to "craft, engineering, and technology", "art, architecture, and archaeology", and "food, drink, and taste".

We argue that enforcing the uniqueness criterion provides several benefits:

1. Repositories in S would form a partial preorder, which would roughly correspond to the notion of granularity. Since mappings are total and cannot be multivalued, the range (or image) of the mapping must have cardinality less than or equal to that of the domain. The cardinality thus reflects a notion of granularity that is measured numerically.[3]

2. There would be more consistency when converting between labels. For example, Izquierdo et al. (2007) mapped each WordNet synset to one Base Level Concept (BLC), so one could consistently convert from the former to the latter. A WSD tool or downstream application that uses BLC-annotated corpora can automatically make use of a WordNet-annotated corpus such as SemCor (Landes et al., 1998), because the labels can be directly converted into BLCs.

3. In a similar vein, evaluation metrics that depend on converted labels would be more reliable. A WSD classifier using BLCs can easily be evaluated according to SemCor, because there is only one correct BLC that each word is mapped to. On the other hand, if WordNet synsets are mapped to multiple BLCs, it is not clear how the classifier should be evaluated. The BLCs might all be considered correct, resulting in inflated scores; or if a random one is chosen, the scores may not accurately reflect the classifier's performance.

4. In conjunction with function composition (see Section 3), the uniqueness criterion would also enforce transitivity. Consider WordNet synsets, WordNet topics, and WordNet Domains in Figure 1: if the mappings between these repositories fulfil the uniqueness criterion, there would only be at most one mapping between each repository, as in Figure 3. Under function composition, $n \circ m = p$ (where n, m, and p correspond to mappings in Figure 3).

One might argue that the domain labels for *amethyst* and *abbatoir* should not be interpreted as separate labels, but instead as a set containing all relevant domains; so one would map WordNet synsets to the *power set* of CSI or Domain labels. However, adapting classifier models (for WSD or otherwise) to handle multiple labels instead of one is not always straightforward, so ideally a sense system should only contain sets of senses, not sets of sets of senses.

Another practical solution is to designate one main CSI or Domain label for each WordNet synset, so that all conversions and comparisons will be made according to one label. This main label could be chosen based on inter-annotator agreement or frequency or another metric, as long as it is consistent across all synsets. Other non-designated labels can still be made available for classifiers that can handle multiple labels.

[3]This correspondence of course only applies to the range, but not the whole co-domain. In practice, mappings are usually surjective (so the co-domain *is* the range) — exceptions are limited to newer or more specialised vocabulary. For example, English WordNet (`https://en-word.net/`) contains the definition of *dab* that refers to the dance move, which is not in Princeton WordNet 3.1.

In either case, formalising the uniqueness criterion explicitly provides a better understanding of the potential problems and associated tradeoffs when the criterion is not met. It also allows researchers to evaluate current and future repositories according to specific needs and resources.

4.1.4. Connectivity

Previous work on WSD have focused on building mappings between repositories rather than a complete sense system, so connectivity is rarely assumed. However, in the few cases where more than two repositories were mapped (Gella et al., 2014; Palmer et al., 2004), the resulting sense systems do fulfil the connectivity criterion.

The connectivity criterion on its own is not very informative, but it enables other criteria by extending their benefits to the rest of the sense system. After all, an unconnected sense system technically fulfils all the other criteria in this paper, but is not very useful. As mentioned above, the previous three criteria each had their own practical and theoretical benefits: 1) correctness preservation allowing cross-checking; 2) candidacy preservation allowing comparison of grain level; and 3) uniqueness allowing consistent label conversion. If the connectivity criterion is fulfilled, these benefits can be extended to any two repositories in $\mathbf{Ob}(S)$.

With a sufficient number of repositories in $\mathbf{Ob}(S)$, one can leverage these benefits on a larger scale, opening up new opportunities for WSD research. For example, ensemble classifiers based on different sense repositories can be built: if there are three WSD classifiers that use senses from R, R', and R'' respectively, their outputs can be aggregated and cross-checked, as long as R, R', and R'' are connected to each other in a single graph.

5. Guiding criteria for sense systems

While all criteria listed in this paper are desirable for various reasons, the **basic** criteria are ones which can be fulfilled both in theory and in practice, while the **guiding** criteria may be impossible to fulfil in certain situations, and should be considered more as approximate guidelines than strict criteria.

