From Word to Sense Embeddings: A Survey on Vector Representations of Meaning

Jose Camacho-Collados CAMACHOCOLLADOSJ@CARDIFF.AC.UK
Cardiff University, United Kingdom

Mohammad Taher Pilehvar MP792@CAM.AC.UK
University of Cambridge, United Kingdom

Abstract

Over the past years, distributed representations have proven effective and flexible keepers of prior knowledge to be integrated into downstream applications. This survey is focused on semantic representation of meaning. We start from the theoretical background behind word vector space models and highlight one of their main limitations: the meaning conflation deficiency, which arises from representing a word with all its possible meanings as a single vector. Then, we explain how this deficiency can be addressed through a transition from word level to the more fine-grained level of word senses (in its broader acceptation) as a method for modelling unambiguous lexical meaning. We present a comprehensive overview of the wide range of techniques in the two main branches of sense representation, i.e., unsupervised and knowledge-based. Finally, this survey covers the main evaluation procedures and provides an analysis of five important aspects: interpretability, sense granularity, adaptability to different domains, compositionality and integration into downstream applications.

1. Introduction

Recently, neural network based approaches which process massive amounts of text data to embed words' semantics into low-dimensional vector spaces, the so-called word embeddings, have become increasingly popular (Mikolov, Chen, Corrado, & Dean, 2013a; Pennington, Socher, & Manning, 2014). Word embeddings have been effective in storing valuable syntactic and semantic information (Mikolov, Yih, & Zweig, 2013b); hence, they have proven beneficial in many Natural Language Processing (NLP) tasks, mainly due to their generalization power (Goldberg, 2016). A wide range of applications have reported improvements upon integrating word embeddings, including Machine Translation (Zou, Socher, Cer, & Manning, 2013), syntactic parsing (Weiss, Alberti, Collins, & Petrov, 2015), text classification (Kim, 2014) and question answering (Bordes, Chopra, & Weston, 2014), to name a few. However, despite their flexibility and success in capturing semantic properties of words, the effectiveness of these representations are generally hampered by an important limitation which we will refer to as *meaning conflation deficiency*: the inability to discriminate among different meanings of a word.

A word can have one meaning (monosemous) or multiple meanings (ambiguous). For instance, the noun *nail* can refer to two different meanings depending on the context: a part of the finger or a metallic object. Hence, the noun *nail* is said to be ambiguous[1]. Each

1. *Nail* can also refer to a unit of cloth measurement or even be used as a verb.

individual meaning of an ambiguous word is called a word sense and a lexical resource that lists different meanings (senses) of words is usually referred to as a sense inventory.[2] While most words in general sense inventories (e.g. WordNet) are often monosemous[3], frequent words tend to have more senses, according to the Principle of Economical Versatility of Words (Zipf, 1949). Therefore, accurately capturing the semantics of ambiguous words plays a crucial role in the language understanding of NLP systems.

In order to deal with the meaning conflation deficiency, a number of approaches have attempted to model individual word senses. In this survey we have tried to synthesize the most relevant works on sense representation learning. The main distinction of these approaches is in how they model meaning and where they obtain it from. *Unsupervised* models directly learn word senses from text corpora, while *knowledge-based* systems exploit the sense inventories of lexical resources as their main source for representing meanings. In this survey we cover these two kinds of approach for learning distributed semantic representations of meaning, including evaluation procedures and an analysis of their main properties. While the survey is intended to be as extensive as possible, given the breadth of the topics covered, some areas may not have received a sufficient coverage to be totally self-contained. However, on these cases we provide relevant pointers so the readers can have a deeper look into those topics on their own. Given the wide audience that this survey is intended to reach, we have tried to make it as understandable as possible. Therefore, not many algorithmic details are provided in full detail, but rather the intuition behind them.

The remainder of this survey is structured as follows. First, in Section 2 we provide a theoretical background on word senses, what they are, why modeling them may be useful and its main paradigms. Then, in Section 3 we describe unsupervised sense vector modeling techniques which learn directly from text corpora, while in Section 4 the representations linked to lexical resources are explained. Common evaluation procedures and benchmarks are presented in Section 5. Finally, we present an analysis and comparison between unsupervised and knowledge-based representations in Section 6 and the main conclusions and future challenges in Section 7.

2. Background

This section provides theoretical foundations which support the move from word level to the more fine-grained level of word senses and concepts. First, we provide a background on vector space models, particularly for word representation learning (Section 2.1). Then, we explain some of the main deficiencies of word representations which led to the development of sense modeling techniques (Section 2.2) and present the main paradigms for representing senses (Section 2.3). Finally, we explain the notation followed throughout the survey (Section 2.4).

2. In order to obtain the list of possible word senses of a target word, lexicographers tend to first collect occurrences of the words from corpora and then manually cluster them semantically and based on their contexts, i.e., *concordance* (Kilgarriff, 1997). Given this procedure, Kilgarriff (1997) suggested that word senses, as defined by sense inventories in NLP, should not be construed as objects but rather as abstractions over clusters of word usages.
3. For instance, around 83% of the 155K words in WordNet 3.0 are listed as monosemous (see Section 4.1 for more information on lexical resources).

2.1 Word Representation Learning

Word representation learning has been one of the main research areas in Semantics since the beginnings of NLP. We first introduce the main theories behind word representation learning based on vector space models (Section 2.1.1) and then move to the emerging theories for learning word embeddings (Section 2.1.2).

2.1.1 Vector Space Models

One of the most prominent methodologies for word representation learning is based on Vector Space Models (VSM), which is supported by research in human cognition (Landauer & Dumais, 1997; Gärdenfors, 2004). The earliest VSM applied in NLP considered a document as a vector whose dimensions were the whole vocabulary (Salton, Wong, & Yang, 1975). Weights of individual dimensions were initially computed based on word frequencies within the document. Different weight computation metrics have been explored, but mainly based on frequencies or normalized frequencies (Salton & McGill, 1983). This methodology has been successfully refined and applied to various NLP applications such as information retrieval (Lee, Chuang, & Seamons, 1997), text classification (Soucy & Mineau, 2005), or sentiment analysis (Turney, 2002), to name a few. Turney and Pantel (2010) provides a comprehensive overview of VSM and their applications.

The document-based VSM has been also extended to other lexical items like words. In this case a word is generally represented as a point in a vector space. A word-based vector has been traditionally constructed based on the normalized frequencies of the co-occurring words in a corpus (Lund & Burgess, 1996), by following the initial theories of Harris (1954). The main idea of these word VSM is that words that share similar context should be close in the vector space (therefore, have similar semantics). Figure 1 shows an example of a word VSM where this underlying proximity axiom is clearly highlighted. These models have been also proven effective in NLP tasks such as information extraction (Laender, Ribeiro-Neto, da Silva, & Teixeira, 2002), semantic role labeling (Erk, 2007), word similarity (Radinsky, Agichtein, Gabrilovich, & Markovitch, 2011), word sense disambiguation (Navigli, 2009) or spelling correction (Jones & Martin, 1997), *inter alia*.

One of the main drawbacks of these approaches is the high dimensionality of the produced vectors. Since the dimensions correspond to words in the vocabulary, this number could easily add to the hundreds of thousands or even millions, depending on the underlying corpus. A common approach for dimensionality reduction make use of the Singular Value Decomposition (SVD) and is known as Latent Semantic Analysis (Hofmann, 2001; Landauer & Dooley, 2002, LSA). In addition to this technique, recent models also leverage neural networks to directly learn low-dimensional word representations. These models are introduced in the following section.

2.1.2 Word embeddings

Learning low-dimensional vectors from text corpora can alternatively be achieved by exploiting neural network models. These models are commonly known as *word embeddings* and have been shown to provide a valuable prior knowledge thanks to their generalization power (Goldberg, 2016). This property has proved decisive for achieving state-of-the-art

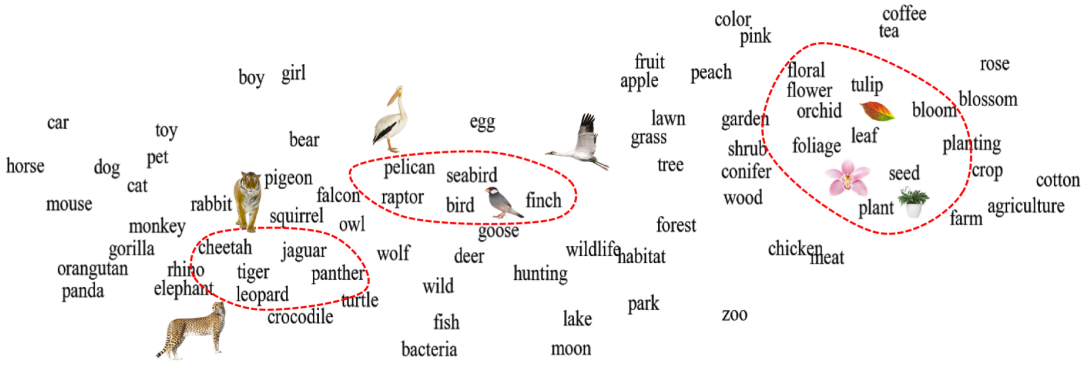


Figure 1: Word vector space reduced to two dimensions using t-SNE (Maaten & Hinton, 2008). In a semantic space, words with similar meanings tend to appear in the proximity of each other, as highlighted by these word clusters (delimited by the red dashed lines) associated with *big cats*, *birds* and *plants*.

performance in many NLP tasks when integrated into a neural network architecture (Zou et al., 2013; Kim, 2014; Bordes et al., 2014; Weiss et al., 2015).

This newer predictive branch was popularized through Word2vec (Mikolov et al., 2013a), a simple but efficient architecture which additionally provides interesting semantic properties of the output vectors (Mikolov et al., 2013b). Two different but related models were proposed: Continuous Bag-Of-Words (CBOW) and Sikp-gram. The CBOW architecture is based on the feedforward neural network language model (Bengio, Ducharme, Vincent, & Janvin, 2003) and aims at predicting the current word using its surrounding context, minimizing the following loss function:

$$E = -\log(p(\vec{w_t}|\vec{W_t})) \tag{1}$$

where w_t is the target word and $W^t = w_{t-n}, ..., w_t, ..., w_{t+n}$ represents the sequence of words in context. Figure 2 shows a simplification of the general architecture of the CBOW and Skip-gram models of Word2vec. The architecture consists of input, hidden and output layers. The input layer has the size of the word vocabulary and encodes the context as a combination of one-hot vector representations of surrounding words of a given target word. The output layer has the same size of the input layer and contains a one-hot vector of the target word during the training phase. The Skip-gram model is similar to the CBOW model but in this case the goal is to predict the words in the surrounding context given the target word, rather than predicting the target word itself. Interestingly, Levy and Goldberg (2014) proved that Skip-gram can be in fact viewed as an implicit factorization of a Point-Mutual Information (PMI) co-occurrence matrix.

Another prominent word embedding architecture is GloVe (Pennington et al., 2014), which combines global matrix factorization and local context window methods through a bilinear regression model. In recent years more complex approaches that attempt to solve some deficiencies of these models have been proposed, including models leveraging subword units (Wieting, Bansal, Gimpel, & Livescu, 2016; Bojanowski, Grave, Joulin, &

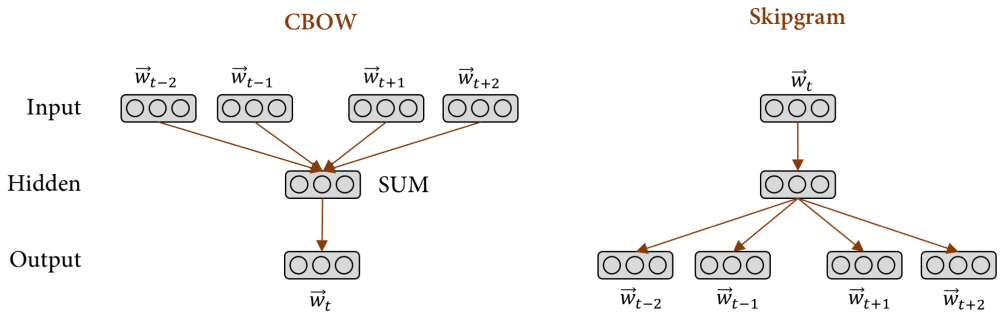


Figure 2: Learning architecture of the CBOW and Skipgram models of Word2vec (Mikolov et al., 2013a).

