Natural Language Engineering **24** (2): 265–312. © Cambridge University Press 2018
doi:10.1017/S135132491700047X

A framework for enriching lexical semantic resources with distributional semantics

CHRIS BIEMANN[1], STEFANO FARALLI[2], ALEXANDER PANCHENKO[1] and SIMONE PAOLO PONZETTO[2]

[1]*Language Technology Group, Department of Informatics, Faculty of Mathematics, Informatics, and Natural Sciences, Universität Hamburg, Germany*
e-mails: {biemann,panchenko}@informatik.uni-hamburg.de
[2]*Data and Web Science Group, School of Business Informatics and Mathematics, Universität Mannheim, Germany*
e-mails: {stefano,simone}@informatik.uni-mannheim.de

(*Received 28 July 2016; revised 12 December 2017; accepted 15 December 2017; first published online 15 January 2018*)

Abstract

We present an approach to combining distributional semantic representations induced from text corpora with manually constructed lexical semantic networks. While both kinds of semantic resources are available with high lexical coverage, our aligned resource combines the domain specificity and availability of contextual information from distributional models with the conciseness and high quality of manually crafted lexical networks. We start with a distributional representation of induced senses of vocabulary terms, which are accompanied with rich context information given by related lexical items. We then automatically disambiguate such representations to obtain a full-fledged proto-conceptualization, i.e. a typed graph of induced word senses. In a final step, this proto-conceptualization is aligned to a lexical ontology, resulting in a hybrid aligned resource. Moreover, unmapped induced senses are associated with a semantic type in order to connect them to the core resource. Manual evaluations against ground-truth judgments for different stages of our method as well as an extrinsic evaluation on a knowledge-based Word Sense Disambiguation benchmark all indicate the high quality of the new hybrid resource. Additionally, we show the benefits of enriching top-down lexical knowledge resources with bottom-up distributional information from text for addressing high-end knowledge acquisition tasks such as cleaning hypernym graphs and learning taxonomies from scratch.

1 Introduction

Automatic enrichment of semantic resources is an important problem (Biemann 2005; Jurgens and Pilehvar 2016) as it attempts to alleviate the extremely costly process of *manual* lexical resource (LR) construction. Distributional semantics (Turney and Pantel 2010; Baroni, Dinu and Kruszewski 2014; Clark 2015) provides an alternative *automatic* meaning representation framework that has been shown to benefit many text-understanding applications.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Recent years have witnessed an impressive amount of work on combining the symbolic semantic information available in manually constructed LRs with distributional information, where words are usually represented as dense numerical vectors, a.k.a. embeddings. Examples of such approaches include methods that modify the Skip-gram model (Mikolov *et al.* 2013) to learn sense embeddings (Chen, Liu and Sun 2014) based on the sense inventory of WordNet, methods that learn embeddings of synsets as given in an LR (Rothe and Schütze 2015) or methods that acquire word vectors by applying random walks over LRs to learn a neural language model (Goikoetxea, Soroa and Agirre 2015). Besides, alternative approaches like NASARI (Camacho-Collados, Pilehvar and Navigli 2015b) and MUFFIN (Camacho-Collados, Pilehvar and Navigli 2015a) looked at ways to produce joint lexical and semantic vectors for a common representation of word senses in text and in LRs. Retrofitting approaches, e.g. Faruqui *et al.* (2015), efficiently 'consume' LRs as input in order to improve the quality of word embeddings, but do not add anything to the resource itself. Other approaches, such as AutoExtend (Rothe and Schütze 2015), NASARI and MUFFIN, learn vector representations for existing synsets that can be added to the resource, however are not able to add missing senses discovered from text.

In these contributions, the benefits of such hybrid knowledge sources for tasks in computational semantics such as semantic similarity and Word Sense Disambiguation (WSD) (Navigli 2009) have been extensively demonstrated. However, none of the existing approaches, to date, have been designed to use distributional information for the enrichment of lexical semantic resources, i.e. adding new symbolic entries.

In this article, we consequently set out to filling this gap by developing a framework for enriching lexical semantic resources with distributional semantic models. The goal of such framework is the creation of a resource that brings together the 'best of both worlds', namely the *precision* and *interpretability* from the lexicographic tradition that describes senses and semantic relations manually, and the *versatility* of data-driven, corpus-based approaches that derive senses automatically.

An LR enriched with additional knowledge induced from text can boost the performance of natural language understanding tasks like WSD or Entity Linking (Mihalcea and Csomai 2007; Rospocher *et al.* 2016), where it is crucial to have a comprehensive list of word senses as well as the means to assign the correct of many possible senses for a given word in context.

Consider, for instance, the following sentence:

Regulator of calmodulin signaling (RCS) knockout mice display anxiety-like behavior and motivational deficits.[1]

No synset for 'RCS' can be found in either WordNet (Fellbaum 1998) or BabelNet (Navigli and Ponzetto 2012a), yet it is distributionally related to other biomedical concepts and thus can help to disambiguate the ambiguous term mice to its 'animal' meaning, as opposed to the 'computer peripheral device'.

Our approach yields a hybrid resource that combines symbolic and statistical meaning representations while (i) staying purely on the lexical-symbolic level,

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3622044

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

(ii) explicitly distinguishing word senses and (iii) being human readable. These properties are crucial to be able to embed such a resource in an explicit semantic data space such as, for instance, the Linguistic Linked Open Data ecosystem (Chiarcos, Hellmann and Nordhoff 2012). According to Norvig (2016), the semantic web and natural language understanding are placed at the heart of current efforts to understand the web on a large scale.

We take the current line of research on hybrid semantic representations one step forward by presenting a methodology for inducing distributionally based sense representations from text, and for linking them to a reference LR. Central to our method is a novel sense-based distributional representation that we call proto-conceptualization (PCZ). A PCZ is a repository of word senses induced from text, where each sense is represented with related senses, hypernymous senses and aggregated clues for contexts in which the respective sense occurs in text. Besides, to further improve interpretability, each sense has an associated image. We build a bridge between the PCZ and lexical semantic networks by establishing a mapping between these two kinds of resources.[2] This results in a new knowledge resource that we refer to as *hybrid aligned resource* (HAR); here, senses induced from text are aligned to a set of synsets from a reference LR, whereas induced senses that cannot be matched are included as additional synsets. By linking our distributional representations to a repository of symbolically encoded knowledge, we are able to complement wide-coverage statistical meaning representations with explicit relational knowledge as well as to extend the coverage of the reference LR based on the senses induced from a text collection. The main contributions of this article are listed as follows:

- **A framework for enriching lexical semantic resources**: We present a methodology for combining information from distributional semantic models with manually constructed lexical semantic resources.
- **A hybrid lexical semantic resource**: We apply our framework to produce a novel hybrid resource obtained by linking disambiguated distributional lexical semantic networks to WordNet and BabelNet.
- **Applications of the hybrid resource**: Besides *intrinsic* evaluations of our approach, we test the utility of our resource *extrinsically* on the tasks of WSD and taxonomy induction, demonstrating the benefits of combining distributional and symbolic lexical semantic knowledge.

The remainder of this article is organized as follows: we first review related work in Section 2 and provide an overview of our framework to build a HAR from distributional semantic vectors and a reference knowledge graph in Section 3. Next, we provide details on our methodology to construct PCZs and to link them to an LR in Sections 4 and 5, respectively. In Section 6, we benchmark the quality of our resource in different evaluation settings, including an intrinsic and an extrinsic evaluation on the task of knowledge-based WSD using a dataset from a SemEval task. Section 7 provides two application-based evaluations that demonstrate how

[2] We use WordNet and BabelNet; however, our approach can be used with similar resources, e.g. those listed at `http://globalwordnet.org/wordnets-in-the-world`.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

our hybrid resource can be used for taxonomy induction. We conclude with final remarks and future directions in Section 8.

2 Related work

2.1 Automatic construction of lexical semantic resources

In the past decade, large efforts have been undertaken to research the automatic acquisition of machine-readable knowledge on a large scale by mining large repositories of textual data (Banko *et al.* 2007; Carlson *et al.* 2010; Fader, Soderland and Etzioni 2011; Faruqui and Kumar 2015). At this, collaboratively constructed resources have been exploited, used either in isolation (Bizer *et al.* 2009; Ponzetto and Strube 2011; Nastase and Strube 2013) or complemented with manually assembled knowledge sources (Suchanek, Kasneci and Weikum 2008; Gurevych *et al.* 2012; Navigli and Ponzetto 2012a; Hoffart *et al.* 2013). As a result of this, recent years have seen a remarkable renaissance of knowledge-rich approaches for many different artificial intelligence tasks (Hovy, Navigli and Ponzetto 2013). Knowledge contained within these very large knowledge repositories, however, has major limitations in that these resources typically do not contain any linguistically grounded probabilistic representation of concepts, instances and their attributes – namely, the bridge between wide-coverage conceptual knowledge and its instantiation within natural language texts. While there are large-scale LRs derived from large corpora such as ProBase (Wu *et al.* 2012), these are usually not sense aware but conflate the notions of term and concept. With this work, we provide a framework that aims at augmenting any of these wide-coverage knowledge sources with distributional semantic information, thus extending them with text-based contextual information.

Another line of research has looked at the problem of Knowledge Base Completion (Nickel *et al.* 2016). Many approaches to Knowledge Base Completion focus on exploiting other KBs (Wang *et al.* 2012; Bryl and Bizer 2014) for acquiring additional knowledge, or rely on text corpora – either based on distant supervision (Snow, Jurafsky and Ng 2006; Mintz *et al.* 2009; Aprosio, Giuliano and Lavelli 2013) or by rephrasing KB relations as queries to a search engine (West *et al.* 2014) that returns results from the web as corpus. Alternative methods primarily rely on existing information in the KB itself (Bordes *et al.* 2011; Jenatton *et al.* 2012; Socher *et al.* 2013; Klein, Ponzetto and Glavaš 2017) to simultaneously learn continuous representations of KB concepts and KB relations by exploiting the KB structure as the ground truth for supervision, inferring additional relations from existing ones. Lexical semantic resources and text are synergistic sources, as shown by complementary work from Faruqui *et al.* (2015), who improve the quality of semantic vectors based on lexicon-derived relational information.

Here, we follow this intuition of combining structured knowledge resources with distributional semantic information from text, but focus instead on providing hybrid semantic representations for KB concepts and entities, as opposed to the classification task of Knowledge Base Completion that aims at predicting additional semantic relations between known entities.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

2.2 Combination of distributional semantics with lexical resources

Several prior approaches combined distributional information extracted from text with information available in LRs like e.g. WordNet. This includes a model (Yu and Dredze 2014) to learn word embeddings based on lexical relations of words from WordNet and PPDB (Ganitkevitch, Van Durme and Callison-Burch 2013). The objective function of this model combines that of the skip-gram model (Mikolov *et al.* 2013) with a term that takes into account lexical relations of target words. In work aimed at retrofitting word vectors (Faruqui *et al.* 2015), a related approach was proposed that performs post-processing of word embeddings on the basis of lexical relations from LRs. Finally, Pham, Lazaridou and Baroni (2015) also aim at improving word vector representations by using lexical relations from WordNet, targeting similar representations of synonyms and dissimilar representations of antonyms. While all these three approaches show excellent performance on word relatedness evaluations, they do not model word senses – in contrast to other work aimed instead at learning sense embeddings using the word sense inventory of WordNet (Jauhar, Dyer and Hovy 2015).

A parallel line of research has recently focused on learning unified statistical and symbolic semantic representations. Approaches aimed at providing unified semantic representations from distributional information and LRs have accordingly received an increasing level of attention (Chen *et al.* 2014; Goikoetxea, Soroa and Agirre 2015; Rothe and Schütze 2015; Camacho-Collados *et al.* 2015b; Nieto Piña and Johansson 2016), *inter alia* (cf. also our introductory discussion in Section 1) and hybrid meaning representations have been shown to benefit challenging semantic tasks such as WSD and semantic similarity at word level and text level.

All these diverse contributions indicate the benefits of hybrid knowledge sources for learning word and sense representations; here, we further elaborate along this line of research and develop a new hybrid resource that combines information from the knowledge graph with distributional sense representations that are human readable and easy to interpret, in contrast to dense vector representations, a.k.a. word embeddings like GloVe (Pennington, Socher and Manning 2014) or word2vec (Mikolov *et al.* 2013). As a result of this, we are able to provide, to the best of our knowledge, the first hybrid knowledge resource that is fully integrated and embedded within the semantic web ecosystem provided by the Linguistic Linked Open Data cloud (Chiarcos, Hellmann and Nordhoff 2012). Note that this complementary to recent efforts aimed at linking natural language expressions in text with semantic relations found within LOD knowledge graphs (Krause *et al.* 2015), in that we focus instead on combining explicit semantic information with statistical, distributional semantic representations of concepts and entities into an augmented resource.