In addition to the 4 basic criteria, we propose two additional guiding criteria:

1. **Non-contradiction: Mappings cannot exist between senses that semantically contradict each other.**

   The non-contradiction criterion forbids mappings between senses whose (strict) implications contradict each other. Examples of such contradictions can easily be found in the literature: the word *monograph* has (at least) two fine-grained senses, one referring to the physical printed volume by an author, another referring to the abstract piece of work instantiated by such a volume. These two senses might be mapped to one coarse-grained sense in a different repository, where it is categorised as a physical object. Thus arises a contradiction where the fine-grained sense referring to the abstract work is mapped to a coarse-grained sense referring to a physical object.

   We formalise the non-contradiction criterion as follows:

$$\begin{aligned} &\forall R, R' \in \mathbf{Ob}(S) \\ &\quad \forall m \in \mathbf{Hom}_S(R, R') \\ &\quad\ \forall s \in R \\ &s \vDash P \ \Rightarrow\ \neg(m(s) \vDash \neg P) \end{aligned} \tag{4}$$

   where $\vDash$ indicates strict entailment and P is any proposition.

   Note that the correctness criterion does not entail the non-contradiction criterion. In the *monograph* example, the mapping fulfils the correctness preservation because a WSD oracle would consider the coarse-grained sense to be correct, despite the contradiction.

2. **Inter-annotator agreement: Mappings should correspond to a partial preorder of inter-annotator agreement levels.**

   It has been observed that, when annotating corpora with senses from a given sense repository, inter-annotator agreement tends to drop when the repository is more fine-grained (Ng et al., 1999; Navigli, 2009). Therefore, if R is coarser-grained than R', one can expect agreement levels to be higher when annotating corpora with senses in R, compared to R'.

   We formalise this criterion as follows:

$$\begin{aligned} &\forall R, R' \in \mathbf{Ob}(S) \\ &(\exists m \in \mathbf{Hom}_S(R, R')) \ \Rightarrow\ (a(R) \leq a(R')) \end{aligned} \tag{5}$$

   where a refers to the inter-annotator agreement, defined by $a : \mathbf{Ob}(S) \rightarrow \mathbb{R}$. $\exists m \in \mathbf{Hom}_S(R, R')$ means that there is at least one mapping from R to R'.

5.1. Motivation

5.1.1. Non-contradiction

Non-contradiction is considered a guiding criterion because, while it is desirable, it is also a difficult criterion to meet. Firstly, some sense representations (such as embeddings) do not come with explicit semantics, so it would be impossible to determine if their implications contradict one another. Secondly, semantic implications are often subtle and difficult to identify: even WordNet, a repository known for its fine-grained senses, does not distinguish the two senses in the *monograph* example above.

However, mappings that do meet the non-contradiction criterion can be useful in downstream tasks that require natural language inference, such as question answering or information extraction. For example, with the correct sense labels, an information extraction tool could eliminate the possibility of an abstract *book* having the same referent as a physical *monograph*. Alternatively, mappings that do not meet the criterion might cause errors in these downstream applications. For the question "When was this monograph created?", a question-answering system might incorrectly assume the physicality of the object in question, and describe the time when the monograph was printed instead of when the text was written.

Some sense repositories that are formed through clustering techniques do not contain any semantic content. For example, clustering WordNet synsets based on confusion matrices (Agirre and Lacalle, 2003) would create clusters that are not explicitly associated with a label or definition. These mappings trivially fulfil the non-contradiction criterion. However, there are also clustering techniques where this criterion does apply: for example, Navigli (2006) makes use of the hierarchical semantic structures in the *Oxford Dictionary of English* to cluster WordNet synsets. As a result, the clusters produced are associated with a textual definition and other semantic information.

5.1.2. Inter-annotator agreement

We previously demonstrated that mapped repositories in a posetal sense system (fulfilling the uniqueness criterion) form a partial preorder of granularity. If the inter-annotator agreement criterion is fulfilled, mapped repositories would also form a partial preorder of inter-annotator agreement levels.

This criterion is considered a guiding criterion because, unlike basic criteria, it cannot be directly enforced — researchers have no reason to artificially inflate or lower inter-annotator agreement. Additionally, this criterion cannot be applied to sense representations that are not used for human annotation, such as word embeddings. Nevertheless, this criterion not only reflects existing expectations for a sense system, but strong violations suggest that the sense distinctions of the coarse-grained sense repository are unnatural, i.e. not in accordance with human linguistic intuitions, since the annotators appear to struggle more despite a reduction in labels.

6. Conclusion

This paper develops a representation of sense systems as categories, and proposes a list of criteria that serve as guidelines for future sense repositories and mappings. The list is by no means exhaustive, as there are other properties that may be desirable depending on the downstream application.