Mikolov, 2017), representing words as probability distributions (Vilnis & McCallum, 2015; Athiwaratkun & Wilson, 2017) or exploiting knowledge resources (more details about these techniques in Section 4.2).[4]

2.2 Meaning Conflation Deficiency

The prevailing objective of representing each word type as a single point in the semantic space has a major limitation: it ignores the fact that words can have multiple meanings and conflates all these meanings into a single representation. Schütze (1998) was one of the first works to identify the meaning conflation deficiency of word vectors. Having different (possibly unrelated) meanings conflated into a single representation can hamper the semantic understanding of an NLP system that uses these at its core. In fact, word embeddings have been proved to not be able to effectively capture different meanings of the same word, even when these meanings occur in the underlying training corpus (Yaghoobzadeh & Schütze, 2016). This meaning conflation can have additional negative impacts on accurate semantic modeling, e.g., semantically unrelated words that are similar to different senses of a word are pulled towards each other in the semantic space (Neelakantan, Shankar, Passos, & McCallum, 2014; Pilehvar & Collier, 2016). For example, the two semantically-unrelated words *rat* and *screen* are pulled towards each other in the semantic space for their similarities to two different senses of *mouse*, i.e., rodent and computer input device. See Figure 3 for an illustration.[5] Moreover, the conflation deficiency violates the triangle inequality of euclidean spaces, which can reduce the effectiveness of word space models (Tversky & Gati, 1982). In order to alleviate this deficiency, a new direction of research has emerged over the past years, which tries to directly model individual meanings of words. In this survey we focus on this new branch of research, which has some similarities and peculiarities with respect to word representation learning.

4. For a more comprehensive overview on word embeddings and their current challenges, please refer to Ruder (2017).
5. Dimensionality was reduced using PCA; visualized with http://projector.tensorflow.org/.

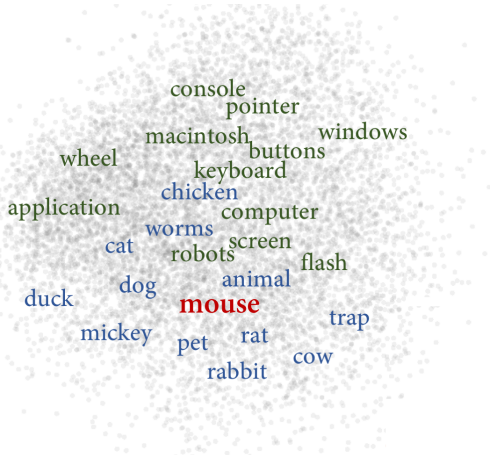


Figure 3: An illustration of the meaning conflation deficiency in a 2D semantic space around the ambiguous word *mouse*. Having the word, with its different meanings, represented as a single point (vector) results in pulling together of semantically unrelated words, such as *computer* and *rabbit*.

2.3 Paradigms

The optimal way of partitioning the meanings of words into multiple senses has long been the point of argument (Erk, McCarthy, & Gaylord, 2009; Erk, 2012; McCarthy, Apidianaki, & Erk, 2016). Traditionally, computational techniques have relied on fixed sense inventories produced by humans, such as WordNet (Fellbaum, 1998). A sense inventory[6] is a lexical resource, such as a dictionary or thesaurus, that lists for each word the possible meanings it can take. Sense distinctions can also be defined through word sense induction, i.e., automatic identification of a word's senses by analyzing the contexts in which it appears.

Generally, sense representations can be divided into two main paradigms depending on how the sense distinctions are defined:

- **Unsupervised.** In these representation models the sense distinctions are induced by analyzing text corpora.
- **Knowledge-based.** These techniques represent word senses as defined by an external sense inventory, e.g., WordNet. As a result, these sense representations are linked to external sense inventories.[7]

In the following two main sections (Sections 3 and 4), we will provide details of each paradigm and their variants.

2.4 Notation

Throughout this survey we use the following notation. Words will generally be referred to as w while senses will be written as s. Concepts, entities and relations will be referred to

6. In Section 4.1 we provide an overview of few of the most popular sense inventories.
7. In this survey we also cover representations directly linked to knowledge resources even if senses are not explicitly listed (e.g., concepts and entities in Wikipedia), including knowledge base embeddings.

as c, e and r, respectively. Following previous work (Navigli, 2009), we use the following interpretable expression for senses as well: $word_n^p$ is the n^{th} sense of *word* with part of speech p. As for synsets as represented in a sense inventory we will use y.[8] A semantic network will be generally represented as N. In order to refer to vectors we will add the vector symbol on the top of each item. For instance, $\vec{w}$ and $\vec{s}$ will refer to the vectors of the word w and sense s, respectively.

In general in this survey we may refer to *sense* as the general term including all representations of meaning beyond the word level, or explicitly to a word associated with a specific meaning[9] (e.g., *bank* with its *financial* meaning), irrespective of whether the meaning belongs to a pre-defined sense inventory or not, and whether it refers to a concept (e.g., *banana*) or an entity (e.g., *France*).

3. Unsupervised Sense Representations

Unsupervised sense representations are constructed on the basis information extracted from text corpora only. Word sense induction, i.e., automatic identification of possible meanings of words, lies at the core of these techniques. An unsupervised model induces different senses of a word by analysing its contextual semantics in a text corpus and represent each sense based on the statistical knowledge derived from the corpus. Depending on the type of text corpus used by the model, we have split unsupervised sense representations into two categories: (1) exploiting monolingual corpora only (Section 3.1) and (2) exploiting multilingual corpora (Section 3.2).

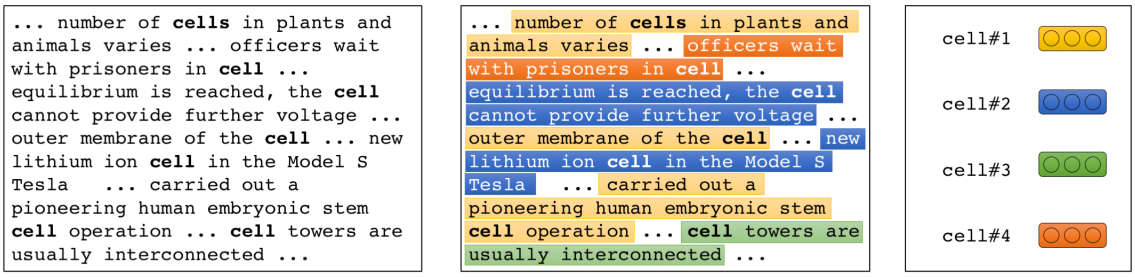
3.1 Sense Representations Exploiting Monolingual Corpora

In this section we present approaches that make use of solely unlabeled monolingual corpora to learn sense representations. These approaches can be divided into two main groups: (1) **clustering (two-stage models)** (Van de Cruys, Poibeau, & Korhonen, 2011; Erk & Padó, 2008; Liu, Qiu, & Huang, 2015a), which first induce senses and then learn representations for these (Section 3.1.1), and (2) **joint training** (Li & Jurafsky, 2015; Qiu, Tu, & Yu, 2016), which perform the induction and representation learning together (Section 3.1.2). In Section 3.1.3, we will briefly overview **contextualized embeddings**, an emerging branch of unsupervised techniques which views sense representation from a different perspective.

3.1.1 Two-stage models

The *context-group discrimination* of Schütze (1998) is one of the pioneering works in sense representation. The approach was an attempt to *automatic* word sense disambiguation in order to address the knowledge-acquisition bottleneck for sense annotated data (Gale, Church, & Yarowsky, 1992) and reliance on external resources. The basic idea of context-group discrimination is to automatically induce senses from contextual similarity, computed by **clustering** the context in which an ambiguous word occurs. Specifically, each context C of an ambiguous word w is represented as a context vector $\vec{v}_C$, computed as the centroid of its content words' vectors $\vec{v}_c$ ($c \in C$). Context vectors are computed for each word in a given

8. See Section 4.1 for more information about the notions of these resource-related concepts.
9. In other works *senses* have also been referred to as *lexemes* on this case (Rothe & Schütze, 2015).



(1) Get occurrences of a word from text corpora

(2) Analyze contexts and induce senses of the word

(3) Compute sense representation

Figure 4: Unsupervised sense representation techniques first induce different senses of a given word (usually by means of clustering occurrences of that word in a text corpus) and then compute representations for each induced sense.

corpus and then clustered into a predetermined number of clusters (context groups) using the Expectation Maximization algorithm (Dempster, Laird, & Rubin, 1977, EM). Context groups for the word are taken as representations for different senses of the word.

Despite its simplicity, the clustering-based approach of Schütze (1998) constitutes the basis for many of the subsequent techniques, which mainly differed in their representation of context or the underlying clustering algorithm. Figure 4 depicts the general procedure followed by the two-stage unsupervised sense representation techniques.

Given its requirement for computing independent representations for all individual contexts of a given word, the context-group discrimination approach is not scalable to large corpora. Reisinger and Mooney (2010) addressed this by directly clustering the contexts, represented as feature vectors of unigrams, instead of modeling contexts as vectors. The approach can be considered as the first new-generation sense representation technique, which is often referred to as *multi-prototype*. In this specific work, contexts were clustered using Mixtures of von Mises-Fisher distributions (movMF) algorithm. The algorithm is similar to k-means but permits controlling the semantic breadth using a per-cluster concentration parameter which would better model skewed distribution of cluster sizes.

Similarly, Huang, Socher, Manning, and Ng (2012) proposed a clustering-based sense representation technique with three differences: (1) context vectors are obtained by a idf-weighted averaging of their word vectors; (2) spherical k-means is used for clustering; and (3) most importantly, occurrences of a word are labeled with their cluster and a second pass is used to learn sense representations. The idea of 2-pass learning has also been employed by Vu and Parker (2016) for another sense representation modeling architecture.

Sense representations can also be obtained from semantic networks. For instance, Pelevina, Arefyev, Biemann, and Panchenko (2016) constructed a semantic graph by connecting each word to the set of its semantically similar words. Nodes in this network were clustered using the Chinese Whispers algorithm (Biemann, 2006) and senses were induced as a weighted average of words in each cluster. A similar sense induction technique was employed by Sense-aware Semantic Analysis (Wu & Giles, 2015, SaSA). SaSA follows Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007, ESA) by representing a word using Wikipedia

concepts. Instead of constructing a nearest neighbour graph, a graph of Wikipedia articles is built by gathering all related articles to a word w and clustering them. The sense induction step is performed on the semantic space of Wikipedia articles.

3.1.2 Joint models

The clustering-based approach to sense representation suffers from the limitation that clustering and sense representation are done independently from each other and as a result the two stages do not take potential advantage of their similarities. The introduction of embedding models was one of the most revolutionary changes to vector space models of word meaning. As a closely related field, sense representations did not remain unaffected. Many researchers have proposed various extensions of the Skip-gram model (Mikolov et al., 2013a) which would enable the capture of sense-specific distinctions. A major limitation of the two-stage models is their computational expensiveness[10]. Thanks to the efficiency of these algorithms and their unified nature (as opposed to the two-phase nature of more conventional techniques) these techniques are generally efficient. Hence, recent works can mostly be grouped into this category.