3 Enriching lexical semantic resources with distributional semantics

The construction of our HAR builds upon methods used to link various manually constructed LRs to construct BabelNet (Ponzetto and Navigli 2010) and UBY (Gurevych *et al.* 2012), among others. In our method, however, linking is performed between two networks that are structurally similar, but have been

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

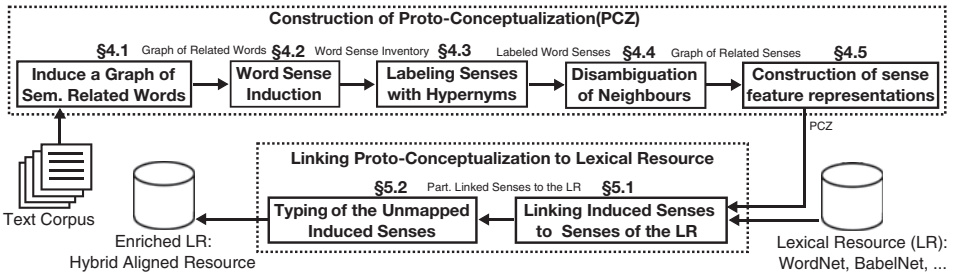


Fig. 1. Overview of the proposed framework for enriching lexical resources: a distributional semantic model is used to construct a disambiguated distributional lexical semantic network (a proto-conceptualization, PCZ), which is subsequently linked to the lexical resource.

Table 1. *Examples of entries of the PCZ resource for words* mouse *and* keyboard *after disambiguation of their related terms and hypernyms (Section 4.4)*

Word sense	Related senses	Hypernyms	Context clues
mouse**:0**	rat**:0**, rodent**:0**, monkey**:0**, ...	animal**:0**, species**:1**, ...	rat:conj_and, white-footed:amod, ...
mouse**:1**	keyboard**:1**, printer**:0**, computer**:0** ...	device**:1**, equipment**:3**, ...	click:-prep_of, click:-nn, ...
keyboard**:0**	piano**:1**, synthesizer**:2**, organ**:0** ...	instrument**:2**, device**:3**, ...	play:-dobj, electric:amod, ..
keyboard**:1**	keypad**:0**, mouse**:1**, screen**:1** ...	device**:1**, technology**:0** ...	computer:nn, qwerty:amod ...

Context clues are represented as typed dependency relations to context words in the input corpus, e.g. keyboard:0 appears as direct object of 'to play'. Trailing numbers indicate automatically induced sense identifiers.

constructed in two completely different ways: one resource is built using an unsupervised bottom-up approach from text corpora, while the second is constructed in a top-down manner using manual labor, e.g. codified knowledge from human experts such as lexicographers (WordNet). In particular, the method consists of following two major phases, as illustrated in Figure 1:

(1) **Construction of a PCZ from text**. Initially, a symbolic disambiguated distributional lexical semantic network, called a PCZ, is induced from a text corpus. To this end, we first create a sense inventory from a large text collection using graph-based word sense induction as provided by the JoBimText project (Biemann and Riedl 2013). The resulting structure contains induced word senses, their yet undisambiguated related terms, as well as context clues per term. First, we obtain sense representations by aggregating context clues over sense clusters. Second, we disambiguate related terms and hypernyms to produce a fully disambiguated resource where all terms have a sense identifier. Hence, the PCZ is a repository of corpus-induced word senses, where each sense is associated with a list of related senses, hypernymous senses, as well as aggregated contextual clues (Table 1).

(2) **Linking a PCZ to an LR.** Next, we align the PCZ with an existing LR. In our work, we make use of lexical semantic resources such as WordNet and BabelNet featuring large vocabularies of lexical units with explicit meaning representations as well as semantic relations between these. In this phase, we

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

create a mapping between the two sense inventories from the PCZ and the LR, and combine them into a new extended sense inventory, our HAR. Finally, to obtain a complete unified resource, we link the 'orphan' PCZ senses for which no corresponding sense could be found by inferring their type (i.e. their most specific generalization) in the LR.

In the following sections, we present each stage of our approach in detail.

4 Construction of a proto-conceptualization

Our method for PCZ construction consists of the four steps illustrated in the upper half of Figure 1, inducing a graph of semantically related words (Section 4.1), word sense induction (Section 4.2), labeling of clusters with hypernyms and images (Section 4.3) and disambiguation of related words and hypernyms with respect to the induced sense inventory (Section 4.4). Further, we describe an additional property of our PCZs, namely the availability of corpus-derived context clues (Section 4.5), as well as alternative ways to construct PCZs on the basis of dense vector representations (Section 4.6).

The PCZ resource with a pipeline as outlined in Figure 1 consists of word senses induced from a corpus. For each word sense, similar and superordinate terms are disambiguated with respect to the induced sense inventory: the structure of a PCZ resembles that of a lexical semantic resource. Sense distinctions and distributions depend on the training corpus, which causes the resource to adapt to its domain. In contrast to manually created resources, each sense also contains context clues that allow disambiguating polysemous terms in context. Table 1 shows example senses for the terms mouse and keyboard. Note that PCZs may contain many entries for the same word, e.g. mouse has two senses, the 'animal' and the 'computer device', respectively. The context clues are not disambiguated, since they are meant for directly matching (undisambiguated) textual context. PCZs can be interpreted by humans at two levels, as exemplified in Figure 2.

(1) The **word sense inventory** is interpretable due to the use of the hypernyms, images and related senses.
(2) The **sense feature representation** is interpretable due to the use of the sparse context clues relevant to the sense.

Note that while in our experiments we rely on a count-based sparse distributional model, the PCZ is a symbolic structure that can be also constructed using alternative distributional models, e.g. word and sense embeddings (cf. Section 4.6).

4.1 Inducing a graph of semantically related words

The goal of this first stage is to build a graph of semantically related words, with edges such as (mouse, keyboard, 0.78), i.e. a distributional thesaurus (DT) (Lin 1998). To induce such graph in an unsupervised way, we rely on a count-based approach to distributional semantics based on the *JoBimText* framework (Biemann and Riedl 2013). Each word is represented by a bag of syntactic dependencies

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

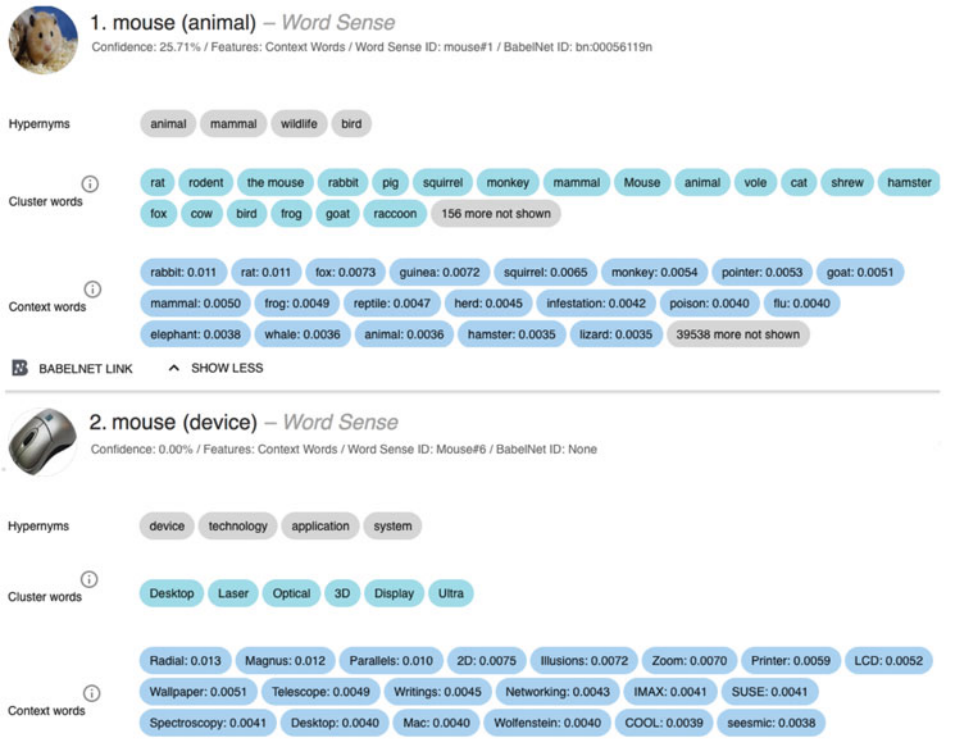


Fig. 2. (Colour online) Word sense representations of the word mouse induced from text generated using the online demo at `http://jobimtext.org/wsd`. The sense labels (device and animal) are obtained automatically based on cluster labeling with hypernyms. The images associated with the senses are retrieved with a search engine using the queries: mouse device and mouse animal. Note the 'BabelNet Link' button, leading to the sense in BabelNet linked to the induced sense with the algorithm described in Section 5.

such as conj_and(rat, •) or prep_of(click,•) extracted from the Stanford Dependencies (De Marneffe, MacCartney and Manning 2006) obtained with the PCFG model of the Stanford parser (Klein and Manning 2003).

Features of each word are weighted and ranked using the Local Mutual Information (LMI) metric (Evert 2005). Subsequently, these word representations are pruned keeping 1,000 most salient features per word and 1,000 most salient words per feature. The pruning reduces computational complexity and noise (Riedl 2016). Finally, word similarities are computed as the number of common features for two words. This is, again, followed by a pruning step in which for every word, only the 200 most similar terms are kept. The resulting graph of word similarities is browsable online (Ruppert *et al.* 2015).[3]

There are many possible ways to compute a graph of semantically similar words, including count-based approaches, such as Lin (1998), Curran (2002) or prediction-based approaches, such as word2vec (Mikolov *et al.* 2013), GloVe (Pennington *et al.*

[3] Word and sense representations used in our experiments can be inspected by selecing the 'Stanford (English)' model in the JoBimViz demo at `http://jobimtext.org/jobimviz/`.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 2. *Comparison of state-of-the-art count- and prediction-based methods to distributional semantics on the basis of the average of the averaged similarity scores between each term in the DT and its top-10 most similar terms using the WordNet path similarity measure (higher means better) averaged over 1,000 high- and low-frequency words*

Method	High	Low
Lin's similarity (Lin 1998)	0.2872	0.2291
t-test (Curran 2002)	0.2589	0.2067
Skip-gram (Mikolov *et al.* 2013)	0.2548	0.2068
Skip-gram with dependency features (Levy and Goldberg 2014)	0.2632	0.1992
LMI with trigram features (Riedl and Biemann 2013)	0.2621	0.2003
LMI with dependency features (Riedl and Biemann 2013)	**0.2933**	**0.2337**

In this article, we use 'LMI with dependency features' as the similarity function.

2014) and word2vecf (Levy and Goldberg 2014). Here, we opt for a count-based approach to distributional semantics based on Local Mutual Information based on two considerations, namely their higher quality of similarity scores and their interpretability.

A thorough experimental comparison of different approaches to computing distributional semantic similarity to build a DT is presented by Riedl (2016, Section 5.7.4) using the WordNet taxonomy as a gold standard. In this evaluation, different DTs are compared by computing, for each term, the average similarity between the term itself and its k most similar terms (based on the DT) using the WordNet path-based similarity measure (Pedersen, Patwardhan and Michelizzi 2004). The overall similarity of the DT with the ground-truth taxonomy (e.g. WordNet) is then given by the average similarity score across all terms. Using this evaluation framework, Riedl is able to compare a wide range of different approaches for the construction of a weighted similarity graph, including state-of-the-art approaches based on sparse vector representations (Lin 1998; Curran 2002), as well as dense representations based on word2vec (Mikolov *et al.* 2013) and word2vecf, which makes use of dependency-based features (Levy and Goldberg 2014). In his experiment, all methods were trained on the same corpus, and all dependency-based models, including the Skip-gram approach trained with the word2vecf tool (Levy and Goldberg 2014), used the same feature representations.

We report some of the results from Riedl's experiments in Table 2. In this experiment, 1,000 infrequent and 1,000 frequent nouns proposed by Weeds, Weir and McCarthy (2004) were used. Dependency-based models (all models except the Skip-gram) used syntactic features extracted using the Stanford Parser. In addition to this dependency-based model, we report results of the same model based on trigram features, where a context is formed by the concatenation of the two adjacent words. All models were trained on the 105 million sentences of newspaper data described in Section 6.1. Further details of the experiment, e.g. parameters of the models, are available in Riedl (2016, Section 5.7.4).