A sense system that fulfils our list of criteria brings multiple benefits and opportunities to the WSD task: not only does it provide theoretical grounding for sense mappings, it also opens up other opportunities to improve existing WSD tools, such as extending them to ensemble classifiers that can crosscheck annotation from multiple sense repositories.

Agirre, E. and Lacalle, O. L. D. (2003). Clustering Wordnet Word Senses. In *Proceedings of the Conference on Recent Advances on Natural Language (RANLP'03)*. http://ixa3.si.ehu.es/cgi-bin/ signatureak/signaturecgi.cgi http://ixa2.si.ehu.es/pub/webcorpus.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 94–101, Geneva, Switzerland, August. COLING.

Camacho-Collados, J. and Navigli, R. (2017). Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, page 223–228. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Dolan, W. B. (1994). Word sense ambiguation: Clustering related senses. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, page 712–716. Association for Computational Linguistics.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press, Cambridge, Mass.

Gella, S., Strapparava, C., and Nastase, V. (2014). Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources. In *LREC*, pages 1117–1121.

Green, R., Pearl, L., Dorr, B. J., and Resnik, P. (2001). Mapping lexical entries in a verbs database to WordNet senses. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 244–251, Toulouse, France, July. Association for Computational Linguistics.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 897–907. Association for Computational Linguistics.

Ide, N. and Wilks, Y. (2007). Making Sense About Sense. In Eneko Agirre et al., editors, *Word Sense Disambiguation: Algorithms and Applications*, Text, Speech and Language Technology, pages 47–73. Springer Netherlands, Dordrecht.

Izquierdo, R., Suarez, A., and Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. In *Recent Advances in Natural Language Processing*, pages 298–302, Borovets, Bulgaria.

Kågebäck, M. and Salomonsson, H. (2016). Word sense disambiguation using a bidirectional lstm. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, page 51–56. The COLING 2016 Organizing Committee, Dec.

Kilgarriff, A. (2003). "I Don't Believe in Word Senses". In Brigitte Nerlich, et al., editors, *Polysemy*. DE GRUYTER MOUTON, Berlin, New York, January.

Lacerra, C., Bevilacqua, M., Pasini, T., and Navigli, R. (2020). CSI: A Coarse Sense Inventory for 85% Word Sense Disambiguation. In *Proc. of AAAI*.

Landes, S., Leacock, C., and Tengi, R. I. (1998). Building semantic concordances. In *WordNet: an electronic lexical database*, chapter 8, pages 199–216. MIT Press, Cambridge, MA.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26. ACM.

Magnini, B. and Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. In *LREC*, pages 1413–1418.

Mihalcea, R. and Faruque, E. (2004). Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, page 155–158. Association for Computational Linguistics, Jul.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc. event-place: Lake Tahoe, Nevada.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, December. Publisher: Oxford Academic.

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

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 105–112, Sydney, Australia. Association for Computational Linguistics.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69, February.

Ng, H. T., Lim, C. Y., and Foo, S. K. (1999). A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. In *SIGLEX99: Standardizing Lexical Resources*.

Oele, D. and Noord, G. v. (2017). Distributional lesk: Effective knowledge-based word sense disambiguation. In *IWCS 2017 — 12th International Conference on Computational Semantics: Short papers*, pages W17–6931.

Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different Sense Granularities for Different Applications. In *Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding (ScaNaLU 2004) at HLT-NAACL 2004*, pages 49–56.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making Fine-Grained and Coarse-Grained Sense Distinctions, Both Manually and Automatically. *Natural Language Engineering*, 13(2):137–163, June.

Peters, W., Peters, I., and Vossen, P. (1998). Automatic Sense Clustering in Eurowordnet. In *Proceedings of LREC'1998*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Pradhan, S. S. and Xue, N. (2009). OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado, May. Association for Computational Linguistics.

Scarlini, B., Pasini, T., and Navigli, R. (2020a). SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proc. of AAAI*.

Scarlini, B., Pasini, T., and Navigli, R. (2020b). With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Seppälä, S., Hicks, A., and Ruttenberg, A. (2016). Semi-automatic mapping of WordNet to Basic Formal Ontology. In Verginica Barbu Mititelu, et al., editors, *Proceedings of the Eighth Global WordNet Conference*, pages 369–376, Bucharest, Romania, January 27-30.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. *ArXiv*.

Piek Vossen, editor. (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Springer Netherlands.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, page 161–170. German Society for Computational Linguistics & Language Technology.

Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL*, page 78–83.