Neelakantan et al. (2014) was the first to propose a multi-prototype extension of the Skip-gram model. Their model, called Multiple-Sense Skip-Gram (MSSG), is similar to earlier work in that it represents the context of a word as the centroid of it words' vectors and clusters them to form the target word's sense representation. Though, the fundamental difference is that clustering and sense embedding learning is performed jointly. During training, the intended sense for each word is dynamically selected as the closest sense to the context and weights are updated only for that sense. In a concurrent work, Tian, Dai, Bian, Gao, Zhang, Chen, and Liu (2014) proposed a Skip-gram based sense representation technique that significantly reduced the number of parameters of Huang et al. (2012). The model replaces word embeddings in the Skip-gram model with a finite mixture model in which each mixture corresponds to a prototype of the word. The EM algorithm was adopted for the training of this multi-prototype Skip-gram model.

Liu, Liu, Chua, and Sun (2015b) argued that the above techniques are limited in that they consider only the local context of a word for inducing its sense representations. To address this limitation, they proposed Topical Word Embeddings (TWE) in which each word is allowed to have different embeddings under different topics, where topics are computed globally using latent topic modelling (Blei, Ng, & Jordan, 2003). Three variants of the model were proposed: (1) TWE-1, which regards each topic as a pseudo word, and learns topic embeddings and word embeddings separately; (2) TWE-2, which considers each word-topic as a pseudo word, and learn topical word embeddings directly; and (3) TWE-3, which assigns distinct embeddings for each word and each topic and builds the embedding of each word-topic pair by concatenating the corresponding word and topic embeddings. Various extensions of the TWE model have been proposed. The Neural Tensor Skip-gram (NTSG) model (Liu et al., 2015a) applies the same idea of topic modeling for sense representation but introduces a tensor to better learn the interactions between words and topics. Another extension is MSWE (Nguyen, Nguyen, Modi, Thater, & Pinkal, 2017), which argues that

10. For instance, the model of Huang et al. (2012) took around one week to learn sense embeddings for a 6,000 subset of the 100,000 vocabulary on a corpus of one billion tokens (Neelakantan et al., 2014).

multiple senses might be triggered for a word in a given context and replaces the selection of the most suitable sense in TWE by a mixture of weights that reflect different association degrees of the word with multiple senses in the context.

These joint unsupervised models, however, suffer from two relevant limitations. First, due to their straightforward implementation, most unsupervised sense representation techniques assume a fixed number of senses per word. This assumption is far from being realistic. Words tend to have a highly variant number of senses, from one (monosemous) to dozens. In a given sense inventory, usually, most words are monosemous. For instance, around 80% of words in WordNet 3.0 are monosemous, with less than 5% having more than 3 senses. However, ambiguous words tend to occur more frequently in a real text which slightly smooths the highly skewed distribution of words across polysemy. Table 1 shows the distribution of word types by their number of senses in SemCor (Miller, Leacock, Tengi, & Bunker, 1993), one of the largest available sense-annotated datasets which comprises around 235,000 semantic annotations for thousands of words. The skewed distribution clearly shows that word types tend to have varying number of senses in a natural text, as also discussed in other studies (Piantadosi, 2014; Bennett, Baldwin, Lau, McCarthy, & Bond, 2016; Pasini & Navigli, 2018). Second, a common strand of most unsupervised models is that they extend the Skip-gram model by replacing the conditioning of a word to its context (as in the original model) with an additional conditioning on the intended senses. However, the context words in these models are not disambiguated. Hence, a sense embedding is conditioned on the word embeddings of its context. In the following we describe some approaches directly targeted to addressing these two limitations of the joint unsupervised models described above:

1. **Dynamic polysemy.** A direct solution to the varying polysemy problem of sense representation models would be to set the number of senses of a word as defined by an external sense inventory. The Skip-gram extension of Nieto Piña and Johansson (2015) follows this procedure. However, by taking external lexicons as groundtruth the approach suffers from two main limitations. First, the model is unable to handle words that are not defined in the lexicon. Second, the model assumes that the sense distinctions defined by the underlying text match that specified by the lexicon, which might not be necessarily true. In other words, not all senses of a word might have occurred in the text data and the lexicon might not cover all different senses of the word in the underlying text data. A better solution would involve dynamic induction of senses from the underlying text. Such model was first implemented in the non-parameteric MSSG (NP-MSSG) system of Neelakantan et al. (2014). The model applies the online non-parametric clustering procedure of Meyerson (2001) to the task by creating a new sense for a word type only if its similarity (as computed using the current context) to existing senses for the word is less than a parameter λ. AdaGram (Bartunov, Kondrashkin, Osokin, & Vetrov, 2016) improves this dynamic behaviour by a more principled nonparametric Bayesian approach. The model, which similarly to previous works builds on Skip-gram, assumes that the polysemy of a word is proportional to its frequency (more frequent words are probably more polysemous).

2. **Pure sense-based models.** Ideally, a model should model the dependency between sense choices in order to address the ambiguity from context words. Qiu et al. (2016)

# Senses	2	3	4	5	6	7	8	9	10	11	12	≥ 12
Nouns	22%	17%	14%	13%	9%	7%	4%	4%	3%	3%	1%	3%
Verbs	15%	16%	14%	13%	9%	7%	5%	4%	4%	3%	1%	9%
Adjectives	23%	19%	15%	12%	8%	5%	2%	3%	3%	1%	2%	6%

Table 1: Distribution of words per number of senses in the SemCor dataset (words with frequency < 10 were pruned).

address this problem by proposing a pure sense-based model. The model also expands the disambiguation context from a small window (as done in the previous works) to whole sentences. MUSE (Lee & Chen, 2017) is another Skip-gram extension that proposes pure sense representations using reinforcement learning. Thanks to a linear-time sense sequence decoding module, the approach provides a more efficient way of searching for sense combinations.

3.1.3 Contextualized word embeddings

Given that unsupervised sense representations are often produced as a result of clustering, their semantic distinctions are unclear and their mapping to well-defined concepts is not straightforward. In fact, one of the main limitations of these models lies in their difficult integration in downstream models (more details about this in Section 6.5). Recently, an emerging branch of research has focused on directly integrating unsupervised embeddings into downstream models. Word embeddings, such as Word2vec and Glove, compute a single representation for each word, which is used to represent words in downstream models independently from the context in which they appear. In contrast, contextualized word embeddings are sensitive to the context, i.e., their representation dynamically changes depending on the context in which they appear. The sequence tagger of Li and McCallum (2005) is one of the pioneering works that employ contextualized representations. The model infers context sensitive latent variables for each word based on a soft word clustering and integrates them, as additional features, to a CRF sequence tagger.

With the introduction of word embeddings and the efficacy of neural networks, and in the light of meaning conflation deficiency of word embeddings, context-sensitive models have once again attracted research attention. Melamud, Goldberger, and Dagan (2016) represent the context of a target word by extracting the output embedding of a multi-layer perceptron built on top of a bi-directional LSTM language model. This work was one of the basis of this new wave of unsupervised contextualized representations, which led to the integration of contextualized word embedding techniques into downstream applications in later works. Figure 5 provides a high-level illustration of the integration of contextualizing word embeddings into an NLP model. For each word (e.g., *cell* in the figure) in a given input text, the language model unit is responsible for processing the context (usually using recurrent neural networks) and producing a context-sensitive representation. The context-sensitive embeddings are in the internal states of a deep recurrent neural network, which is trained independently from the main task on a large unlabeled text corpus. The input to the main system is usually the concatenation of the two types of representations.

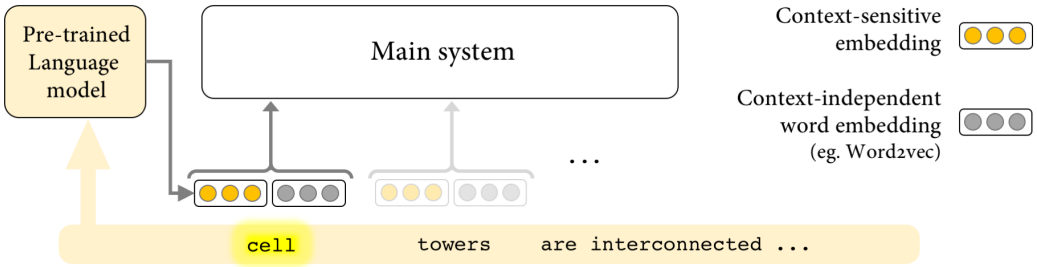


Figure 5: A general illustration of contextualized word embeddings and how they are integrated in NLP models (*Main system* in the figure). A language modelling component is responsible for analyzing the context of the target word (*cell* in the figure) and generating its dynamic embedding. Unlike (context-independent) word embeddings, which have static representations, contextualized embeddings have dynamic representations that are sensitive to the context in which they appear.

The TagLM model of Peters, Ammar, Bhagavatula, and Power (2017) is a recent example of this type, where a multi-layer bidirectional LSTM (Hochreiter & Schmidhuber, 1997) language model is trained on monolingual text. Embeddings from Language Models (Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018, ELMo) are similar in the approach with the exception that some weights are shared between the two directions of the language modeling unit. The assumption in these works is that the language model unit implicitly disambiguates the intended meaning of words according to their contexts.

3.2 Sense Representations Exploiting Multilingual Corpora

Sense distinctions defined by a sense inventory such as WordNet might not be optimal for some downstream applications, such as Machine Translation (MT). Given that ambiguity does not necessarily transfer across languages, sense distinctions for MT should ideally be defined based on the translational differences across a specific language pair. The usual approach to do this is to cluster possible translations of a source word in the target language, with each cluster denoting a specific sense of the source word.

Such translation-specific sense inventories have been used extensively in MT literature (Ide, Erjavec, & Tufis, 2002; Carpuat & Wu, 2007b; Bansal, Denero, & Lin, 2012; Liu, Lu, & Neubig, 2018). The same inventory can be used for the creation of sense embeddings that are suitable for MT. Guo, Che, Wang, and Liu (2014) induced a sense inventory in the same manner by clustering words' translations in parallel corpora. Words in the source language were tagged with their corresponding senses and the automatically annotated data was used to compute sense embeddings using standard word embedding techniques. Ettinger, Resnik, and Carpuat (2016) followed the same sense induction procedure but used the retrofitting-based sense representation of Jauhar, Dyer, and Hovy (2015) [11], by replacing the standard sense inventory used in the original model (WordNet) with a translation-specific inventory.

11. See Section 4.3 for more details about this model.

Similarly, Šuster, Titov, and van Noord (2016) exploited translation distinctions as supervisory signal in an autoencoder for inducing sense representations. At the encoding stage, the discrete-state autoencoder assigns a sense to the target word and at decoding recovers the context given the word and its sense. At the training time, the encoder uses words as well as their translations (from aligned corpora). This bilingual model was extended by Upadhyay, Chang, Zou, Taddy, and Kalai (2017) to a multilingual setting, better benefiting from multilingual distributional information.

4. Knowledge-based Semantic Representations

In addition to unsupervised techniques which only learn from text corpora, there is another branch of research which take advantage of the knowledge available in external resources. This section covers those techniques that exploit knowledge resources for constructing sense and concept representations. First, we will give an overview on currently used knowledge resources (Section 4.1). Then, we will briefly describe some approaches which have made use of knowledge resources for improving word vectors (Section 4.2). Finally, we will focus on the construction of knowledge-based representations of senses (Section 4.3) and concepts or entities (Section 4.4).