The performance figures indicate that the method we use here yields the overall best performance in terms of semantic similarity compared to other count-based or

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

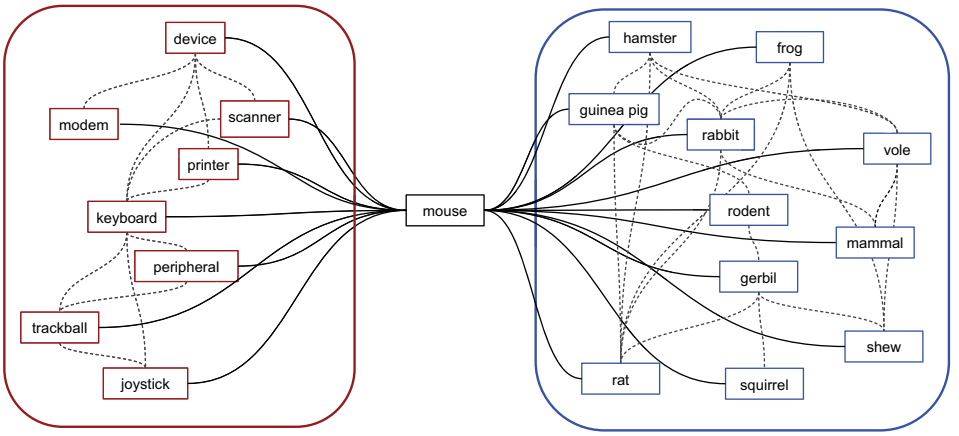


Fig. 3. (Colour online) Example of graph-based word sense induction for the word mouse: the two circles denote two induced word senses, as found by analysis of the ego graph of mouse.

word-embedding approaches (including both word2vec and word2vecf). Besides, the results generally indicate the advantage of using dependency-based context representations over the bag-of-words representations. This is in line with prior studies on semantic similarity (Padó and Lapata 2007; Van de Cruys 2010; Panchenko and Morozova 2012). For this reason, we use syntactic features in our experiments but would like to stress that the overall framework also allows simpler context representations, giving rise to its application to resource-poor languages.

The second reason for using the Local Mutual Information approach to compute a graph of semantically related words is the fact that the resulting word representations are human interpretable, since words are represented by sparse features – as opposed to dense features such as those found within word embeddings. Besides being a value on its own, this feature enables a straightforward implementation of word sense disambiguation methods on the basis of the learned representations (Panchenko *et al.* 2017a, 2017c).

4.2 Word sense induction

In the next stage, we induce a sense inventory on top of the DT by clustering ego-networks of similar words. In our case, an inventory represents senses by a word cluster, such as {rat, rodent, monkey, ...} for the 'animal' sense of the word mouse.

Sense induction is conducted one word t at the time on the DT. First, we retrieve nodes of the ego-network G of t being the N most similar words of t according to the DT. Figure 3 presents a sample ego network of related words.[4] Note that the target word t itself is excluded during clustering. Second, we connect each node in G to its n most similar words according to DT. The n parameter regulates the granularity of the induced sense inventory: we experiment with $n \in \{200, 100, 50\}$ and $N = 200$.

[4] See the `Serelex.org` system for further visualizations of ego networks of semantically related words (Panchenko *et al.* 2013).

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Finally, the ego network is clustered with Chinese Whispers (Biemann 2006), a non-parametric algorithm that discovers the number of clusters (word senses, in our case) automatically. The algorithm is iterative and proceeds in a bottom-up fashion. Initially, all nodes have distinct cluster identifiers. At each step, a node obtains the cluster identifier from the *dominant* cluster in its direct neighborhood, which is the cluster with the highest sum of edge weights to the current node.

The choice of Chinese Whispers among other algorithms, such as HyperLex (Véronis 2004) or Markov Cluster Algorithm (Van Dongen 2008), was motivated by the absence of meta-parameters, its state-of-the-art performance on Word Sense Induction tasks (Di Marco and Navigli 2013), as well as its efficiency (time-linear in the number of edges), see Cecchini, Riedl and Biemann (2017) for a comparative evaluation.

4.3 Labeling induced senses with hypernyms and images

At the third stage, each sense cluster is automatically labeled to characterize it in more detail and to improve its interpretability. First, we extract hypernyms from the input corpus. Here, we rely on the Hearst (1992) patterns, yet the approach we use can benefit also from more advanced methods for extraction of hypernyms, e.g. HypeNet (Shwartz, Goldberg and Dagan 2016) or the Dual Tensor Model (Glavaš and Ponzetto 2017). Note that despite their simplicity, Hearst patterns still are a strong baseline, used for applications like, for instance, taxonomy induction (Panchenko *et al.* 2016; Bordea, Lefever and Buitelaar 2016).

Second, we rank the quality of a hypernym h to act as generalization for the meaning expressed by cluster c on the basis of the product of two scores, namely *hypernym relevance* and *coverage*:

$$relevance(c,h) = \sum_{w \in c} rel(t,w) \cdot freq(w,h)$$

$$coverage(c,h) = \sum_{w \in c} \min(freq(w,h), 1)$$

where $rel(t,w)$ is the relatedness of the cluster word w to the target word t (the ambiguous word in an ego network, cf. Figure 3) and $freq(w,h)$ is the frequency of the hypernymy relation (w,h) as extracted via patterns. Thus, a highly ranked hypernym h needs to be observed frequently in a hypernym pattern, but also needs to be confirmed by several cluster words. This stage results in a ranked list of labels that specify the word sense, which we add to the PCZ. The highest scoring hypernym is further used in the title of the word sense, e.g. mouse (device) or mouse (animal).

Finally, to further improve the interpretability of the induced senses, we add images to our sense clusters as follows. Previous work (Faralli and Navigli 2012) showed that web search engines can be used to bootstrap sense-related information. Consequently, we assign an image to each word in the cluster querying the Bing image search API[5] using the query composed of the target word and its highest

[5] https://azure.microsoft.com/en-us/services/cognitive-services/search

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Algorithm 1 Disambiguation of related terms and hypernyms

Require: WSI, a word sense inventory in the form of a set of tuples (*word*, *sense_id*, *cluster*, *isas*), where *cluster* and *isas* have no sense identifiers.
Ensure: PCZ, a proto-conceptualization in the form of a set of tuples (*word*, *sense_id*, *cluster′*, *isas′*), where *cluster′* and *isas′* are disambiguated with respect to sense inventory of the WSI.

$PCZ = \emptyset$
for all $(word, sense_id, cluster, isas) \in WSI$ **do**
  $cluster' =$ DISAMBIGUATE $(cluster, word, WSI)$
  $isas' =$ DISAMBIGUATE $(isas, word, WSI)$
  $PCZ = PCZ \cup (word, sense_id, cluster', isas')$

function DISAMBIGUATE $(cluster, cword, WSI)$
$cluster' = \emptyset$
$context = cluster \cup (cword, 1.0)$
for all $word, sim \in cluster$ **do**
  $sense_id = -1, max_sim = 0$
  for all $(dword, dsense_id, dcluster, disas) \in$ GETSENSES$(word, WSI)$ **do**
    if $sim(context, dcluster) > max_sim$ **then**
      $sense_id = dsense_id$
      $max_sim = sim$
  $cluster' = cluster' \cup (word, sim, sense_id)$
return *cluster′*

scoring hypernym, e.g. mouse device. The first image result of this query is selected to represent the induced word sense. This step is optional in our pipeline, and is primarily aimed at improving the user interaction with the word sense inventory.

The resulting sense representation is illustrated in Figure 2 for two induced senses of the word mouse. Providing to the user hypernyms, images, list of related senses and the list of the most salient context clues ensures interpretability of each sense. Note that all these elements are obtained without manual intervention, see Panchenko *et al.* (2017b) for more details.

4.4 Disambiguation of related terms and hypernyms

Next, we disambiguate the lexical graphs induced in the previous step. Each word in the induced inventory has one or more senses; however, the list of related words and hypernyms of each induced sense does not carry sense information yet. In our example (Table 1), the sense of mouse for the entry keyboard:1 could have either referred to the 'animal' or the 'electronic device'. Consequently, we apply a semantic closure procedure to arrive at a resource in which all terms get assigned a unique, best-fitting sense identifier. Our method assigns each disambiguation target word w – namely, a similar or superordinate term from each sense of the induced word sense inventory – the sense $\hat{s}$ whose context (i.e. the list of similar or superordinate terms) has the maximal similarity with the target word's context (i.e. the other words in the target word's list of similar or superordinate items). We use the cosine similarity between context vectors to find the most appropriate sense $\hat{s}$ matching the 'context' of an ambiguous word *cluster* one of its 'definitions' $WSI(w').cluster$:

$$\hat{s} = \underset{(w', _, cluster, _) \in WSI(w)}{\operatorname{argmax}} cos(WSI(w').cluster, cluster). \quad (1)$$

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

This way we are able to link, for instance, keyboard in the list of similar terms for mouse:1 to its 'device' sense (keyboard:1), since mouse:1 and keyboard:1 share a large amount of terms from the information technology domain. This simple, local approach is scalable (cf. the complexity analysis at the end of this section) and it performs well, as we show later in the evaluation.

Algorithm 1 presents our method to compute the semantic closure. The input is a JoBimText model as a set of tuples $(word, sense_id, cluster, isas)$, where *cluster* is a list of similar terms in the format $(word_i, sim_i)$ with sim_i being the similarity value between *word* and $word_i$, and *isas* is a list of hypernym terms in the same format. The algorithm outputs a PCZ in the form of a set of tuples $(word, sense_id, cluster', isas')$, where $cluster'$ is a list of disambiguated similar terms and $isas'$ is a list of disambiguated hypernym terms both in the format $(word_i, sim_i, sense_id_i)$. The algorithm starts by creating an empty PCZ structure PCZ. For each entry of an input JoBimText model, we disambiguate related words (*cluster*) and hypernym terms (*isas*) with the function DISAMBIGUATE (lines 3–4). This function retrieves for each *word* in a *cluster* the set of its senses with the GETSENSES function. Next, we calculate similarity between the *cluster* of the *word* and the cluster of the candidate sense (denoted as *dcluster*). The $word_i$ obtains the *sense_id* of the candidate sense that maximizes this similarity (lines 8–13).

Our disambiguation approach is a rather straightforward algorithm based on similarity computations. Despite its simplicity, we are able to achieve a disambiguation accuracy in the high ninety per cent range for noun word senses, while at the same time having a time-linear complexity in the number of word senses, as we will show in the evaluation below (Section 6.2). We can assume, in fact, that the function GETSENSES has a run-time complexity of $O(1)$ and the function *cos* (Equation (1)) has complexity $O(m)$, where m is the average number of neighbors of each word sense. Then, the run-time complexity of the algorithm is $O(n * m^2 * k)$, where n is the overall number of induced word senses, and k is the average polysemy of a word. Since k is small and m is bound by the maximum number of neighbors (200 in our case), the amortized run time is linear in the vocabulary size. This makes our approach highly scalable: in recent experiments, we have been accordingly able to apply our method at web scale on the CommonCrawl,[6] the largest existing public repository of web content.

4.5 Construction of sense feature representations

Finally, we calculate feature representations for each sense in the induced inventory – that is, grammatical dependency features that are meant to provide an aggregated representation of the contexts in which a word sense occurs.

We assume that a word sense is a composition of cluster words that represent the sense and accordingly define a sense vector as a function of word vectors representing cluster items. Let W be a set of all words in the training corpus and let $S_i = \{w_1, \ldots, w_n\} \subseteq W$ be a sense cluster obtained in a previous step. Consider

[6] https://commoncrawl.org

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 3. *Sense inventories derived from the Wikipedia corpus via a sparse count-based (JoBimText) and dense predict-based (Skip-gram) distributional models*

	Sparse vectors (JoBimText)	Dense vectors (word2vec)
Mouse (animal)	Rat, rodent, monkey, pig, animal, human, rabbit, cow	Rat, hamster, hedgehog, mole, monkey, kangaroo, skunk
Mouse (device)	Keyboard, computer, printer, joystick, stylus, modem	Cursor, keyboard, AltGr, chording, D-pad, button, trackball

a function $vec_w : W \rightarrow \mathbb{R}^m$ that maps words to their vectors and a function $\gamma_i : W \rightarrow \mathbb{R}$ that maps cluster words to their weight in the cluster S_i. The sense vector representation (the context clues) is then a weighted average of word vectors:

$$S_i = \frac{\sum_{k=1}^{n} \gamma_i(w_k) vec_w(w_k)}{\sum_{k=1}^{n} \gamma_i(w_k)}. \quad (2)$$

Table 1 (column 4) provides an example of such feature representations. While the averaged word vectors are ambiguous and can contain features related to various senses, features with high weights tend to belong to the target sense as the secondary senses of the averaged words vectors rarely match semantically, hence the aggregation amplifies the correct sense.

This concludes the description of steps we use to construct PCZs from text corpora.

4.6 Inducing PCZs with dense vector representations

In this section, we briefly describe alternative routes to the construction of a PCZ from text in an unsupervised way. In the remainder of this article, we will rely on the results of the approach described above. The goal of this section is to show that our overall framework is agnostic to the type of underlying distributional semantic model. In this section, we consider three approaches to generating a PCZ using word or sense embeddings.

Option 1: Inducing PCZs using word embeddings with explicit disambiguation. As illustrated in Figure 1, the first stage of our approach involves the computation of a graph of semantically similar words. Above, the graph was induced using a count-based model, however, any of the models listed in Table 2 can be used to generate such a graph of ambiguous words including models based on dense vector representations, such as the Skip-gram model. In this strategy, one would need to generate top nearest neighbors of word on the basis of cosine similarity between word embeddings. Table 3 shows an excerpt of nearest neighbors generated using the JoBimText and word2vec toolkits. The obtained word graphs can be subsequently used to induce word senses using the graph-based approach described in Section 4.2. The obtained clusters can be labeled using a database of hypernyms exactly in the same way as for the models based on the count-based JoBimText framework

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 4. *A Skip-gram-based PCZ model by Pelevina et al. (2016): Neighbors of the word* mouse *and the induced senses*

Vector	Nearest neighbors
Mouse	Rat, keyboard, hamster, hedgehog, monkey, kangaroo, cursor, button
Mouse:0	Rat:0, hamster:0, hedgehog:1, mole:0, monkey:0, kangaroo:0, skunk:0
Mouse:1	Cursor:0, keyboard:1, AltGr:0, chording:1, D-pad:0, button:0

The neighbors of the initial vector belong to both senses, while those of sense vectors are sense specific.

(cf. Section 4.3). All further steps of the workflow presented in Figure 1 remain the same.

The main difference between the approach described above and the methods based on dense representations of words is the representation of the context clues of PCZ (cf. Table 1). In the case of an underlying sparse count-based representation, context clues remain human readable and interpretable, whereas in case of dense representations, context clues are represented by a dense vector embedding, and it is not straightforward to aggregate context clues over sense clusters.

Option 2: Inducing PCZs using word embeddings without explicit disambiguation. The first two stages of this approach are the same compared to the previous strategy. Namely, first one needs to generate a graph of ambiguous semantically related words (Section 4.1) and then to run ego-network clustering to induce word senses (Section 4.2). However, instead of explicit disambiguation of nearest neighbors (Section 4.4), the third stage could obtain vector sense representations by averaging the word embeddings of sense clusters (Section 4.5). Finally, disambiguated nearest neighbors can be obtained by calculating nearest neighbors of each sense vector in the space of word sense embeddings. This step is equivalent to the computation of a DT (Section 4.1); however, it directly yields disambiguated nearest neighbors (cf. Table 4). Note, however that, disambiguation of hypernyms using Algorithm 1 is still required when using this approach.

This approach was explored in our previous work (Pelevina *et al.* 2016), where we showed that words sense embeddings obtained in this way can be successfully used for unsupervised WSD, yielding results comparable to the state of the art.

Option 3: Inducing PCZ using word sense embeddings. Finally, a PCZ can be also induced using sparse (Reisinger and Mooney 2010) and dense (Neelakantan *et al.* 2014; Li and Jurafsky 2015; Bartunov *et al.* 2016) multi-prototype vector space models (the latter are also known as word sense embeddings). These models directly induce sense vectors from a text corpus, not requiring the word sense induction step of our method (Section 4.2). Instead of ego-network-based sense induction, these methods rely on some form of context clustering, maintaining several vector representations for each word type during training. To construct a PCZ using such models within our framework, we need to compute a list of nearest neighbors (Section 4.1), label the obtained sense clusters with hypernyms (Section 4.3) and

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Algorithm 2 Linking induced senses to senses of a lexical resource

Require: $T = \{(j_i, R_{j_i}, H_{j_i})\}$, W, th, m
Ensure: $M = (source, target)$
$M = \emptyset$
for all $(j_i, R_{j_i}, H_{j_i}) \in T.monosemousSenses$ **do**
  $C(j_i) = W.getSenses(j_i.lemma, j_i.POS)$
  if $|C(j_i)| == 1$, let $C(j_i) = \{c_0\}$ **then**
    if $sim(j_i, c_0, \emptyset) \geq th$ **then**
      $M = M \cup \{(j_i, c_0)\}$
for $step = 1$; $step \leq m$; $step = step + 1$ **do**
  $M_{step} = \emptyset$
  for all $(j_i, R_{j_i}, H_{j_i}) \in T.senses/M.senses$ **do**
    $C(j_i) = W.getSenses(j_i.lemma, j_i.POS)$
    for all $c_k \in C(j_i)$ **do**
      $rank(c_k) = sim(j_i, c_k, M)$
    if $rank(c_k)$ has a single top value for c_t **then**
      if $rank(c_t) \geq th$ **then**
        $M_{step} = M_{step} \cup \{(j_i, c_t)\}$
  $M = M \cup M_{step}$
for all $(j_i, R_{j_i}, H_{j_i}) \in T.senses/M.senses$ **do**
  $M = M \cup \{(j_i, j_i)\}$
return M

disambiguate these hypernyms using Algorithm 1. The sense vectors replace the aggregated context clues, so the stage described in Section 4.5 is superfluous for this option as well.

We also experimented in previous work with the construction of PCZs using this approach (Panchenko 2016), showing how to use sense embeddings for building PCZs, reaching satisfactory levels of recall and precision of matching as compared to a mapping defined by human judges.

While an empirical comparison of these options would be interesting, it is beyond the scope of this paper, where our main point is to demonstrate the benefits of linking manually created LRs with models induced by distributional semantics (by example of a count-based model).

5 Linking a proto-conceptualization to a lexical semantic resource

This section describes how a corpus-induced semantic network (a PCZ) is linked to a manually created semantic network, represented by an LR.

5.1 Linking induced senses to senses of the lexical resource

Now, we link each sense in our PCZ to the most suitable sense (if any) of an LR (see Figure 1 step 3). There exist many algorithms for knowledge base linking (Pavel and Euzenat 2013); here, we build upon simple, yet high-performing previous approaches to linking LRs that achieved state-of-the-art performance. These rely at their core on computing the overlap between the bags of words built from the LRs' concept lexicalizations, e.g. Navigli and Ponzetto (2012a) and Gurevych *et al.* (2012) (*inter alia*). Specifically, we develop (i) an iterative approach – so that the linking can benefit from the availability of linked senses from previous

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

iterations – (ii) leveraging the lexical content of the source and target resources. Algorithm 2 takes as input

(1) a PCZ $T = \{(j_i, R_{j_i}, H_{j_i})\}$ where j_i is a sense identifier (i.e. mouse:1), R_{j_i} the set of its semantically related senses (i.e. $R_{j_i} = \{$keyboard:1, computer:0, ...$\}$ and H_{j_i} the set of its hypernym senses (i.e. $H_{j_i} = \{$equipment:3, ...$\}$;
(2) an LR W: we experiment with: WordNet, a lexical database for English and BabelNet, a very large multilingual 'encyclopedic dictionary';
(3) a threshold *th* over the similarity between pairs of concepts and a number m of iterations as a stopping criterion.

The algorithm outputs a mapping M, which consists of a set of pairs of the kind $(source, target)$ where $source \in T.senses$ is a sense of the input PCZ T and $target \in W.senses \cup source$ is the most suitable sense of W or *source* when no such sense has been identified.

The algorithm starts by creating an empty mapping M (line 1). Then for each monosemous sense (e.g. Einstein:0 is the only sense in the PCZ for the term Einstein) it searches for a candidate monosemous sense (lines 2–6). If such monosemous candidate senses exist (line 4), we compare the two senses (line 5) with the following similarity function:

$$sim(j, c, M) = \frac{|T.BoW(j, M, W) \cap W.BoW(c)|}{|T.BoW(j, M, W)|}, \tag{3}$$

where

(1) $T.BoW(j, M, W)$ is the set of words containing all the terms extracted from related/hypernym senses of j and all the terms extracted from the related/hypernym (i.e. already linked in M) synsets in W. For each synset from the LR, we use all synonyms and content words of the gloss.
(2) $W.BoW(c)$ contains the synonyms and the gloss content words for the synset c and all the related synsets of c.

Then a new link pair (j_i, c_0) is added to M if the similarity score between j_i and c_0 meets or exceeds the threshold *th* (line 5). At this point, we collected a first set of disambiguated (monosemous) senses in M and start to iteratively disambiguate the remaining (polysemous) senses in T (lines 7–16). This iterative disambiguation process is similar to the one we described for the monosemous case (lines 2–6), with the main difference that, due to the polysemy of the candidates synsets, we instead use the similarity function to rank all candidate senses (lines 11–12) and select the top-ranked candidates for the mapping (lines 13–15). At the end of each iteration, we add all collected pairs to M (line 16). Finally, all unlinked j of T, i.e. induced senses that have no corresponding LR sense, are added to the mapping M (lines 17–18).

Comparison with other mapping algorithms. Previous work for the construction of BabelNet (Navigli and Ponzetto 2012a) and UBY (Gurevych *et al.* 2012) looked at the related problem of matching heterogeneous lexical semantic resources, i.e. Wikipedia and WordNet. In our scenario, however, we aim instead at establishing

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Algorithm 3 Typing of the unmapped induced senses

Require: $M = (source, target)$, W
Ensure: $H = (source, type)$
1: $H = \emptyset$
2: **for all** $(source, target) \in M$ **do**
3: **if** $target \notin W$ **then**
4: $Rank = 0$
5: **for all** $related \in R_{source}, \exists (related, trelated) \in M, trelated \in W$ **do**
6: **for all** $hop \in (1, 2, 3)$ **do**
7: **for all** $ancestor \in W.ancestors(trelated, hop)$ **do**
8: $Rank(ancestor) = Rank(ancestor) + 1.0/hop$
9: **for all** $ntype \in Rank.top(top_h)$ **do**
10: $H = H \cup (source, ntype)$
11: **return** H

a bridge between any of these latter reference KBs and a PCZ – i.e. a fully disambiguated distributional semantic representation of distributionally induced word senses (cf. Section 4.5). Since we are working with a PCZ on the source side, as opposed to using Wikipedia, we cannot rely on graph-based algorithms such as PageRank (Niemann and Gurevych 2011) 'out of the box': while PCZs can be viewed as graphs, these are inherently noisy and require cleaning techniques in order to remove cycles and wrong relations (cf. Section 7 where we accordingly address the topic of taxonomy induction and cleaning within our framework). Similarly, the fact that PCZs are automatically induced from text – and hence potentially noisier that clean collaboratively generated content from Wikipedia – forces us to limit evidence for generating the mapping to local information, as opposed to, e.g. graph-based expansions used to boost the recall of BabelNet–WordNet mappings, cf. Navigli and Ponzetto (2012a). To overcome this 'locality' constraint, we develop an iterative approach to indirectly include non-local evidence based on previous mapping decisions. Our algorithm, in fact, uses previous mappings to additionally expand the bag-of-word of the candidate PCZ sense to be mapped, based on related/hypernym synsets linked in the previous iterations only (i.e. to keep the expansion 'safe', cf. Equation (3)).

5.2 Typing of the unmapped induced senses

An approach based on the bag-of-words from concept lexicalizations has the advantage of being simple, as well as high performing as we show later in the evaluation – cf. also findings from Navigli and Ponzetto (2012a). However, there could be still PCZ senses that cannot be mapped to the target LR, e.g. because of vocabulary mismatches, sparse concepts' lexicalizations, or because they are simply absent in the resource.

Consequently, in the last phase of our resource creation pipeline, we link these 'orphan' PCZ senses (i.e. those from lines 17–18 of Algorithm 2), in order to obtain a unified resource, and propose a method to infer the type of those concepts that were not linked to the target LR. For example, so far we were not able to find a BabelNet sense for the PCZ item Roddenberry:10 (the author of 'Star Trek'). However, by looking at the linked related concepts that share the same BabelNet hypernym – e.g. the PCZ items Asimov:3 *is-a* $\text{author}_{BabelNet}$, Tolkien:7 *is-a* $\text{author}_{BabelNet}$, Heinlein:8

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 5. *Text corpora used in our experiments to induce distributional disambiguated semantic networks (proto-conceptualizations, PCZs)*

Name	Language	Source of texts	Genre	Size
Wiki	English	Text of Wikipedia articles	Encyclopedic	35 million sent.
News	English	News articles: Gigaword and LCC	Narrative, publicistic	105 million sent.

is-a $\text{author}_{BabelNet}$, etc. – we can infer that Roddenberry:10 *is-a* author:1, since the latter was linked to the Babel synset $\text{author}_{BabelNet}$.