4.1 Knowledge Resources

Knowledge resources exist in many flavors. In this section we give an overview of knowledge resources that are mostly used for sense and concept representation learning. The nature of knowledge resources vary with respect to several factors. Knowledge resources can be broadly split into two general categories: expert-made and collaboratively-constructed. Each type has its own advantages and limitations. Expert-made resources (e.g., WordNet) feature accurate lexicographic information such as textual definitions, examples and semantic relations between concepts. On the other hand, collaboratively-constructed resources (e.g., Wikipedia, Wiktionary) provide features such as encyclopedic information, wider coverage and up-to-dateness.

4.1.1 WordNet

A prominent example of expert-made resource is **WordNet** (Miller, 1995), which is one of the most widely used resources in NLP and semantic representation learning. WordNet can be described as a semantic network whose basic units are synsets. A synset represents a concept which may be expressed through nouns, verbs, adjectives or adverbs and is composed of one or more lexicalizations (i.e., synonyms that are used to express the concept). For example, the synset of the concept defined as "the series of vertebrae forming the axis of the skeleton and protecting the spinal cord" comprises six lexicalizations: *spinal column*, *vertebral column*, *spine*, *backbone*, *back*, and *rachis*. Therefore, WordNet is also used as a sense inventory for sense representation learning by following this structure.

Synsets may also be seen as nodes in a semantic network. These nodes are connected to each other by means of lexical or semantic relations such as hypernymy or meronymy. These relations are seen as the edges in the WordNet semantic network. The most recent version of WordNet version (3.1, released on 2012) covers 155,327 words and 117,979 synsets. In

its way to become a multilingual resource, WordNet has also been extended to languages other than English through the Open Multilingual WordNet project (Bond & Foster, 2013) and related efforts.

4.1.2 Wikipedia, Freebase, Wikidata and DBpedia

Wikipedia is one of the most prominent examples of collaboratively-constructed resource. Wikipedia is the largest multilingual encyclopedia of world and linguistic knowledge, with individual pages for millions of concepts and entities in over 250 languages. Its coverage is steadily growing, thanks to continuous updates by collaborators. For instance, the English Wikipedia alone receives approximately 750 new articles per day. Each of these articles provides, for its corresponding concept, a great deal of information in the form of textual information, tables, infoboxes, and various relations such as redirections, disambiguations, and categories.

A similar collaborative effort was **Freebase** (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008). Partly powered by Wikipedia, Freebase was a large collection of structured data, in the form of a knowledge base. As of January 2014, Freebase contained around over 40 million entities and 2 billion relations. Freebase was finally shut down in May 2016 but its information was partially transferred to Wikidata and served in the construction of Google's Knowledge Graph. **Wikidata** (Vrandečić, 2012) is a project operated directly by the Wikimedia Foundation with the goal of turning Wikipedia into a fully structured resource, thereby providing a common source of data that can be used by other Wikimedia projects. It is designed as a document-oriented semantic database based on *items*, each representing a topic and identified by a unique identifier. Knowledge is encoded with *statements* in the form of property-value pairs, among which definitions (descriptions) are also included. **DBpedia** (Bizer, Lehmann, Kobilarov, Auer, Becker, Cyganiak, & Hellmann, 2009) is a similar effort towards structuring the content of Wikipedia. In particular, DBpedia exploits Wikipedia infoboxes, which constitutes its main source of information.

4.1.3 BabelNet and ConceptNet

The types of knowledge available in the expert-based and collaboratively-constructed resources make them often complementary. This has motivated researchers to combine various lexical resources across the two categories (Niemann & Gurevych, 2011; McCrae, Aguado-de Cea, Buitelaar, Cimiano, Declerck, Gómez-Pérez, Gracia, Hollink, Montiel-Ponsoda, Spohr, et al., 2012; Pilehvar & Navigli, 2014). A prominent example is **BabelNet** (Navigli & Ponzetto, 2012), which provides a mapping of WordNet to a number of collaboratively-constructed resources, including Wikipedia. The structure of BabelNet[12] is similar to that of WordNet. Synsets are the main linguistic units and are connected to other semantically related synsets, whose lexicalizations are multilingual in this case. The relations between synsets come from WordNet plus new semantic relations coming from other resources such as Wikipedia hyperlinks and Wikidata. The combination of these resources make BabelNet a large multilingual semantic network, containing 13,789,332 synsets and 354,538,633 relations for 271 languages in its 3.0 release version.

12. `http://babelnet.org/`

ConceptNet (Speer, Chin, & Havasi, 2017) is a similar resource that combines semantic information from heterogeneous sources. In particular, ConceptNet includes relations from resources like WordNet, Wiktionary and DBpedia, as well as common-sense knowledge from crowdsourcing and games with a purpose. The main difference between ConceptNet and BabelNet lies in the main semantic unit, as ConceptNet models words isntead of the synset unit used in BabelNet and other knowledge resources like WordNet.

4.1.4 PPDB: The Paraphrase Database

A different kind of resource is the ParaPhrase DataBase (Ganitkevitch, Van Durme, & Callison-Burch, 2013; Pavlick, Rastogi, Ganitkevitch, Van Durme, & Callison-Burch, 2015, PPDB). PPDB is a lexical resource containing over 150 million paraphrases at different linguistic levels: lexical (single word), phrasal (multiword), and syntactic. In addition to gathering paraphrases, PPDB also has a graph structure where words are viewed as nodes and the edges represent mutual paraphrase connections.

4.2 Knowledge-enhanced Word Representations

As explained in Section 2, word vector representations (e.g., word embeddings) are in the main constructed by exploiting information from text corpora only. However, there is also a line of research which tries to combine the information available in text corpora with the knowledge encoded in lexical resources. This knowledge can be leveraged to include additional information not available in text corpora in order to improve the semantic coherence or coverage of existing word vector representations. Moreover, knowledge-enhanced word representation techniques are closely related to knowledge-based sense representation learning (see next section), as various models make use of similar techniques interchangeably.

The earlier attempts to improve word embeddings using lexical resources modified the objective function of a neural language model for learning word embeddings (e.g., Skip-gram of Word2vec) in order to integrate relations from lexical resources into the learning process (Xu, Bai, Bian, Gao, Wang, Liu, & Liu, 2014; Yu & Dredze, 2014). A more recent class of techniques, usually referred to as retrofitting (Faruqui, Dodge, Jauhar, Dyer, Hovy, & Smith, 2015), attempts at improving pre-trained word embeddings as a post-processing step. Given any pre-trained word embeddings, the main idea of retrofitting is to move closer words which are connected via a relationship in a given semantic network[13]. The main objective function to minimize in the retrofitting model is the following:

$$\sum_{i=1}^{|V|} \Big(\alpha_i \|\vec{w_i} - \vec{\hat{w}_i}\| + \sum_{(w_i,w_j)\in \mathrm{N}} \beta_{i,j} \|\vec{w_i} - \vec{w_j}\| \Big) \tag{2}$$

where $|V|$ represents the size of the vocabulary, N is the input semantic network represented as a set of word pairs, w_i and w_j correspond to word embeddings in the pre-trained model, α_i and $\beta_{i,j}$ are adjustable control values, and $\vec{\hat{w}_i}$ represents the output word embedding.

Building upon retrofitting, Speer and Lowry-Duda (2017) exploited the multilingual relational information of ConcepNet for constructing embeddings on a multilingual space,

13. FrameNet (Baker, Fillmore, & Lowe, 1998), WordNet and PPDB are used in their experiments.

and Lengerich, Maas, and Potts (2017) generalized retrotting methods explicitly modeling pairwise relations. Some similar approaches are Pilehvar and Collier (2017) and Goikoetxea, Soroa, Agirre, and Donostia (2015), which analyze the structure of semantic networks via Personalized Page Rank (Haveliwala, 2002) for extending the coverage and quality of pre-trained word embeddings, respectively. Finally, Bollegala, Alsuhaibani, Maehara, and Kawarabayashi (2016) modified the loss function of a given word embedding model to learn vector representations exploiting cues from both co-occurrences and a semantic network simultaneously.

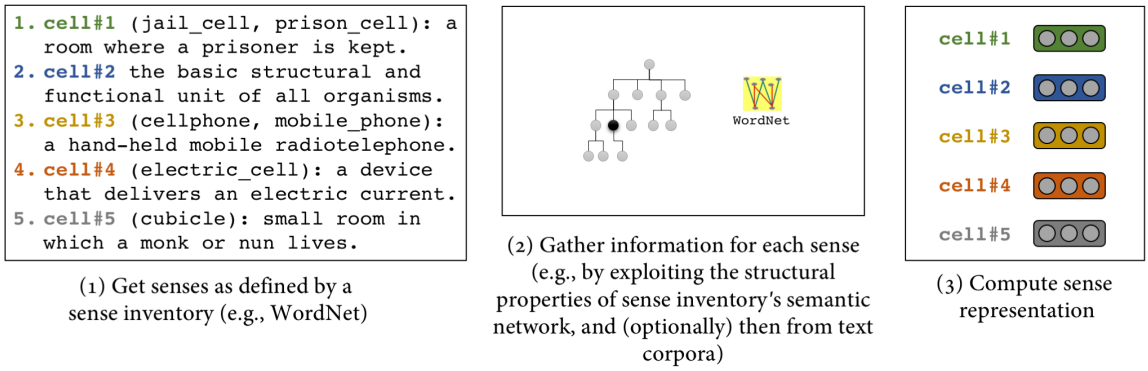
Recently, a new branch that focuses on specializing word embeddings for specific applications have emerged. For instance, Kiela, Hill, and Clark (2015) investigate two variants of retrofitting to specialize word embeddings for similarity or relatedness. Mrksic, Vulić, Séaghdha, Leviant, Reichart, Gai, Korhonen, and Young (2017) improve the quality of bilingual word embeddings exploiting a number of linguistic constraints (e.g., synonymy and antonymy) from resources such as PPDB and BabelNet. The use of these specific linguistic constraints proved effective for monolingual semantic similarity and the dialogue state tracking task.

In fact, knowledge resources play an important role in the construction of multilingual vector spaces. The use of external resources avoids the need of compiling a large parallel corpora, which has been traditionally been the main source for learning cross-lingual word embeddings in the literature (Upadhyay, Faruqui, Dyer, & Roth, 2016; Ruder, Vulić, & Søgaard, 2017). These alternative models for learning cross-lingual embeddings exploit knowledge from lexical resources such as WordNet or BabelNet (Goikoetxea, Soroa, & Agirre, 2018), bilingual dictionaries (Artetxe, Labaka, & Agirre, 2016) or comparable corpora extracted from Wikipedia (Vulić & Moens, 2015).

4.3 Knowledge-based Sense Representations

This section provides an overview of the state of the art in knowledge-based sense representations. These representations are usually obtained as a result of *de-conflating* a word into its individual sense representations, as defined by an external sense inventory. Figure 6 depicts the main workflow for knowledge-based sense vector representation modeling techniques. The learning signal for these techniques vary, but in the main two different types of information available in lexical resources are leveraged: textual definitions (or *glosses*) and semantic networks.

Textual definitions are used as main signals for initializing sense embeddings by several approaches. Chen, Liu, and Sun (2014) proposed an initialization of word sense embeddings by averaging pre-trained word embeddings trained on text corpora. Then, these initialized sense representations are utilized to disambiguate a large corpus. Finally, the training objective of Skip-gram from Word2vec (Mikolov et al., 2013a) is modified in order to learn both word and sense embeddings from the disambiguated corpus. In contrast, Chen, Xu, He, and Wang (2015) exploit a convolutional neural network architecture for initializing sense embeddings using textual definitions from lexical resources. Then, these initialized sense embeddings are fed into a variant of the Multi-sense Skip-gram Model of Neelakantan et al. (2014) (see Section 3.1) for learning knowledge-based sense embedddings. Finally, in



(1) Get senses as defined by a sense inventory (e.g., WordNet)

(2) Gather information for each sense (e.g., by exploiting the structural properties of sense inventory's semantic network, and (optionally) then from text corpora)

(3) Compute sense representation

Figure 6: Knowledge-based sense representation techniques take sense distinctions for a word as defined by an external lexical resource (sense inventory). For each sense, relevant information is gathered and a representation is computed.

Yang and Mao (2016) word sense embeddings are learned by exploiting an adapted Lesk[14] algorithm (Vasilescu, Langlais, & Lapalme, 2004) over short contexts of word pairs.

A different line of research has experimented with the graph structure of lexical resources for learning knowledge-based sense representations. As explained in Section 4.1, some lexical resources can be viewed as a **semantic network** on which nodes are concepts and edges represent the relations among concepts. Semantic networks constitute an scalable alternative to disambiguate large amounts of text (Agirre, de Lacalle, & Soroa, 2014; Moro, Raganato, & Navigli, 2014). Therefore, a straightforward method to learn sense representations would be to disambiguate large amounts of text corpora and apply a word representation learning method on the sense-annotated text (Iacobacci, Pilehvar, & Navigli, 2015). Following this direction, Mancini, Camacho-Collados, Iacobacci, and Navigli (2017) proposed a shallow graph-based disambiguation procedure and modified the objective functions of Word2vec in order to simultaneously learn word and sense embeddings in a shared vector space. The objective function is in essence similar to the objective function proposed by Chen et al. (2014) explained before, which also learns both word and sense embeddings in the last step of their learning process.

Similarly to the post-processing of word embeddings by using knowledge resources (see Section 4.2), recent works have made use of pre-trained word embeddings for not only improving them but also de-conflating them into senses. Approaches that **post-process pre-trained word embeddings** for learning sense embeddings are listed below:

1. A simple method to obtain sense representations from a semantic network is to directly apply the Personalized Page Rank algorithm (Haveliwala, 2002), as in Pilehvar and Navigli (2015). By applying the Personalized Page Rank algorithm in the semantic network, a vector based on the output of the algorithm can be learned for each synset. By following this procedure, Pilehvar and Collier (2016) put forward DeConf,

14. The original Lesk algorithm (Lesk, 1986) and its variants exploit the similarity between textual definitions and a target word's context for disambiguation.

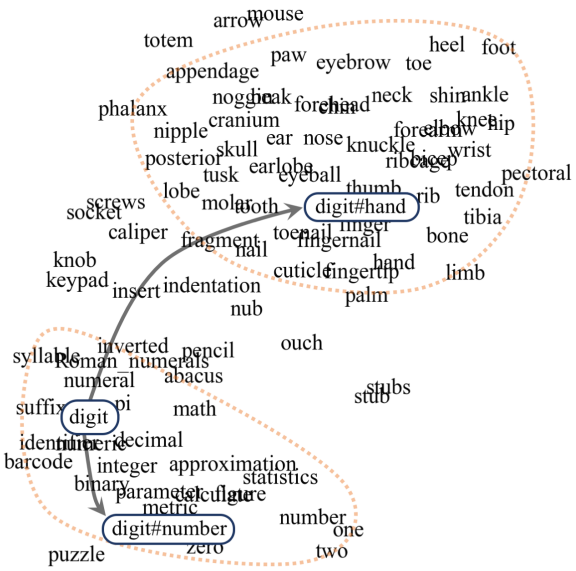


Figure 7: Vector space of words and senses (Pilehvar & Collier, 2016, DeConf). In the figure the word *digit* and its two *mathematical* and *body part* associated senses are displayed in their corresponding regions of the vector space.

a model based on an optimization function so that pre-trained word embeddings can be plugged-in, thereby learning a shared space of word and sense embeddings as a result. By using this method sense embeddings are integrated into the word vector space so they appear near the words which are close in the semantic network (in this case WordNet was used). Figure 7 shows the word *digit* and its *hand* and *number* senses in the vector space.

2. Jauhar et al. (2015) proposed an extension of Retrofitting[15] (Faruqui et al., 2015) for learning representations for the senses of the underlying sense inventory (e.g., WordNet). They additionally presented a second approach which adapt the training objective of Word2vec to include senses within the learning process. The training objective is optimized using EM.

3. Johansson and Pina (2015) post-processed pre-trained word embeddings by means of a optimization problem applying two main constraints: polysemous word embeddings can be decomposed as a combination of its sense embeddings and sense embeddings should be close to their neighbours in the semantic network. In their experiments SALDO (Borin, Forsberg, & Lönngren, 2013), a Swedish semantic network, was used, although their approach may be directly extensible to different semantic networks as well.

4. Finally, Rothe and Schütze (2015) is another method using pre-trained word embeddings as input. In this case, they put forward an autoencoder architecture based on

15. See Section 4.2 for more information on retrofitting.

two main constraints: word vectors correspond to the sum of its senses and synsets to the sum of its lexicalizations (senses). For example, the vector of the word *crane* would correspond to the sum of the vector of its senses $crane_n^1$, $crane_n^2$ and $crane_v^1$ (using WordNet as reference). Similarly, the vector of the synset defined as "arrange for and reserve (something for someone else) in advance" in WordNet would be equal to the sum of the vectors of its corresponding senses *reserve*, *hold* and *book*. Equation 3 displays these constraints mathematically:

$$\vec{w} = \sum_{i=1}^{n} \vec{s_i}; \vec{y} = \sum_{j=1}^{m} \vec{s_j} \tag{3}$$

where s_i and s_j refer to the senses of word w and synset y, respectively.

4.4 Knowledge-based Concept and Entity Representations

In this section we present approaches which rely solely on the relational information of knowledge bases (Section 4.4.1) and hybrid models which combine cues from text corpora and knowledge resources (Section 4.4.2).

4.4.1 Knowledge base embeddings

This section provides a review of those representation techniques targeting concepts and named entities from knowledge bases only. A large body of research in this area takes knowledge graphs (or semantic networks) as signals to construct embedding representations of entities (and relations) specifically targeted to the knowledge completion task.

A pioneer work in this was Bordes, Usunier, Garcia-Duran, Weston, and Yakhnenko (2013), who put forward TransE, a method to embed both entities and relations. In this model relations are viewed as translations which operate in the same vector space of entities. Given a knowledge base represented as a set of triples $\{(e_1, r, e_2)\}$), where e_1 and e_2 are entities and r the relation between them, the main goal is to approach the entities in a way that $\vec{e_1} + \vec{r} \approx \vec{e_2}$ for all triples in the space (i.e., $\forall(e_1, r, e_2) \in N$). Figure 8 illustrates the main idea of the model. This objective may be achieved by exploiting different learning architectures and constraints. In the original work of Bordes et al. (2013) the optimization is carried out by stochastic gradient descent with an L_2 normalization of embeddings as an additional constraint. Following this underlying idea, various approaches have proposed improvements of different parts of the learning architecture:

1. TransP (Wang, Zhang, Feng, & Chen, 2014b) was a similar model that provided improvements on the relational mapping by dealing with specific properties present in the knowledge graph.

2. Lin, Liu, Sun, Liu, and Zhu (2015) proposed to learn embeddings of entities and relations in separate spaces (TransR).

3. Ji, He, Xu, Liu, and Zhao (2015) introduced a dynamic mapping for each entity-relation pair in separated spaces (TransD).

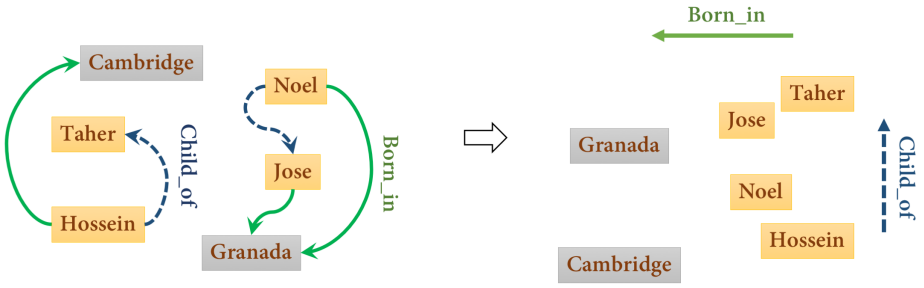


Figure 8: From a knowledge graph to entity and relation embeddings. Illustration based on the slides of Weston and Bordes (2014).

4. Luo, Wang, Wang, and Guo (2015) put forward a two-stage architecture using pre-trained word embeddings for initialization.

5. A unified learning framework that generalize TransE and NTN (Socher, Perelygin, Wu, Chuang, Manning, Ng, & Potts, 2013) was presented in Yang, Yih, He, Gao, and Deng (2015).

6. Finally, Ebisu and Ichise (2018) discussed regularization issues from TransE and proposed TorusE, consisting of a new regularization method solving TransE's regularization problems.

Alternatively, a branch of research focuses specifically on modeling entities only (not relations) and computes embeddings for individual nodes in the graph. DeepWalk (Perozzi, Al-Rfou, & Skiena, 2014) is one of the prominent techniques in this branch. The core idea in this algorithm is to use random graph walks to represent a given graph as a series of artificial sentences. Similarly to natural language in which semantically similar words tend to co-occur, consecutive words in these artificial sentences correspond to neighbouring (topologically related) vertices in the graph. These sentences are then used as input to the Skip-gram model (see Section 2.1.2) and embeddings for individual words (i.e., concept nodes) are computed. Node2vec (Grover & Leskovec, 2016) is an extension of DeepWalk which better controls the depth-first and breadth-first property of random walks. In contrast, Nickel and Kiela (2017) put forward a newer form of representation by embedding words into a Poincaré ball[16] which takes into account both similarity and the hierarchical structure of the taxonomy given as input[17].

These have been some of the most relevant works on knowledge base embeddings in recent years, but given the large amount of literature in this topic, this list is by no means comprehensive. A broader overview of knowledge graph embeddings, including more in-depth explanations, can be found at Cai, Zheng, and Chang (2018) or Nguyen (2017), the latter focusing on the knowledge base completion task.

16. A Poincaré ball is a hyperbolic space in which all points are inside the unit disk.
17. WordNet is used as the reference taxonomy in the original work.

4.4.2 Hybrid models exploiting knowledge bases and text corpora

In addition to these models relying entirely on the information available in knowledge bases, there are approaches that combine cues from knowledge bases and text corpora into the same model. Given its semi-structured nature and the textual content provided, Wikipedia has been the main source for these kinds of approach (Wang, Zhang, Feng, & Chen, 2014a; Sherkat & Milios, 2017; Cao, Huang, Ji, Chen, & Li, 2017). Other approaches have also combined knowledge from heterogeneous resources like Wikipedia and WordNet, exploiting the mapping and multilinguality provided by BabelNet[18] (Camacho-Collados, Pilehvar, & Navigli, 2016).

Given their hybrid nature, these models can be targeted to textual applications as well. A straightforward application is word or named entity disambiguation, for which the embeddings can be used as initialization in the embedding layer on a neural network architecture (Fang, Zhang, Wang, Chen, & Li, 2016; Eshel, Cohen, Radinsky, Markovitch, Yamada, & Levy, 2017) or used directly as a knowledge-based disambiguation system exploiting semantic similarity (Camacho-Collados et al., 2016).

5. Evaluation

In this section we present the most common evaluation benchmarks for assessing the quality of meaning representations. Depending on their nature, evaluation procedures are generally divided in intrinsic (Section 5.1) and extrinsic (Section 5.2).

5.1 Intrinsic Evaluation

Intrinsic evaluation refers to a class of benchmarks that provide a generic evaluation of the quality and coherence of vector space, independently from their performance in downstream applications. Semantic similarity has been viewed as the most straightforward method to evaluate sense representations. In particular, the semantic similarity of small lexical units such as words and senses in which compositionality is not required have taken the most attention.