The input of Algorithm 3 consist of the mapping M of a PCZ to an LR W (cf. Algorithm 2). The output is a new mapping H containing pairs of the kind $(source, type)$ where $type$ is a type in W for the concept $source \in PCZ$. We first initialize the new mapping H as an empty set (line 1). Then for all the pairs $(source, target)$ where the target is a concept not included in the target LR W (line 3), we compute a rank of all the ancestors of each related sense that has a counterpart $trelated$ in W (lines 5–8). In other words, starting from linked related senses $trelated$, we traverse the taxonomy hierarchy (at most for three hops) in W and each time we encounter a sense $ancestor$ we increment its rank by the inverse of the distance to $trelated$. Finally, we add the pairs $(source, ntype)$ to H for all the $ntype$ at the top top_h in the $Rank$.

Finally, our final resource consists of (i) the PCZ; (ii) the mapping M of PCZ entries to the LR (e.g. WordNet or BabelNet); (iii) the mapping H of suggested types for the PCZ entries not mapped in M.

6 Experiments

In this section, we present results of four experiments, which intrinsically and extrinsically evaluate the quality of our HAR.

6.1 Corpora used for the induction of proto-conceptualizations

We evaluate our method using texts of different genres, namely standard newswire text vs. encyclopedic texts in order to examine performance in different settings. The corpora, described in Table 5, are a 105 million sentence news corpus composed of Gigaword (Parker *et al.* 2011) and the Leipzig Corpora Collection, (Richter *et al.* 2006)[7] and a 35 million-sentence Wikipedia corpus[8] from a 2011 dump.

We opt for these text collections because they were previously extensively used for the evaluation of distributional models based on the JoBimText framework (Biemann and Riedl 2013; Riedl and Biemann 2013). Specifically, previous work (Riedl and Biemann 2013) experimented with the induction of distributional models

[7] `http://corpora.uni-leipzig.de`

[8] The *wiki* corpus is downloadable (cf. Section 8). The *news* corpus is available by request due to license restrictions of the Gigaword corpus.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 6. *Structural analysis of our five word sense inventories of the proto-conceptualizations (PCZs) used in our experiments*

	Words			Senses		Polysemy		Rel.senses		Hyper.	
PCZ	*n*	#	Mono	Poly	#	Avg.	Max	#	Avg.	#	Avg.
News-p1.6	200	207 k	137 k	69 k	332 k	1.6	18	234 k	63.9	15 k	6.9
News-p2.3	50	200 k	99 k	101 k	461 k	2.3	17	298 k	44.3	15 k	5.8
Wiki-p1.8	200	206 k	120 k	86 k	368 k	1.8	15	300 k	59.3	15 k	4.4
Wiki-p6.0	30	258 k	44 k	213 k	1.5 M	6.0	36	811 k	16.9	52 k	1.7
Wiki-p1.6-mwe	200	465 k	288 k	176 k	765 k	1.6	13	662 k	46.6	30 k	3.2

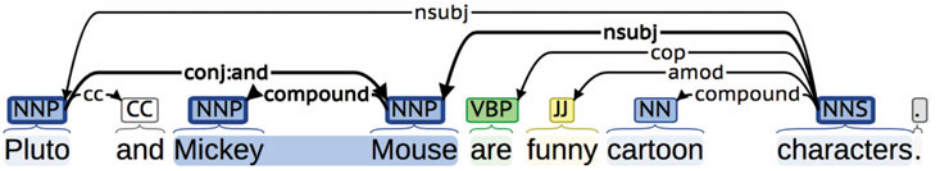


Fig. 4. (Colour online) Extraction of distributional dependency features for a multiword expression Mickey Mouse: all outgoing dependencies are used as features. This image was created using the Stanford dependency visualizer (`http://nlp.stanford.edu:8080/corenlp`).

on the basis of both corpora, and showed that the quality of semantic similarity (which, in turn, is used to build the DT, cf. Section 4.1) increases with corpus size. Since 'more data' helps, we experiment in this work with the full-sized corpora. Further description of the *wiki* and *news* text collections can be found in Riedl and Biemann (2013) and Riedl (2016, p. 94).

We experiment with different parameterizations of the sense induction algorithm to obtain PCZs with different average sense granularities, since *a priori*, there is no clear evidence for what the 'right' sense granularity of a sense inventory should be. Chinese Whispers sense clustering with the default parameters ($n = 200$) produced an average number of 2.3 (news) and 1.8 (wiki) senses per word with the usual power-law distribution of sense cluster sizes. Decreasing connectivity of the ego network via the *n* parameter leads to more fine-grained inventories (cf. Table 6).

Finally, we use the method described in Riedl and Biemann (2015) to compute a dataset that includes automatically extracted multiword terms using the Wikipedia corpus (*wiki-p1.6-mwe*). Since most of the multiwords are monosemous, average polysemy of this dataset decreased from 1.8 to 1.6 for the analogous model without multiwords (*wiki-p1.8*). To obtain a feature representation of a multiword expression, we gathered all outgoing dependency relations of this term as illustrated in Figure 4.

In Table 6, we present statistics for the five different resources we induce from our corpora. For each dataset, we report the counts of overall number of words (vocabulary size), including monosemous words and polysemous ones, respectively. For each PCZ, we report the cardinality, the average polysemy and the maximum polysemy. Finally, we report the overall and the average number of related senses

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

and hypernyms. Numbers vary across datasets due to the different nature of the two source corpora and the selection of different parameter values for sense induction.

While inducing senses directly from corpus data allows for large coverage and flexibility, it also makes it difficult to evaluate the quality of the resulting sense clusters (Agirre and Soroa 2007). Since we do not *a priori* know the sense granularity of the PCZs, the sense inventory input to our disambiguation and linking algorithms cannot be fixed in advance, e.g. in order to produce a static gold standard. Therefore, in our intrinsic evaluations (Sections 6.2–6.4), we assess the quality of our resources by manually validating a sample of the output of the different steps of our method. Later, in Section 6.5, we perform an extrinsic evaluation against a gold standard on a WSD benchmarking dataset from a SemEval task.

The PCZ described above were subsequently linked to WordNet 3.1 and BabelNet 2.5 using the method described above. All the models described above as well as the induced sense inventories and word similarity graphs can be accessed online (cf. Section 8).

6.2 Experiment 1: Quality of disambiguation of the related terms

Experimental setting. In this experiment, we evaluate the quality of Algorithm 1 for the disambiguation of related words (cf. Table 1) by performing a *post-hoc* evaluation using manual judgments on a sample of sense-disambiguated terms from one of our PCZ resources.

We manually selected a set of frequent nouns and proper nouns, such that each word has least two homonymous (as opposed to polysemous) word senses. We deliberately avoided words with polysemous senses, as word sense induction algorithms are known to robustly extract mostly coarse-grained inventories of homonymous words (Di Marco and Navigli 2013; Cecchini *et al.* 2017). The words were selected according to two criteria. First, each of the two homonymous word senses should have a comparable frequency – compare, for instance, the fairly common senses of python (animal) and python (language), as opposed to boa (animal) and boa (language), where the 'language' sense of the word boa is much rarer as compared to its 'animal' sense. Second, each of these senses should be common enough to be recognizable without the need of consulting a dictionary by a non-native, graduate-level speaker of English. We tested for sense frequencies and popularity by checking that selected senses were found among the top ones as listed in BabelNet. Using these criteria, we manually selected a set of seventeen nouns, such as apple, java, python, etc.[9]

Since our resources only partially overlap in terms of sense inventory, and there is no *a priori* reference sense granularity, we cannot perform evaluation on a shared

[9] The full list of words, where senses are denoted in the brackets using the respective hypernyms: apache (tribe|software), apple (fruit|company), bank (river|institution), commercial (ad|business), corvette (car|ship), jaguar (animal|car), java (island|technology), lotus (flower|car), mustang (horse|car), pascal (person|language), port (sea-related|computer-related), puma (animal|brand), python (snake|language), ruby (gem|language), sun (star|company), tiger (animal|tank) and viper (snake|car).

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 7. *Accuracy of Algorithm 1 for disambiguation of related words evaluated on a set of seventeen frequent words each having two non-marginal homonymous word senses, e.g. as in* mouse (keyboard) *and* mouse (animal)

Part of speech	# Word forms	# Senses	# Contexts	Accuracy
Nouns	15	30	1,055	0.94
Proper nouns	17	49	1,177	0.85
Adjectives	6	6	5,66	0.63
Verbs	4	6	86	0.76
All	42	91	2,284	0.84

gold standard. Consequently, we opt instead for a *post-hoc* evaluation of the accuracy of the disambiguation step, namely the fraction of correctly disambiguated related words among all disambiguated words. Post-hoc validations have major limitations in that they are time consuming, do not scale and hinder direct comparability across methods – nevertheless, they are commonly used in the field of knowledge acquisition to estimate the quality of knowledge resources (Banko *et al.* 2007; Suchanek *et al.* 2008; Carlson *et al.* 2010; Velardi, Faralli and Navigli 2013) (*inter alia*).

We performed manual validation as follows. We first collected all disambiguated entries of the *wiki-p1.6* model (cf. Table 1), where these seventeen target words appear and randomly sampled 15% of these entries to make annotation feasible, resulting in a dataset of 2,884 ambiguous-related words in context. We restrict evaluation to the *wiki-p1.6* model for two reasons: an encyclopedic source is expected to provide better sense coverage *a priori*, thus providing us with more evaluation instances, while a low number of clusters is in line with findings that graph-based sense induction methods can produce rather coarse high-quality clusterings (Cecchini *et al.* 2017).

Table 7 presents statistics of our dataset: note that we gathered word senses of all parts of speech that correspond to the selected seventeen words, including verbs and adjectives, for the sake of completeness of our study. An annotator with previous experience in lexicographic annotation performed the judgment of the 2,884 contexts in a curated way (using several rounds of feedback on random samples of annotations). The annotator was presented with a table containing four columns: (i) the target word, (ii) a sense cluster defining the sense of the target word, (iii) a sense cluster that defines the context of the target word. The last column collected the binary answer on the question whether the 'definition' cluster is semantically compatible with the 'context' cluster. Table 8 illustrates two examples of semantically compatible and incompatible clusters. The reasons of incompatibility of sense clusters are either the absence of obvious semantic relations between the words in the clusters (cf. the 'planet' vs. 'basketball' sense of sun) or simply incoherence of one or both sense clusters – i.e. the case when the annotator cannot figure out the meaning of the sense denoted by a cluster, such as the case for the context cluster of tiger. The annotator was instructed to consider a sense cluster to be interpretable if it was possible to grasp a dominant meaning by looking at the top twenty words, while allowing for a small fraction of spurious terms (since sense clusters are automatically generated).

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 8. *Examples provided to the assessor participating in the study with correct judgments*

Word	Sense "Definition"	Sense "Context"	Related
Java	UNIX, Linux, Symbian, Unix, OS, Android, Mobile, Solaris, MS-DOS, Windows, iOS	screen, I/O, multiprocessor, IDE, repository, pak, Blu-ray, Graphics, Video, Itanium, ...	Yes
Python	hamster, lemurs, turtle, constrictor, lizard, orca, rhinoceros, cobra, ...	turtle, breeds, cattle, breed, cobra, Bulbul, Kingfisher, Mammals, starling, ...	Yes
Sun	Hearth, mirror, orb, soil, star, spotlight, temperature, water, solstice, burst, ...	eel, brave, Celtics, Wrangler, rockies, Chargers, Expos, Cavaliers, Cougars, padre, ...	No
Tiger	Macaque, deer, rhinoceros, Falcon, mascot, Whale, Gibbon, Hyena, boar, deer, ...	Nighthawk, Cessna, F-16, Valiant, Corsair, Maurer, Mirage, Reaper, Scorpion, ...	No

The subject was asked to determine if the first sense cluster (representing a sense definition of the ambiguous-related word) is semantically related to the second sense cluster (representing the context of the ambiguous-related word).

Results and discussion. The results of the experiment are summarized in Table 7.[10] Performance of the disambiguation procedure for the proper names and nouns ranges from 0.85 to 0.94, thus indicating an overall high quality of the procedure. Note that the word senses of adjectives and verbs are mostly the result of part-of-speech tagging errors, since in the seed set of seventeen words, we added only nouns and proper nouns. Wrongly tagged words have in general more noisy, uninterpretable clusters.