Word similarity datasets exist in many flavors. It is also worth distinguishing the notions of similarity and relatedness. While words that are similar are prone to be replaced in similar contexts, related words only needs to co-occur in the same documents. WordSim-353 (Finkelstein, Evgeniy, Yossi, Ehud, Zach, Gadi, & Eytan, 2002) is a dataset that conflates these two notions. Genuine similarity datasets include RG-65 (Rubenstein & Goodenough, 1965), which only contains 65 word pairs, or SimLex-999 (Hill, Reichart, & Korhonen, 2015), consisting of 999 word pairs. Moreover, there are multilingual benchmarks which include word similarity datasets for several languages. For instance, the translations and reannotations of WordSim-353 and SimLex-999 (Leviant & Reichart, 2015) and the datasets from the SemEval-2017 task on multilingual word similarity (Camacho-Collados, Pilehvar, Collier, & Navigli, 2017) provide evaluation datasets for languages other than English as well.

In order to adapt these word-based evaluation benchmarks to sense vectors, various strategies have been proposed (Reisinger & Mooney, 2010). Among these, the most popular

18. See Section 4.1.3 for more information about BabelNet.

is to take the most similar pair of senses across the two words, also known as *MaxSim* (Resnik, 1995; Pilehvar & Navigli, 2015):

$$sim(w_1, w_2) = \max_{s \in S_{w_1}, s' \in S_{w_2}} \cos(\vec{s}_1, \vec{s}_2) \quad (4)$$

where S_{w_i} is a set including all senses of w_i and $\vec{s}_i$ represents the sense vector representation of the sense s_i. Another strategy, known as AvgSim, simply averages the pairwise similarities of all possible senses of w_1 and w_2. Cosine similarity (cos) is the most prominent metric for computing the similarity between sense vectors.

In all these benchmarks, words are paired in isolation. However, we know that for a specific meaning of an ambiguous word to be triggered, the word needs to appear in a context. In fact, Kilgarriff (1997) argued that representing a word with a fixed set of senses may not be the best way for modelling word senses but instead, word senses should be defined according to a given context. To this end, Huang et al. (2012) presented a different kind of similarity dataset in which words are provided with their corresponding contexts. The task consists of assessing the similarity of two words by taking the contexts in which they occur into consideration. The dataset is known as Stanford Contextual Word Similarity (SCWS) and has been established as one of the main intrinsic evaluations for sense representations. Exploiting the appropriate sense given a context has proved important in this task, a property that sense representations are well suited to as proved by the fact that they generally outperform word-based models in this dataset. However, a pre-disambiguation step is required to leverage sense representations in this task. Simple similarity measures such as MaxSimC or AvgSimC are generally utilized. Unlike MaxSim and AvgSim, MaxSimC and AvgSimC take the context of the target word into account. First, the confidence for selecting the most appropriate sense within the sentence is computed (e.g., by computing the average of word embeddings from the context and selecting the sense which is closest to the average context vector in terms of cosine similarity). Then, the final score corresponds to the similarity between the selected senses (i.e., MaxSimC) or to a weighted average among all senses (i.e., AvgSimC).

Finally, in addition to semantic similarity, there are other intrinsic evaluation procedures such as synonymy selection (Landauer & Dumais, 1997; Turney, 2001; Jarmasz & Szpakowicz, 2003; Reisinger & Mooney, 2010), outlier detection (Camacho-Collados & Navigli, 2016; Blair, Merhav, & Barry, 2016) or sense clustering (Snow, Prakash, Jurafsky, & Ng, 2007; Dandala, Hokamp, Mihalcea, & Bunescu, 2013). For more information, Bakarov (2018) provides a more comprehensive overview of intrinsic evaluation benchmarks.

5.2 Extrinsic Evaluation

Extrinsic evaluation procedures aim at assessing the quality of meaning representations within a downstream task. In addition to intrinsic evaluation procedures, extrinsic evaluation is necessary to understand the effectiveness of different sense representation techniques in real-word applications. This is especially relevant because intrinsic evaluation protocols do not always correlate with downstream performance (Tsvetkov, Faruqui, Ling, Lample, & Dyer, 2015; Chiu, Korhonen, & Pyysalo, 2016). However, while extrinsic evaluation is definitely important to assess the effectiveness of integrating sense representations in down-

stream tasks, there is also a higher variability in terms of tasks, pipelines and benchmarks in comparison to intrinsic procedures, which are more straightforward.

Some of the most common tasks that have been used as extrinsic evaluation procedures for sense representations in natural language processing are text categorization and sentiment analysis (Liu et al., 2015b; Li & Jurafsky, 2015; Pilehvar, Camacho-Collados, Navigli, & Collier, 2017), document similarity (Wu & Giles, 2015), and word sense induction (Pelevina et al., 2016; Panchenko, Ruppert, Faralli, Ponzetto, & Biemann, 2017b) and disambiguation (Chen et al., 2014; Rothe & Schütze, 2015; Camacho-Collados et al., 2016; Peters et al., 2018). As mentioned in Section 4.4.1, knowledge base embeddings are also frequently evaluated on the knowledge base completion task (Bordes et al., 2013). In Section 6.5 we will explain in more detail some of the applications to which sense representations have been applied to date.

6. Analysis

This section provides an analysis and comparison of knowledge-based and unsupervised representation techniques, highlighting the advantages and limitations of each, while suggesting the applications and scenarios for which each technique is suited. We focus on five important aspects: interpretability (Section 6.1), adaptability to different domains (Section 6.2), sense granularity (Section 6.3), compositionality (Section 6.4) and integration into downstream applications (Section 6.5).

6.1 Interpretability

One of the main reasons behind moving from word to sense level is the semantically-grounded nature of word senses, which may enable a better interpretability. In this particular aspect, however, there is a considerable difference between unsupervised and knowledge-based models. Unsupervised models learn senses directly from sense corpora, which results in model-specific sense interpretations. In many cases, these induced senses do not correspond to human notions of senses, or are not easily distinguishable. On this respect, Panchenko et al. (2017b) worked on improving the interpretability of unsupervised sense representations by extracting their hypernyms and images corresponding to the specific meanings.

In contrast, knowledge-based representations are already linked to entries in a sense inventory, which enables a higher interpretability, as these entries are generally associated with definitions, examples, images and often relations with other concepts (e.g., WordNet) and translations (e.g., BabelNet). This, in turn, enables the direct injection of extra prior information from lexical resources, which may be useful to supply end models with a deeper background knowledge (Young, Kunze, Basile, Cabrio, Hawes, & Caputo, 2017). As a drawback, knowledge-based representations are generally constrained to the underlying sense inventories and, hence, may fail to provide an accurate representation of unseen novel senses in text corpora. This is partially solved by keeping sense inventories updated, thought not generally a straightforward process. As explained in Section 4.1, collaborative resources like Wikipedia are less prone by this issue.

6.2 Adaptability to Different Domains

One feature which has been praised in word embeddings is their adaptability to general and specialized domains (Goldberg, 2016). In this case, unsupervised models are more similar to these word embedding techniques, as they learn directly from text corpora. This provides them with a theoretical advantage with respect to knowledge-based models, as only a text corpus from a given domain is needed to learn the senses belonging to a particular domain. From this perspective, knowledge-based systems generally learn representations for all senses from a given sense inventory, and they are not specialized to a given domain.

Knowledge-enhanced approaches like those proposed by Mancini et al. (2017) or Fang et al. (2016), which directly learn from text corpora, may partially alleviate this limitation of knowledge-based models. However, the senses should still appear in the semantic network as input to the model. In other words, knowledge-based approaches are not able to learn new senses, which may be an important limitation in some specific domains and tasks. Moreover, the accurate representation of certain domains would require suitable knowledge resources, which might not be available for specialized domains or low-resource languages.

6.3 Sense Granularity

A sense inventory may list a few dozen different senses for words such as *run*, *play* and *get*. Words with multiple senses (i.e., ambiguous) are generally classified into two categories: polysems and homonyms. Polysemous words have multiple related meanings. For instance the word *mark* can refer to a "distinguishing symbol" as well as a "visible indication made on a surface". In this case the distinctions of these two senses are also said to be fine-grained, as these two meanings are difficult to be torn apart. Homonymous words[19] have meanings that are completely unrelated. For instance, the geological and financial institution senses of the word *bank*.[20] This would also be a case of a coarse-grained distinction of senses, as these two meaning of *bank* are clearly different.

In general, the fine granularity of certain sense inventories has always been a point of argument in NLP (Kilgarriff, 1997; Navigli, 2009; Hovy, Navigli, & Ponzetto, 2013). It has been pointed out that sense distinctions in WordNet may be too fine-grained to be useful for many NLP applications (Navigli, 2006; Snow et al., 2007; Hovy et al., 2013). For instance, WordNet 3.0 (see Section 4.1.1) lists 41 different senses for the verb *run*. However, most of these senses are translated to either *correr* or *operar* in Spanish. Therefore, a cross-lingual task such as Machine Translation might not benefit from the additional distinctions provided by the sense inventory. In fact, a merging of these fine-grained distinctions into more coarse-grained classes (referred to as *supersenses* in WordNet) has been shown to be beneficial in downstream applications (Flekova & Gurevych, 2016; Pilehvar et al., 2017).

This discussion is also relevant for unsupervised techniques. The dynamic learning of senses, instead of fixing the number of senses for all words, have shown to provide a more

19. According to the Cambridge Dictionary, a homonym is "a word that sounds the same (homophone) or is spelled the same (homograph) as another word but has a different meaning". Given that NLP focuses on written forms, a homonym in this context usually refers to the latter condition, i.e., homographs with different meanings.

20. The distinction between homonyms and polysems can sometime be subtle. For instance, research in historical linguistics has shown that the two meanings of the word *bank* could have been related to each other earlier in the Italian language, since the bankers used to do their business on the riverbanks.

realistic distribution of senses (see Section 3.1.2). Moreover, there has been discussions about whether all occurrences of words can be effectively partitioned into senses (Kilgarriff, 1997; Hanks, 2000; Kilgarriff, 2007), which has led to a new scheme on which meanings of a word are described on a graded fashion (Erk et al., 2009; McCarthy et al., 2016). While this scheme is not covered in this survey, it has been shown that a graded scale to assess senses may correlate better on how humans perceive different meanings. Although not exactly the same conclusions, these findings are also related to the criticisms about the fine granularity of current sense inventories, which has in turn shown to be harmful in certain downstream applications.

6.4 Compositionality

Compositional methods model the semantics of a complex expression based on the meanings of its constituent (e.g., words). Typically, constituent words are represented as their word vector with all the meanings conflated. However, for an ambiguous word in an expression, usually only a single meaning is triggered and other senses are irrelevant. Therefore, pinpointing the meaning of a word to the given context may be a reasonable idea for compositionality. This can be crucial to applications such as information retrieval in which query ambiguity can be an issue (Allan & Raghavan, 2002; Di Marco & Navigli, 2013).

Different works have tried to introduce sense representations in the context of compositionality (Köper & im Walde, 2017; Kober, Weeds, Wilkie, Reffin, & Weir, 2017), with different degrees of success. The main idea is to select the intended sense of a word and only introduce that specific meaning into the composition, either through context-based sense induction (Thater, Fürstenau, & Pinkal, 2011), exemplar-based representation (Reddy, Klapaftis, McCarthy, & Manandhar, 2011), or with the help of external resources, such as WordNet (Gamallo & Pereira-Fariña, 2017). Cheng and Kartsaklis (2015) also proposed a recurrent neural network in which word embeddings were split into multiple sense vectors. The network was applied to paraphrase detection with positive results. However, sense distinction in the context of compositionality has often been evaluated on generic benchmarks, such as paraphrase detection. Despite the potential benefit in tasks such as question answering and information retrieval, there have been no attempts at integrating sense representations as components of neural compositional models.

6.5 Integration into Downstream Applications

As mentioned in Section 5, the ideal goal for sense representations is to be effectively integrated into downstream applications. Unlike word representations (and more specifically embeddings), sense representations are still in their infancy in this regard. This is also due to the non-immediate integration of these representations, which generally require an additional word sense disambiguation or induction step. However, as with word embeddings, sense representations can be theoretically applied to multiple applications. **Information retrieval** has been one of the first applications investigated (Schütze & Pedersen, 1995). The authors showed how document-query similarity based on word senses could lead to considerable improvements with respect to word-based models.

Since then, word senses (and in particular sense representations) have been integrated into various NLP tasks, with varying degree of success. Li and Jurafsky (2015) proposed

a framework to integrate unsupervised sense embeddings into various natural language processing tasks. The research concluded that these unsupervised representations did not provide a significant influence. They suggested that an increase in the dimensionality of word embeddings can lead to similar result. However, the disambiguation step was a simple procedure based on the similarity between sense embedddings and a embedding representation of the input text (computed as the average of the content words' embedddings). In order to avoid this disambiguation step, Peters et al. (2018) put forward contextualized word representations (see Section 3.1.3) which can be seamlessly integrated into neural architectures. Their work showed how significant improvements can be gained by integrating these contextualized representations into downstream task such as **question answering**, **textual entailment** and **sentiment analysis**.

As far as knowledge-based representations are concerned, an explicit or implicit word sense disambiguation step is required to transform words into their intended senses. Pilehvar et al. (2017) proposed a method based on a shared space of word and knowledge-based sense embeddings, introducing a simple graph-based disambiguation step prior to their integration into a neural network architecture for **text classification**. The inclusion of senses was shown to improve when the input text was large enough, but the inclusion of pre-trained sense embedddings in this setting did not significantly improve the use of word embeddings in most datasets. The major benefits of using sense representations were observed when using supersenses (see Section 6.3), a conclusion which was also observed by Flekova and Gurevych (2016) on other downstream classification tasks.

Machine Translation (MT) has been a longstanding field in NLP which has had several attempts at exploiting sense representations. Since a word may have different translations depending on its sense in context, sense identification has been traditionally believed to be able to potentially improve word-based MT models. Carpuat et al. analyzed the impact of WSD in the performance of standard MT systems at the time (Carpuat & Wu, 2005, 2007a, 2007b). The studies were inconclusive, but generally reflected the difficulty to successfully integrate semantically-grounded models into an MT pipeline. This was also partially due to the lack of sense-annotated corpora, producing the so-called knowledge-acquisition bottleneck (Gale et al., 1992). A few recent efforts have been proposed to directly address this issue by disambiguating large amounts of parallel corpora (Taghipour & Ng, 2015; Otegi, Aranberri, Branco, Hajic, Neale, Osenova, Pereira, Popel, Silva, Simov, & Agirre, 2016; Delli Bovi, Camacho-Collados, Raganato, & Navigli, 2017). Alternatively, Liu et al. (2018) recently proposed an end-to-end neural architecture which explicitly models homographs (i.e., ambiguous words) with context-aware embeddings, showing overall gains and improvements of the translation of such ambiguous words.

In addition to these tasks, there have been other applications on which sense and concept representations, in their broader meaning, have been effectively integrated: **word sense or named entity disambiguation** (Chen et al., 2014; Rothe & Schütze, 2015; Camacho-Collados et al., 2016; Fang et al., 2016; Panchenko, Faralli, Ponzetto, & Biemann, 2017a; Peters et al., 2018), **knowledge base completion** (Bordes et al., 2013) or **unification** (Delli Bovi, Espinosa-Anke, & Navigli, 2015), **common-sense reasoning** (Lieto, Radicioni, Rho, & Mensa, 2017), **lexical substitution** (Cocos, Apidianaki, & Callison-Burch, 2016), **hypernym discovery** (Espinosa-Anke, Camacho-Collados, Delli Bovi, & Saggion,

2016), **lexical entailment** (Nickel & Kiela, 2017), or **visual object discovery** (Young et al., 2017).

7. Conclusions

In this survey we have presented an extensive overview of semantically-grounded models for constructing distributed representations of meaning. Word embeddings have been shown to provide interesting semantic properties that can be applied to most language applications. However, these models tend to conflate different meanings into a single representation. Therefore, an accurate distinction of senses is often required for a deep understanding of lexical meaning. To this end, we provided a comprehensive survey of models that learn representation for senses which are either directly induced from text corpora (i.e., unsupervised) or defined by external sense inventories (i.e., knowledge-based).

Some of these models have already proven effective in practise, but there is still much room for improvement. For example, even though semantically-grounded information is captured (to different degrees) by almost all models, common-sense reasoning has not yet been deeply explored. Also, most of these models have been tested on English only, whereas only a few have proposed models for other languages or attempted multilinguality. Finally, the integration of these theoretical models into downstream applications is the next step forward, as it is not clear now what the best integration strategy would be, and if a pre-disambiguation step is necessary. For instance, approaches such as the contextualised embeddings of Peters et al. (2018) have shown a new possible direction in which senses are learned dynamically for each context, without the need for an explicit pre-disambiguation step.

Although not exactly distributed representations of meaning, modelling relations in a flexible way is also another possible avenue for future work. Relations are generally modeled in works targeting knowledge-based completion. Moreover, a recent line of research is experimenting with embedding these relations exploiting text corpora as well (Toutanova, Chen, Pantel, Poon, Choudhury, & Gamon, 2015; Jameel, Bouraoui, & Schockaert, 2018), which paves the way for new approaches integrating these relations into downstream text applications.

From this perspective, the definition of sense and the correct paradigm is certainly still an open question. Do senses need to be discrete? Should they need to be tied to a knowledge resource or sense inventory? Should they be learned dynamically depending on the context? These are the questions that are yet to be explored according to the many studies on this topic. As also explained in our analysis, some approaches are more suited to certain applications or domains, without any clear general conclusion. These open questions are certainly still relevant and encourage further research on distributed representations of meaning, with many areas yet to be explored.

References

Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, *40*(1), 57–84.

Allan, J., & Raghavan, H. (2002). Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the* 25^{th} *annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 307–314, Tampere, Finland.

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294.

Athiwaratkun, B., & Wilson, A. (2017). Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1645–1656.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 86–90. Association for Computational Linguistics.

Bansal, M., Denero, J., & Lin, D. (2012). Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pp. 773–782, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bartunov, S., Kondrashkin, D., Osokin, A., & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Vol. 51 of *Proceedings of Machine Learning Research*, pp. 130–138, Cadiz, Spain. PMLR.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, *3*, 1137–1155.

Bennett, A., Baldwin, T., Lau, J. H., McCarthy, D., & Bond, F. (2016). Lexsemtm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Proceedings of ACL*, pp. 1513–1524.

Biemann, C. (2006). Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pp. 73–80. Association for Computational Linguistics.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, *7*(3), 154–165.

Blair, P., Merhav, Y., & Barry, J. (2016). Automated generation of multilingual clusters for the evaluation of distributed representations. *arXiv preprint arXiv:1611.01547*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, *5*(1), 135–146.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250. ACM.

Bollegala, D., Alsuhaibani, M., Maehara, T., & Kawarabayashi, K.-i. (2016). Joint word representation learning using a corpus and a semantic lexicon. In *AAAI*, pp. 2690–2696.

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet.. In *ACL (1)*, pp. 1352–1362.

Bordes, A., Chopra, S., & Weston, J. (2014). Question answering with subgraph embeddings. In *EMNLP*.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pp. 2787–2795.

Borin, L., Forsberg, M., & Lönngren, L. (2013). Saldo: a touch of yin to wordnets yang. *Language resources and evaluation*, *47*(4), 1191–1211.

Cai, H., Zheng, V. W., & Chang, K. (2018). A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*.

Camacho-Collados, J., & Navigli, R. (2016). Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 43–50.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 15–26.

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, *240*, 36–64.

Cao, Y., Huang, L., Ji, H., Chen, X., & Li, J. (2017). Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1623–1633.

Carpuat, M., & Wu, D. (2005). Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 387–394. Association for Computational Linguistics.

Carpuat, M., & Wu, D. (2007a). How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. *Proceedings of TMI*, 43–52.

Carpuat, M., & Wu, D. (2007b). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Chen, T., Xu, R., He, Y., & Wang, X. (2015). Improving distributed representation of word sense via wordnet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing – Short Papers*, pp. 15–20, Beijing, China.

Chen, X., Liu, Z., & Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pp. 1025–1035, Doha, Qatar.

Cheng, J., & Kartsaklis, D. (2015). Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1531–1542. Association for Computational Linguistics.

Chiu, B., Korhonen, A., & Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, Berlin, Germany.

Cocos, A., Apidianaki, M., & Callison-Burch (2016). Word sense filtering improves embedding-based lexical substitution. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 99–104. Association for Computational Linguistics.

Dandala, B., Hokamp, C., Mihalcea, R., & Bunescu, R. C. (2013). Sense clustering using Wikipedia.. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 164–171, Hissar, Bulgaria.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., & Navigli, R. (2017). EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of ACL*, Vol. 2, pp. 594–600.

Delli Bovi, C., Espinosa-Anke, L., & Navigli, R. (2015). Knowledge base unification via sense embeddings and disambiguation. In *Proceedings of EMNLP*, pp. 726–736. Association for Computational Linguistics.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Di Marco, A., & Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, *39*(3), 709–754.

Ebisu, T., & Ichise, R. (2018). Toruse: Knowledge graph embedding on a lie group. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of ACL*, Prague, Czech Republic.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Erk, K., McCarthy, D., & Gaylord, N. (2009). Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the*

ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 10–18. Association for Computational Linguistics.

Erk, K., & Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 897–906.

Eshel, Y., Cohen, N., Radinsky, K., Markovitch, S., Yamada, I., & Levy, O. (2017). Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 58–68.

Espinosa-Anke, L., Camacho-Collados, J., Delli Bovi, C., & Saggion, H. (2016). Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of EMNLP*, pp. 424–435.

Ettinger, A., Resnik, P., & Carpuat, M. (2016). Retrofitting sense-specific word vectors using parallel text. In *Proceedings of NAACL-HLT*, pp. 1378–1383, San Diego, California.

Fang, W., Zhang, J., Wang, D., Chen, Z., & Li, M. (2016). Entity disambiguation by knowledge and text jointly embedding. In *Proceedings of CoNLL*, pp. 260–269.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, pp. 1606–1615.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Finkelstein, L., Evgeniy, G., Yossi, M., Ehud, R., Zach, S., Gadi, W., & Eytan, R. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

Flekova, L., & Gurevych, I. (2016). Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of ACL*.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pp. 1606–1611, Hyderabad, India.

Gale, W. A., Church, K., & Yarowsky, D. (1992). A method for disambiguating word senses in a corpus. *Computers and the Humanities*, *26*, 415–439.

Gamallo, P., & Pereira-Fariña, M. (2017). Compositional semantics using feature-based models from wordnet. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 1–11, Valencia, Spain. Association for Computational Linguistics.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). Ppdb: The paraphrase database.. In *Proceedings of NAACL-HLT*, pp. 758–764.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Goikoetxea, J., Soroa, A., & Agirre, E. (2018). Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*.

Goikoetxea, J., Soroa, A., Agirre, E., & Donostia, B. C. (2015). Random walks and neural network language models on knowledge bases. In *Proceedings of NAACL*, pp. 1434–1439.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345–420.

Grover, A., & Leskovec, J. (2016). Node2Vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 855–864, New York, NY, USA.

Guo, J., Che, W., Wang, H., & Liu, T. (2014). Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*, pp. 497–507.

Hanks, P. (2000). Do word meanings exist?. *Computers and the Humanities*, *34*(1-2), 205–215.