To better understand the amount of spurious items in our sense clusters, we performed an additional manual evaluation where, for a sample of hundred randomly sampled noun PCZ items, we counted the ratio between wrong (e.g. rat for the computer sense of mouse) and correct (keyboard, computer, etc.) related sense that were found within the PCZs. We obtained a macro average of 0.0495 and a micro average of 0.0385 wrongly related senses within the PCZs. Moreover, eighty-three per cent of the above sample has no unrelated senses, and only two per cent have only a single unrelated sense with a macro average ratio between the wrong and correct related PCZs of 0.067. This indicates that, overall, the amount of spurious senses within clusters is indeed small, thus providing a high-quality context for an accurate disambiguation of noun DT clusters.

6.3 Experiment 2: Linking induced senses to lexical resources

Experimental setting. In this experiment, we evaluate the performance of our linking component (Section 5.1). For this, we choose two lexical-semantic networks:

[10] The judgments are available for download (cf. Section 8).

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

WordNet (Fellbaum 1998), which has a high coverage on English common nouns, verbs and adjectives, and BabelNet (Navigli and Ponzetto 2012a), which also includes a large amount of proper nouns and senses gathered from multiple other sources, including Wikipedia.

We follow standard practices, e.g. Navigli and Ponzetto (2012a), and create five evaluation test sets, one for each dataset from Section 6.1, by randomly selecting a subset of 300 induced word senses for each dataset, and manually establishing a mapping from these senses to WordNet and BabelNet senses (senses that cannot be mapped are labeled as such in the gold standard).

We compare against the following two most frequent sense (MFS) baselines, which select from all the possible senses for a given term t:

(1) The MFS in WordNet, where frequencies of senses are observed on a manually annotated semantic concordance (Miller *et al.* 1993).
(2) The MFS in BabelNet. Since BabelNet combines WordNet and Wikipedia, this amounts to (i) the WordNet MFS for senses originally found in WordNet and (ii) the most cited (i.e. internally hyperlinked) Wikipedia page for senses derived from Wikipedia.

The quality and correctness of the mapping is estimated as accuracy on the ground-truth judgments, namely the amount of true mapping decisions among the total number of (potentially, empty) mappings in the gold standard. Each pair (j, c) in a mapping M created with Algorithm 2 is evaluated as (i) true positive (TP) when c is the most suitable sense in the LR for the induced word sense j; (ii) true negative (TN) when c refers to j itself and there are no senses in the LR to capture the meaning expressed by j; (iii) false positive (FP) when c is not the most suitable sense in the LR for the sense t; (iv) false negative (FN) when c refers to j itself and there is a sense in the LR that captures the same meaning of j.

We also evaluate our mapping by quantifying coverage and extra-coverage on the reference resource:

$$Coverage(A, B) = \frac{|A \cap B|}{|B|} \qquad ExtraCoverage(A, B) = \frac{|A/B|}{|B|} \tag{4}$$

where A is the set of LR synsets or induced word senses mapped in M using Algorithm 2, and B is the whole set of LR synsets. That is, Coverage indicates the percentage of senses of the LR sense inventory covered by the mapping M, whereas ExtraCoverage indicates the ratio of senses in M not linked to the LR sense inventory over the total number of senses in an LR. That is, ExtraCoverage is a measure of novelty to quantify the amount of senses discovered in T and not represented by the LR: it indicates the amount of 'added' knowledge we gain with our resource based on the amount of senses that cannot be mapped and are thus included as novel senses.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 9. *Results on linking to lexical semantic resource: number of linked induced word senses, Coverage, ExtraCoverage, accuracy of our method and of the MFS baseline for our five datasets*

	WordNet-linked				
PCZ	#linked senses	Cov.	ExtraCov.	Accuracy	MFS baseline
News-p1.6	88 k	34.5%	206.0%	86.9%	85.5%
News-p2.3	145 k	38.2%	267.0%	93.3%	85.0%
Wiki-p1.8	91 k	35.9%	234.7%	94.8%	80.5%
Wiki-p6.0	400 k	49.9%	919.9%	93.5%	74.2%
Wiki-mw-p1.6	81 k	30.7%	581.2%	95.3%	89.7%
	BabelNet-linked				
	# linked senses	Cov.	ExtraCov.	Accuracy	MFS baseline
News-p1.6	164 k	1.3%	2.9%	81.8%	52.3%
News-p2.3	236 k	1.4%	3.9%	85.1%	57.2%
Wiki-p1.8	232 k	1.9%	2.4%	86.4%	41.0%
Wiki-p6.0	737 k	2.8%	1.3%	82.2%	54.7%
Wiki-mw-p1.6	589 k	4.7%	1.8%	83.8%	59.4%

Results and discussion. In Table 9, we present the results using the optimal parameter values (i.e. $th = 0.0$ and $m = 5$ of Algorithm 2).[11] For all datasets, the number of linked senses, Coverage and ExtraCoverage are directly proportional to the number of entries in the dataset – i.e. the finer the sense granularity, as given by a lower sense clustering n parameter, the lower the number of mapped senses, Coverage and ExtraCoverage.

In general, we report rather low coverage figures: the coverage in WordNet is always lower than fifty per cent (thirty per cent in one setting) and coverage in BabelNet is in all settings lower than five per cent. Low coverage is due to different levels of granularities between the source and target resource. Our target LRs, in fact, have very fine-grained sense inventories. For instance, BabelNet lists seventeen senses of the word python including two (arguably obscure ones) referring to particular roller coasters. In contrast, word senses induced from text corpora tend to be coarse and corpus specific. Consequently, the low coverage comes from the fact that we connect a coarse and a fine-grained sense inventory – cf. also previous work (Faralli and Navigli 2013) showing comparable proportions between coverage and extra-coverage of automatically acquired knowledge (i.e. glosses) from corpora.

Finally, our results indicate differences between the order of magnitude of the Coverage and ExtraCoverage when linking to WordNet and BabelNet. This high difference is rooted in the cardinality of the two sense inventories, whereas BabelNet encompasses millions of senses, WordNet contains hundreds of thousands – many

[11] To optimize m, we prototyped our approach on a dev set consisting of a random sample of 300 senses, and studied the curves for the number of linked induced senses to WordNet resp. BabelNet. The th value was then selected as to maximize the accuracy.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

of them not covered in our corpora. Please note that an ExtraCoverage of about three per cent in BabelNet corresponds to about 300k novel senses. Overall, we take our results to be promising in that, despite the relative simplicity of our approach (i.e. almost parameter-free unsupervised linking), we are able to reach high accuracy figures in the range of around 87–95 per cent for WordNet and accuracies consistently above eighty per cent for BabelNet. This compares well against a random linking baseline that is able to achieve 44.2 per cent and 40.6 per cent accuracy on average when mapping to WordNet and BabelNet, respectively. Also, we consistently outperform the strong performance exhibited by the MFS baselines, which, in line with previous findings on similar tasks (Suchanek *et al.* 2008; Ponzetto and Navigli 2009) provide a hard-to-beat competitor. Thanks to our method, in fact, we are able to achieve an accuracy improvement over the MFS baseline ranging from 1.4 per cent to 14.3 per cent on WordNet mappings, and from 24.4 per cent to 45.4 per cent on BabelNet. Despite not being comparable, our accuracy figures are in the same ballpark as those reported by Navigli and Ponzetto (2012a) (cf. Table 1), who use a similar method for linking Wikipedia to WordNet.

Error analysis. To gain insights into the performance of our approach, as well as its limitations, we performed a manual error analysis of the output on the WordNet mappings, identifying a variety of sources of errors that impact the quality of the output resource. These include

- **part-of-speech tagging errors**, which may produce wrong senses such as non-existent 'verbs' (e.g. tortilla:VB) (about ten per cent of the errors);
- **Hearst patterns errors** that may extract wrong hypernyms such as issue for the entry emotionalism (about twenty per cent of the errors);
- **linking errors** where the accuracy strongly depends on the granularity of senses and relationships of the target LR (about seventy per cent of the errors).

More specifically, false positives are often caused by the selection of a synset that is slightly different from the most suitable one (i.e. semantic shift), whereas false negatives typically occur due to the lack of connectivity in the semantic network.

Even if the high values of the estimated accuracy (see Table 9) of our mapping approach indicate that we are generally performing well over all the classes of test examples (i.e. true positive, true negative, false positive and false negative), the performance figures exhibit a different order of magnitude between the count of true positives and true negatives. True negatives are senses in the ExtraCoverage that we estimate to be correct new senses not contained in the reference LR. For a sample of such senses, we performed an additional manual analysis, and identified the following reasons that explain our generally high ExtraCoverage scores:

- **Named entities and domain-specific senses** (about forty per cent of the true negatives): true negative senses are due to correct new senses not contained in the target LR. This holds in particular for WordNet, where encyclopedic content occurs in a spotty fashion in the form of a few examples for some classes.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 10. *Statistics and performance on typing unmapped PCZ items: number of induced senses counting for ExtraCoverage, number of typed and untyped induced senses, accuracy of our method and accuracy of the MFS baseline for our five datasets*

WordNet-linked					
PCZ	#extra senses	w types	w/o types	Accuracy	MFS baseline
News-p1.6	244 k	184 k	59 k	83.3%	80.7%
News-p2.3	316 k	226 k	90 k	91.4%	89.5%
Wiki-p1.8	277 k	225 k	51 k	89.2%	89.0%
Wiki-p6.0	1 M	675 k	487 k	81.2%	78.2%
Wiki-p1.6-mwe	683 k	538 k	144 K	78.8%	77.3%
BabelNet-linked					
PCZ	#extra senses	w types	w/o types	Accuracy	MFS baseline
News-p1.6	168 k	73 k	95 k	91.2%	87.2%
News-p2.3	225 k	89 k	135 k	90.3%	89.8%
Wiki-p1.8	208 k	143 k	65 k	87.2%	85.0%
Wiki-p6.0	1,4 M	278 k	1.1 M	41.2%	40.3%
Wiki-p1.6-mwe	552 k	342 k	209 k	89.6%	88.0%

- **Sense granularity misalignment** (about sixty per cent of the true negatives): true negatives that derive from excessively fine clustering, and should have been combined with other senses to represent a more generic sense.

6.4 Experiment 3: Typing of the unmapped induced senses

Experimental setting. The high ExtraCoverage rates from Section 6.3 show that our resource contains a large number of senses that are not contained in existing LRs such as WordNet and BabelNet. Besides, high accuracy scores in the evaluation of the quality of the sense clusters from Section 6.2 seem to indicate that such extra items are, in fact, of high quality. Crucially, for our purposes, information found among the extra coverage has enormous potential, e.g. to go beyond 'Wikipedia-only' sense spaces. Consequently, we next evaluate our semantic typing component (Section 5.2) to assess the quality of our method to include also these good extra clusters that, however, have no perfect mapping in the reference LR (WordNet, BabelNet).

Similarly to the experiments for the resource mapping (Section 6.3), we manually create five test sets, one for each dataset from Section 6.1, by randomly selecting 300 unmapped PCZ items for each dataset, and manually identifying the most appropriate type of each induced sense among WordNet or BabelNet senses. Given these gold standards, performance is then computed as standard accuracy on each dataset.

Results and discussion. In Table 10, we report the statistics and the estimated accuracy for the task of typing the previously unmapped senses found among

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

the ExtraCoverage. For each dataset and LR, we report the number of senses in the ExtraCoverage, the number of senses for which we inferred the type, the number of senses for which we were not able to compute a type, and the estimated accuracy for the types inferred by our method on the basis of either the links generated using our approach from Section 5.1, or those created using the MFS linking baseline. The results show that accuracy decreases for those datasets with higher polysemy. In particular, we obtain a low accuracy of 41.2 per cent for the 'wiki-p6.0' where the disambiguated thesaurus contains only a low number of related senses, resulting in sparsity issues. For the other settings, the accuracy ranges from 78.8 per cent to 91.4 per cent (WordNet) and from 87.2 per cent to 91.2 per cent (BabelNet). The MFS baseline accuracies of typing the unmapped induced senses (see Section 6.3) are lower, scoring 0.2 per cent to 2.7 per cent less accuracy for WordNet and 0.5 per cent to 4.2 per cent less accuracy for BabelNet: these results corroborate the previous ones on linking, where the MFS was shown to be a tough baseline. Besides, the higher performance figures achieved by the MFS on typing when compared to linking indicate that the typing task has a lower degree of difficulty in the respect that popular (i.e. frequent) types provide generally good type recommendations.

6.5 Experiment 4: Evaluation of enriched lexical semantic resources

Experimental setting. In our next experiment, we follow previous work (Navigli and Ponzetto 2012a) and benchmark the quality of our resources by making use of the evaluation framework provided by the SemEval-2007 task 16 (Cuadros and Rigau 2007) on the 'Evaluation of wide-coverage knowledge resources'. This SemEval task is meant to provide an evaluation benchmark to assess wide-coverage LRs on the basis of a traditional lexical understanding task, namely WSD (Navigli 2009). The evaluation framework consists of the following two main phases:

(1) **Generation of sense representations.** From each LR, sense representations, also known as 'topic signatures', are generated, which are sets of terms that are taken to be highly correlated with a set of target senses. In practice, a sense representation consists of a weighted vector, where each element corresponds to a term that is deemed to be related to the sense, and the corresponding weight quantifies its strength of association.
(2) **WSD evaluation.** Next, sense representations are used as weighted bags of words in order to perform monolingual WSD using a Lesk-like method (cf. Lesk 1986) applied to standard lexical sample datasets. Given a target word in context and the sense representations for each of the target word's senses, the WSD algorithm selects the sense with the highest lexical overlap (i.e. the largest number of words in common) between the sense representation and the target word's textual context.

This SemEval benchmark utilizes performance on the WSD task as an indicator of the quality of the employed LR. This approach makes it possible to extrinsically compare the quality of different knowledge resources, while making as few assumptions as possible over their specific properties – this is because knowledge resources

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 11. *Sample entries of the hybrid aligned resource (HAR) for the words* mouse *and* keyboard

PCZ ID	WordNet ID	PCZ related terms	PCZ context clues
mouse:0	mouse:wn1	rat:0, rodent:0, monkey:0, ...	rat:conj_and, gray:amod, ...
mouse:1	mouse:wn4	keyboard:1, computer:0, printer:0 ...	click:-prep_of, click:-nn, ...
keyboard:0	keyboard:wn1	piano:1, synthesizer:2, organ:0 ...	play:-dobj, electric:amod, ..
keyboard:1	keyboard:wn1	keypad:0, mouse:1, screen:1 ...	computer, qwerty:amod ...

Trailing numbers indicate sense identifiers. To enrich WordNet sense representations, we rely on related terms and context clues.

are simply viewed as sense representations, namely weighted bags of words. Besides, to keep the comparison fair, it uses a common and straightforward disambiguation strategy (i.e. Lesk-like word overlap) and a knowledge representation formalism (i.e. sense representations) that is equally shared across all LRs when evaluating them on the same reference dataset. Specifically, the evaluation is performed on two lexical sample datasets, from the Senseval-3 (Mihalcea, Chklovski and Kilgarriff 2004) and SemEval-2007 Task 17 (Pradhan *et al.* 2007) evaluation campaigns. The first dataset has coarse-grained and fine-grained sense annotations, while the second contains only fine-grained annotations. In all experiments, we follow the original task formulation and quantify WSD performance using standard metrics of recall, precision and balanced F-measure.

Here, we use the SemEval task to benchmark the 'added value' in knowledge applicable for WSD that can be achieved by enriching a standard resource like WordNet with disambiguated distributional information from our PCZs on the basis of our linking. To this end, we experiment with different ways of enriching WordNet-based sense representations with contextual information from our HAR. For each WordNet sense of a disambiguation target, we first build a 'core' sense representation from the content and structure of WordNet, and then expand it with different kinds of information that can be collected from the PCZ sense that is linked to it (cf. Table 11):

- **WordNet.** The baseline model relies solely on the WordNet LR. It builds a sense representation for each sense of interest by collecting synonyms and definition terms from the corresponding WordNet synset, as well as all synsets directly connected to it (we remove stop words and weigh words with term frequency).
- **WordNet + Related (news).** We augment the WordNet-based representation with related terms from the PCZ. That is, if the WordNet sense is linked to a corresponding induced sense in our resource, we add all related terms found in the linked PCZ sense to the sense representation.
- **WordNet + Related (news) + Context (news/wiki).** Sense representations of this model are built by taking the previously generated ones, and additionally including terms obtained from the context clues of either the news (+ Context (news)) or Wikipedia (+ Context (wiki)) corpora we use (see Section 6.1).

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

In the last class of models, we used up to 5,000 most relevant context clues per word sense. This value was set experimentally: performance of the WSD system gradually increased with the number of context clues, reaching a plateau at the value of 5,000. During aggregation, we excluded stop words and numbers from context clues. Besides, we transformed syntactic context clues to terms, stripping the dependency type, so they can be added to other lexical representations. For instance, the context clue rat:conj_and of the entry mouse:0 was reduced to the feature rat.

Table 12 shows a complete example from our dataset that demonstrate how, thanks to our HAR, we are able to expand WordNet-based sense representations with many relevant terms from the related terms of our sense-disambiguated PCZs.

We compare our approach to four state-of-the-art systems: KnowNet (Cuadros and Rigau 2008), BabelNet, WN+XWN (Cuadros and Rigau 2007) and NASARI. KnowNet builds sense representations based on snippets retrieved with a web search engine. We use the best configuration reported in the original paper (KnowNet-20), which extends each sense with twenty keywords. BabelNet in its core relies on a mapping of WordNet synsets and Wikipedia articles to obtain enriched sense representations; here, we consider both original variants used to generate sense representations, namely collecting all BabelNet synsets that can be reached from the initial synset at distance one (BabelNet-1) or two (BabelNet-2) and then outputting all their English lexicalizations. The WN + XWN system is the top-ranked unsupervised knowledge-based system of Senseval-3 and SemEval-2007 datasets from the original competition (Cuadros and Rigau 2007). It alleviates sparsity by combining WordNet with the eXtended WordNet (Mihalcea and Moldovan 2001). The latter resource relies on parsing of WordNet glosses. For all these resources, we use the scores reported in the respective original publications.

NASARI provides hybrid semantic vector representations for BabelNet synsets, which are a superset of WordNet. Consequently, we follow a procedure similar to the one we use to expand WordNet-only sense representations with information from our PCZs – namely, for each WordNet-based sense representation, we add all features from the lexical vector of NASARI that corresponds to it.[12]

Thus, we compare our method to three hybrid systems that induce sense representations on the basis of WordNet and texts (KnowNet, BabelNet, NASARI) and one purely knowledge-based system (WN + XWN). Note that we do not include the supervised TSSEM system in this comparison, as in contrast to all other considered methods including ours, it relies on a large sense-labeled corpus.

Results and discussion. Table 13 presents results of the evaluation, which generally indicate the high quality of the sense representations in our HAR. Expanding WordNet-based sense representations with distributional information provides, in fact, a clear advantage over the original representation on both Senseval-3 and SemEval-2007 datasets. Using related words specific (via linking) to a given WordNet

[12] We used the version of lexical vectors (July 2016) featuring 4.4 million of BabelNet synsets, yet covering only seventy two per cent of word senses of the two datasets used in our experiments.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 12. *WordNet-only and PCZ-enriched sense representations for the fourth WordNet sense of the word* disk *(i.e. the 'computer science' one): the 'core' WordNet sense representation is additionally enriched with related words from our hybrid aligned resource*

Model	Sense representation
WordNet-only	memory, device, floppy, disk, hard, disk, disk, computer, science, computing, diskette, fixed, disk, floppy, magnetic, disc, magnetic, disk, hard, disc, storage, device
WordNet + Related (wiki)	recorder, disk, floppy, console, diskette, handset, desktop, iPhone, iPod, HDTV, kit, RAM, Discs, Blu-ray, computer, GB, microchip, site, cartridge, printer, tv, VCR, Disc, player, LCD, software, component, camcorder, cellphone, card, monitor, display, burner, web, stereo, internet, model, iTunes, turntable, chip, cable, camera, iphone, notebook, device, server, surface, wafer, page, drive, laptop, screen, pc, television, hardware, YouTube, dvr, DVD, product, folder, VCR, radio, phone, circuitry, partition, megabyte, peripheral, format, machine, tuner, website, merchandise, equipment, gb, discs, MP3, hard-drive, piece, video, storage device, memory device, microphone, hd, EP, content, soundtrack, webcam, system, blade, graphic, microprocessor, collection, document, programming, battery, keyboard, HD, handheld, CDs, reel, web, material, hard-disk, ep, chart, debut, configuration, recording, album, broadcast, download, fixed disk, planet, pda, microfilm, iPod, videotape, text, cylinder, cpu, canvas, label, sampler, workstation, electrode, magnetic disc, catheter, magnetic disk, Video, mobile, cd, song, modem, mouse, tube, set, ipad, signal, substrate, vinyl, music, clip, pad, audio, compilation, memory, message, reissue, ram, CD, subsystem, hdd, touchscreen, electronics, demo, shell, sensor, file, shelf, processor, cassette, extra, mainframe, motherboard, floppy disk, lp, tape, version, kilobyte, pacemaker, browser, Playstation, pager, module, cache, DVD, movie, Windows, cd-rom, e-book, valve, directory, harddrive, smartphone, audiotape, technology, hard disk, show, computing, computer science, Blu-Ray, blu-ray, HDD, HD-DVD, scanner, hard disc, gadget, booklet, copier, playback, TiVo, controller, filter, DVDs, gigabyte, paper, mp3, CPU, dvd-r, pipe, cd-r, playlist, slot, VHS, film, videocassette, interface, adapter, database, manual, book, channel, changer, storage

sense provides substantial improvements in the results. Further expansion of sense representations with context clues (cf. Table 1) provides a modest improvement on the SemEval-2007 dataset only. Consequently, the results seem to indicate that context clues generally do not provide additional benefits over the expansions provided from the related terms of the linked PCZ items for task of generating sense representations.

On the Senseval-3 dataset, our hybrid models show better performance than all unsupervised knowledge-based approaches considered in our experiment. On the

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 13. *Comparison of our approach to the state of the art unsupervised knowledge-based methods on the SemEval-2007 Task 16 (weighted setting)*

	Senseval-3 fine-grained			SemEval-2007 fine-grained		
Model	Precision	Recall	F_1	Precision	Recall	F_1
Random	19.1	19.1	19.1	27.4	27.4	27.4
WordNet (WN)	29.7	29.7	29.7	44.3	21.0	28.5
WN + Related (news)	**47.5**	**47.5**	**47.5**	54.0	50.0	51.9
WN + Related (news) + Context (news)	47.2	47.2	47.2	54.8	51.2	52.9
WN + Related (news) + Context (wiki)	46.9	46.9	46.9	**55.2**	**51.6**	**53.4**
BabelNet-1	**44.3**	**44.3**	**44.3**	52.2	46.3	49.1
BabelNet-2	35.0	35.0	35.0	**56.9**	**53.1**	**54.9**
KnowNet	44.1	44.1	44.1	49.5	46.1	47.7
NASARI (lexical vectors)	32.3	32.2	32.2	49.3	45.8	47.5
WN + XWN	38.5	38.0	38.3	54.9	51.1	52.9

The best results per section (i.e. the ones using our resources vs. those from the previous literature) are boldfaced, the best results overall are underlined.

SemEval-2007 dataset instead, we perform on a par, yet slightly below BabelNet's best setting. Error analysis of the sense representations suggests that the extra performance of BabelNet on the SemEval data seems to derive from an aggressive graph-based expansion technique that leverages semantic relations harvested from Wikipedias in many languages – cf. the overall lower performance obtained by collecting sense representations from all Babel synsets at depth 1 only (BabelNet-1) vs. those that can be reached with two hops (BabelNet-2). This 'joint multilingual' approach has also been shown to benefit WSD in general (Navigli and Ponzetto 2012b), and represents an additional source of semantic information not present in our resource (we leave multilinguality for future work).

The results generally indicate the high quality of our HAR in a downstream application scenario, where we show competitive results while being much less resource intensive. This is because our method relies only on a relatively small LR like WordNet and raw, unlabeled text, as opposed to huge LRs like BabelNet or KnowNet. That is, while our method shows competitive results better or on a par with other state-of-the-art systems, it does not require access to web search engines (KnowNet), the structure and content of a very large collaboratively generated resource and texts mapped to its sense inventory (BabelNet, NASARI), or even a machine translation system or multilingual interlinked Wikipedias (BabelNet).

Related work on unsupervised WSD using the induced sense inventory. This article is focused on the HAR, e.g. a PCZ linked to an LR, as in the *knowledge-based* WSD experiment described above. However, the induced sense inventory is a valuable resource on its own and can be used to perform *unsupervised knowledge-free* WSD. In this case, the induced sense representations, featuring context clues, related words

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

and hypernyms, are used as features representing the induced senses. In our related experiments with sparse count-based (Panchenko *et al.* 2016, 2017c) and dense prediction-based (Pelevina *et al.* 2016) distributional models, we show that such unsupervised knowledge-free disambiguation models yield state-of-the-art results as compared to the unsupervised systems participated in SemEval 2013 and the AdaGram (Bartunov *et al.* 2016) sense embeddings model. An interactive demo that demonstrates our model developed in these experiments is described in Figure 2 and in Panchenko *et al.* (2017b).