Harris, Z. (1954). Distributional structure. *Word*, *10*, 146–162.

Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, pp. 517–526, Hawaii, USA.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Compasutation*, *9*(8), 1735–1780.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1), 177–196.

Hovy, E. H., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, *194*, 2–27.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pp. 873–882, Jeju Island, Korea.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pp. 95–105, Beijing, China.

Ide, N., Erjavec, T., & Tufis, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of ACL-02 Workshop on WSD: Recent Successes and Future Directions*, pp. 54–60, Philadelphia, USA.

Jameel, S., Bouraoui, Z., & Schockaert, S. (2018). Unsupervised learning of distributional relation vectors. In *Proceedings of ACL*, Melbourne, Australia.

Jarmasz, M., & Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 212–219, Borovets, Bulgaria.

Jauhar, S. K., Dyer, C., & Hovy, E. (2015). Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL*, pp. 683–693, Denver, Colorado.

Ji, G., He, S., Xu, L., Liu, K., & Zhao, J. (2015). Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, pp. 687–696.

Johansson, R., & Pina, L. N. (2015). Embedding a semantic network in a word space. In *Proceedings of NAACL*, pp. 1428–1433, Denver, Colorado.

Jones, M. P., & Martin, J. H. (1997). Contextual spelling correction using latent semantic analysis. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pp. 166–173.

Kiela, D., Hill, F., & Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2044–2048.

Kilgarriff, A. (1997). "i don't believe in word senses".. *Computers and the Humanities*, *31*(2), 91–113.

Kilgarriff, A. (2007). Word senses. In *Word Sense Disambiguation*, pp. 29–46. Springer.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pp. 1746–1751, Doha, Qatar.

Kober, T., Weeds, J., Wilkie, J., Reffin, J., & Weir, D. (2017). One representation per word - does it make sense for composition?. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 79–90, Valencia, Spain. Association for Computational Linguistics.

Köper, M., & im Walde, S. S. (2017). Applying multi-sense embeddings for german verbs to determine semantic relatedness and to detect non-literal language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2, pp. 535–542.

Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *SIGMOD Rec.*, *31*(2), 84–93.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.. *Psychological Review*, *104*(2), 211.

Landauer, T., & Dooley, S. (2002). Latent semantic analysis: theory, method and application. In *Proceedings of CSCL*, pp. 742–743.

Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *IEEE software*, *14*(2), 67–75.

Lee, G.-H., & Chen, Y.-N. (2017). Muse: Modularizing unsupervised sense embeddings. In *Proceedings of EMNLP*, Copenhagen, Denmark.

Lengerich, B. J., Maas, A. L., & Potts, C. (2017). Retrofitting distributional embeddings to knowledge graphs with functional relations. *arXiv preprint arXiv:1708.00112*.

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 Conference on Systems Documentation,* Toronto, Ontario, Canada, pp. 24–26.

Leviant, I., & Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pp. 2177–2185.

Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding?. In *Proceedings of EMNLP*, pp. 683–693, Lisbon, Portugal.

Li, W., & McCallum, A. (2005). Semi-supervised sequence modeling with syntactic topic models. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, pp. 813–818. AAAI Press.

Lieto, A., Radicioni, D., Rho, V., & Mensa, E. (2017). Towards a unifying framework for conceptual represention and reasoning in cognitive systems. *Intelligenza Artificiale*, *11*(2), 139–153.

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion.. In *Proceedings of AAAI*, pp. 2181–2187.

Liu, F., Lu, H., & Neubig, G. (2018). Handling homographs in neural machine translation. In *Proceedings of NAACL*, New Orleans, LA, USA.

Liu, P., Qiu, X., & Huang, X. (2015a). Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 1284–1290.

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015b). Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2418–2424.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Luo, Y., Wang, Q., Wang, B., & Guo, L. (2015). Context-dependent knowledge graph embedding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1656–1661.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mancini, M., Camacho-Collados, J., Iacobacci, I., & Navigli, R. (2017). Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of CoNLL*, pp. 100–111, Vancouver, Canada.

McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability. *Computational Linguistics*.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, *46*(4), 701–719.

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61, Berlin, Germany.

Meyerson, A. (2001). Online facility location. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pp. 426–432, Washington, DC, USA. IEEE Computer Society.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations.. In *HLT-NAACL*, pp. 746–751.

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

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. (1993). A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pp. 303–308, Plainsboro, N.J.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, *2*, 231–244.

Mrksic, N., Vulić, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gai, M., Korhonen, A., & Young, S. (2017). Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *TACL*.

Navigli, R. (2006). Meaningful clustering of senses helps boost Word Sense Disambiguation performance. In *Proceedings of COLING-ACL*, pp. 105–112, Sydney, Australia.

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

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

Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, pp. 1059–1069, Doha, Qatar.

Nguyen, D. Q., Nguyen, D. Q., Modi, A., Thater, S., & Pinkal, M. (2017). A mixture model for learning multi-sense word embeddings. In *Proceedings of *SEM 2017*.

Nguyen, D. Q. (2017). An overview of embedding models of entities and relationships for knowledge base completion. *arXiv preprint arXiv:1703.08098*.

Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 6341–6350. Curran Associates, Inc.

Niemann, E., & Gurevych, I. (2011). The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and Wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 205–214.

Nieto Piña, L., & Johansson, R. (2015). A simple and efficient method to generate word sense representations. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 465–472, Hissar, Bulgaria.

Otegi, A., Aranberri, N., Branco, A., Hajic, J., Neale, S., Osenova, P., Pereira, R., Popel, M., Silva, J., Simov, K., & Agirre, E. (2016). QTLeap WSD/NED Corpora: Semantic Annotation of Parallel Corpora in Six Languages. In *Proc. of LREC*, pp. 3023–3030.

Panchenko, A., Faralli, S., Ponzetto, S. P., & Biemann, C. (2017a). Using linked disambiguated distributional networks for word sense disambiguation. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 72–78.

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., & Biemann, C. (2017b). Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation. In *Proceedings of EACL*, pp. 86–98.

Pasini, T., & Navigli, R. (2018). Two knowledge-based methods for high-performance sense distribution learning. In *Proceedings of AAAI*, New Orleans, United States.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015). Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL(2)*, pp. 425–430, Beijing, China. Association for Computational Linguistics.

Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2016). Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 174–183.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pp. 1532–1543.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 701–710.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL*, New Orleans, LA, USA.

Peters, M., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1756–1765. Association for Computational Linguistics.

Piantadosi, S. T. (2014). Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.

Pilehvar, M. T., Camacho-Collados, J., Navigli, R., & Collier, N. (2017). Towards a Seamless Integration of Word Senses into Downstream NLP Applications. In *Proceedings of ACL*, Vancouver, Canada.

Pilehvar, M. T., & Collier, N. (2016). De-conflated semantic representations. In *Proceedings of EMNLP*, pp. 1680–1690, Austin, TX.

Pilehvar, M. T., & Collier, N. (2017). Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 388–393, Valencia, Spain. Association for Computational Linguistics.

Pilehvar, M. T., & Navigli, R. (2014). A robust approach to aligning heterogeneous lexical resources. In *Proceedings of ACL*, pp. 468–478.

Pilehvar, M. T., & Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, *228*, 95–128.

Qiu, L., Tu, K., & Yu, Y. (2016). Context-dependent sense embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 183–191.

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 337–346.

Reddy, S., Klapaftis, I. P., McCarthy, D., & Manandhar, S. (2011). Dynamic and static prototype vectors for semantic composition. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pp. 705–713.

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Proceedings of ACL*, pp. 109–117.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pp. 448–453.

Rothe, S., & Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*, pp. 1793–1803, Beijing, China.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Ruder, S. (2017). On word embeddings, part 1. *URL: http://ruder.io/word-embeddings-2017/(visited on 1/04/2018)*.

Ruder, S., Vulić, I., & Søgaard, A. (2017). A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, *24*(1), 97–123.

Schütze, H., & Pedersen, J. (1995). Information retrieval based on word senses. In *Proceedings of SDAIR'95*, pp. 161–175, Las Vegas, Nevada.

Sherkat, E., & Milios, E. E. (2017). Vector embedding of wikipedia concepts and entities. In *International Conference on Applications of Natural Language to Information Systems*, pp. 418–428. Springer.

Snow, R., Prakash, S., Jurafsky, D., & Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of EMNLP*, pp. 1005–1014, Prague, Czech Republic.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Parsing with compositional vector grammars. In *Proceedings of EMNLP*, pp. 455–465, Sofia, Bulgaria.

Soucy, P., & Mineau, G. W. (2005). Beyond tfidf weighting for text categorization in the vector space model. In *Proceedings of IJCAI*, Vol. 5, pp. 1130–1135.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4444–4451.

Speer, R., & Lowry-Duda, J. (2017). Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 85–89. Association for Computational Linguistics.

Šuster, S., Titov, I., & van Noord, G. (2016). Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of NAACL-HLT*, pp. 1346–1356, San Diego, California.

Taghipour, K., & Ng, H. T. (2015). One million sense-tagged instances for word sense disambiguation and induction. *CoNLL 2015*, 338.

Thater, S., Fürstenau, H., & Pinkal, M. (2011). Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 1134–1143, Chiang Mai, Thailand.

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., & Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pp. 151–160.

Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP (2)*, pp. 2049–2054, Lisbon, Portugal.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pp. 491–502.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pp. 417–424.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality.. *Psychological Review*, *89*(2), 123.

Upadhyay, S., Chang, K.-W., Zou, J., Taddy, M., & Kalai, A. (2017). Beyond bilingual: Multi-sense word embeddings using multilingual context. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada.

Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1661–1670.

Van de Cruys, T., Poibeau, T., & Korhonen, A. (2011). Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1012–1022, Edinburgh, Scotland, UK.

Vasilescu, F., Langlais, P., & Lapalme, G. (2004). Evaluating variants of the lesk approach for disambiguating words. In *Proceddings of LREC*.

Vilnis, L., & McCallum, A. (2015). Word representations via gaussian embedding. In *Proceedings of ICLR*.

Vrandečić, D. (2012). Wikidata: A New Platform for Collaborative Data Collection. In *Proceedings of WWW*, pp. 1063–1064.

Vu, T., & Parker, D. S. (2016). K-embeddings: Learning conceptual embeddings for words using context. In *Proceedings of NAACL-HLT*, pp. 1262–1267.

Vulić, I., & Moens, M.-F. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2, pp. 719–725.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014a). Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1591–1601.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014b). Knowledge graph embedding by translating on hyperplanes. In *Proceedings of AAAI*, pp. 1112–1119.

Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured training for neural network transition-based parsing. In *Proceedings of ACL*, pp. 323–333, Beijing, China.

Weston, J., & Bordes, A. (2014). Embedding methods for nlp. In *EMNLP Tutorial*.

Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1504–1515.

Wu, Z., & Giles, C. L. (2015). Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia.. In *AAAI*, pp. 2188–2194. Citeseer.

Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., & Liu, T.-Y. (2014). Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1219–1228. ACM.

Yaghoobzadeh, Y., & Schütze, H. (2016). Intrinsic subspace evaluation of word embedding representations. In *Proceedings of ACL*, pp. 236–246.

Yang, B., Yih, W.-t., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

Yang, X., & Mao, K. (2016). Learning multi-prototype word embedding from single-prototype word embedding with integrated knowledge. *Expert Systems with Applications*, *56*, 291 – 299.

Young, J., Kunze, L., Basile, V., Cabrio, E., Hawes, N., & Caputo, B. (2017). Semantic web-mining and deep vision for lifelong object discovery. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 2774–2779. IEEE.

Yu, M., & Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In *ACL (2)*, pp. 545–550.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA.

Zou, W. Y., Socher, R., Cer, D. M., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation.. In *Proceedings of EMNLP*, pp. 1393–1398, Seattle, USA.