7 Applications

Linked distributional disambiguated resources carry a great potential to positively impact many knowledge-rich scenarios. In this section, we leverage our resource for a few downstream applications of knowledge acquisition, namely (i) noise removal in automatically acquired knowledge graphs and (ii) domain taxonomy induction from scratch.

7.1 Linking knowledge resources helps taxonomy construction

Task definition. We examine a crucial task in learning a taxonomy (i.e. the *isa* backbone of a lexical-semantic resource) from scratch (Bordea *et al.* 2015, 2016), namely the induction of clean taxonomic structures from noisy hypernym graphs such as, for instance, those obtained from the extractions of hyponym–hypernym relations from text. In this task, we are given as input a list of subsumption relations between terms or, optionally, word senses – e.g. those from our PCZs (Figure 1) – which can be obtained, for instance, by exploiting lexical-syntactic paths (Hearst 1992; Snow, Jurafsky and Ng 2004), distributional representations of words (Baroni *et al.* 2012; Roller, Erk and Boleda 2014) or a combination of both (Shwartz *et al.* 2016). Due to the automatic acquisition process, such lists typically contain noisy, inconsistent relations – e.g. multiple inheritances and cycles – which do not conform to the desired, clean hierarchical structure of a taxonomy. Therefore, the task of *taxonomy construction* focuses on bringing order among these extractions, and on removing noise by organizing them into a directed acyclic graph (Kozareva and Hovy 2010).

Related work. State-of-the-art algorithms differ by the amount of human supervision required and their ability to respect some topological properties while pruning the noise. Approaches like those of Kozareva and Hovy (2010), Velardi *et al.* (2013) and Kapanipathi *et al.* (2014), for instance, apply different topological pruning strategies that require to specify the root and leaf concept nodes of the KB in advance – i.e. a predefined set of abstract top-level concepts and lower terminological nodes, respectively. The approach of Faralli, Stilo and Velardi (2015) avoids the need of such supervision with an iterative method that uses an efficient variant of topological sorting (Tarjan 1972) for cycle pruning. Such lack of supervision, however, comes at the cost of not being able to preserve the original connectivity between the

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

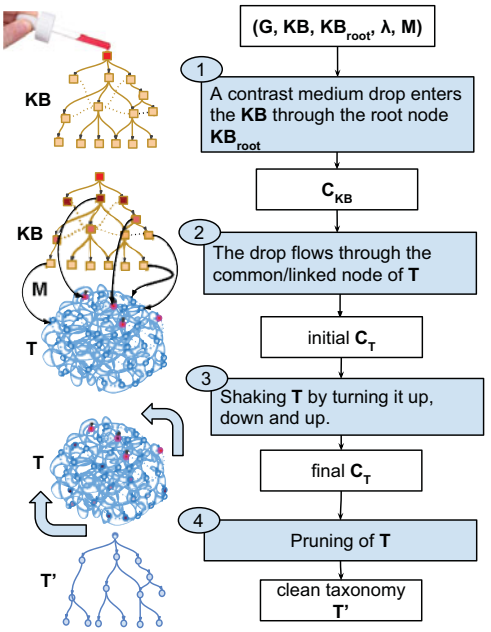


Fig. 5. (Colour online) The ContrastMedium (CM) algorithm for taxonomy construction.

top (abstract) and the bottom (instance) concepts. Random edge removal (Faralli *et al.* 2015), in fact, can lead to disconnected components, a problem shared with the OntoLearn Reloaded approach (Velardi *et al.* 2013), which cannot ensure such property when used to approximate a solution on a large noisy graph.

ContrastMedium algorithm. Links between heterogeneous knowledge resources, like those found within our HAR (Section 5), can be leveraged, together with a specialized algorithm, in order to advance the state of the art in taxonomy construction. To this end, we use ContrastMedium, a novel algorithm (Faralli *et al.* 2017) that is able to extract a clean taxonomy from a noisy knowledge graph without needing to know in advance – that is, having to manually specify – the top-level and leaf concepts of the taxonomy, while preserving the overall connectivity of the graph. ContrastMedium achieves this by projecting the taxonomic structure from a *reference taxonomy* (e.g. WordNet or the taxonomic *isa* backbone of BabelNet) onto a *target (noisy) hypernym graph* – for instance, the graph built from the set of hypernym relations in our HAR (Section 3) – on the basis of links found between the two resources, e.g. those automatically generated using our method from Section 5.

Metaphorically, in the context of clinical analysis, a contrast medium (CM) is injected into the human body to highlight specific complex internal body structures (in general, the cardiovascular system). In a similar fashion, our approach, which is summarized in Figure 5, starts by detecting the topological structure of the reference taxonomy by propagating a certain amount of CM that we initially inject through its root node (step 1). The highlighted structure indicates the distance of a node with respect to the root, with the lowest values of CM indicating the leaf terminological

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

nodes. The observed quantities are then transferred to corresponding nodes of the target hypernym graph by following the links between the two resources (step 2). Next, the medium is propagated by 'shaking' the noisy graph. We let the fluid reach all its components by alternating two phases of propagation: letting the CM flow via both incoming (shake up) and outgoing (shake down) edges (step 3). Finally, we use the partial order induced by the level of CM observed in each node to drive the pruning phase, and we 'stretch' the linked noisy knowledge graph into a proper taxonomy, namely a directed acyclic graph (step 4).

Evaluation. We benchmark ContrastMedium by comparing the quality of its output taxonomies against those obtained with the state-of-the-art approach of Faralli *et al.* (2015). The latter relies on Tarjan's topological sorting, which iteratively searches for a cycle (until no cycle can be found) and randomly removes an edge from it. We applied the two approaches to our linked resources and evaluated the performance on a three-way classification task to automatically detect the level of granularity of a concept. Pruning accuracy is estimated on the basis of a dataset of ground-truth judgments that were created using double annotation with adjudication from a random sample of 1,000 nodes for each noisy hypernym graph ($\kappa = 0.657$ (Fleiss 1971)). To produce a gold-standard, coders were asked to classify concepts from the random sample as (i) a root, top-level abstract concept – i.e. any of entity, object, etc. and more in general nodes that correspond to abstract concepts that we can expect to be part of a core ontology such as, for instance, DOLCE (Gangemi *et al.* 2002); (ii) a leaf terminological node (i.e. instances such as Pet Shop Boys) or (iii) a middle-level concept (e.g. celebrity), namely concepts not fitting into any of the previous classes.

We compute standard accuracy for each of the three classes. That is, we compare the system outputs against the gold standards and obtain three accuracy measures: one for the root nodes (A_R), one for the nodes 'in the middle' (A_M) and finally one for the leaf nodes (A_L). In Table 14, we show some of the results of the evaluation. Thanks to ContrastMedium, we are able to achieve, even despite the baseline already reaching very high performance levels (well above ninety per cent accuracy), improvements of up to six percentage points, with an overall error reduction between around forty per cent and sixty per cent. This performance improvements are due to the fact that ContrastMedium is able to (i) identify important topological clues among ground-truth taxonomic relations from the reference taxonomy and (ii) project them onto the noisy graph on the basis of the links found in the mapping between the two resources. That is, the availability of a mapping between knowledge resources helps us to project the supervision information from the clean source taxonomy into the target noisy graph without the need of further supervision. The reference taxonomy provides us with ground-truth taxonomic relations – this renders our method as knowledge-based, not knowledge-free. However, the availability of resources like, for instance, WordNet for English or the multilingual BabelNet implies that these requirements are nowadays trivially satisfied. The mapping, in turn can be automatically generated with high precision using any of the existing solutions for KB mapping, e.g. our algorithm from Section 5, or by relying on

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 14. *Pruning accuracy of the CM*

ContrastMedium			
	A_R	A_M	A_L
News-p1.6	**98.9%**	**98.3%**	**99.3%**
News-p2.3	**98.7%**	**98.7%**	**99.9%**
Wiki-p1.8	**97.6%**	**94.7%**	**97.3%**
Wiki-p6.0	**95.9%**	**94.3%**	**98.3%**
Tarjan (baseline)			
	A_R	A_M	A_L
News-p1.6	93.3%	94.6%	95.3%
News-p2.3	95.7%	94.7%	95.6%
Wiki-p1.8	93.1%	87.3%	94.1%
Wiki-p6.0	89.5%	90.1%	92.8%

ground-truth information from the Linguistic Linked Open Data cloud (Chiarcos, Hellmann and Nordhoff 2012).

What is perhaps the most interesting bit in our approach is the fact that by combining our unsupervised framework for knowledge acquisition from text (Section 3) with ContrastMedium, we are able to provide an *end-to-end solution for high-quality, unsupervised taxonomy induction from scratch*, i.e. without any human effort.

7.2 Inducing taxonomies from scratch using the hybrid aligned resource

Task definition. We now look at how ContrastMedium can be used as component within a larger system to enable end-to-end taxonomy acquisition from text. In general, taxonomy learning from scratch (Bordea *et al.* 2015, 2016) consists of the task of inducing a hypernym hierarchy from text alone (Biemann 2005), this typically starts with an initial step of finding hypernymy relations from texts, which is followed by a taxonomy construction phase in which local semantic relations are arranged together within a proper global taxonomic structure (cf. previous Section 7.1).

Related work. Existing approaches like Kozareva and Hovy (2010) (among others) use Hearst-like patterns (Hearst 1992) to bootstrap the extraction of terminological sister terms and hypernyms. Instead, in Velardi *et al.* (2013), the extraction of hypernymy relations is performed with a classifier, which is trained on a set of manually annotated definitions from Wikipedia (Navigli and Velardi 2010), being able to detect definitional sentences and to extract the definiendum and the hypernym. In these systems, the harvested hypernymy relations are then arranged into a taxonomy structure, e.g. using cycle pruning and 'longest path' heuristics to

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

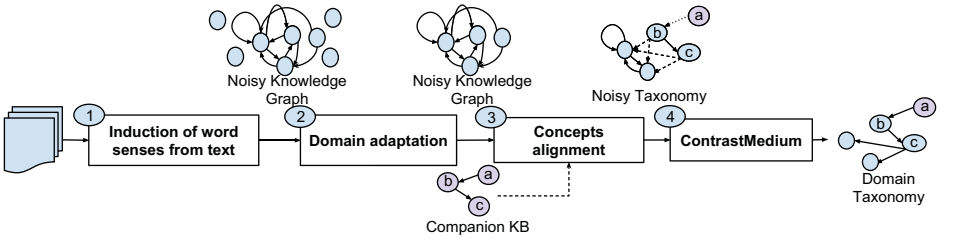


Fig. 6. (Colour online) Our full end-to-end pipeline for taxonomy induction from scratch.

induce a directed acyclic graph structure (Kozareva and Hovy 2010) or by relying on a variant of Chu-Liu Edmonds' optimal branching algorithm (Velardi *et al.* 2013). In general, all such lexical-based approaches suffer from the limitation of not being sense aware, which results in spurious taxonomic structures. Now, our fully disambiguated sense inventories could potentially overcome this problem and enable to a step forward toward the induction of high-quality, full-fledged taxonomies. In fact, we now show how the linked/semantic nature of our resources enables the development of a complete approach for taxonomy induction from scratch that achieves state-of-the-art performance with virtually no explicit supervision.

Using HAR to learn taxonomies from scratch. We now focus on the task of taxonomy induction by exploiting our HAR. Our approach is based on a five-stages pipeline as follows (see Figure 6):

(1) Create PCZs as described in Section 4.
(2) Filter out-of-domain concepts from our PCZs on the basis of domain-terminology-based heuristics. We construct domain-specific PCZs for a target domain by a simple lexical filtering. First, we build an extended lexicon of each domain on the basis of a seed vocabulary of the domain – i.e. domain terminologies such as those provided from the TExEval challenge (see below). Namely, for each seed term, we retrieve all semantically similar terms on the basis of the PCZ.
(3) Build an HAR by linking the PCZs to a companion taxonomy (e.g. WordNet, BabelNet, etc.) based on the methods from Section 5.
(4) Build a noisy hypernym graph by taking the union of the hypernym relations found within our PCZs.
(5) Apply ContrastMedium (Section 7.1) to remove noise from the graph and obtain a proper taxonomic structure.

That is, the combination of all our methods we presented so far provides us with a full end-to-end pipeline for taxonomy induction from scratch. Arguably, our approach is unsupervised in that it does not require any explicit human effort other than the knowledge encoded within the reference LRs. Both PCZs and links to reference knowledge resources are automatically induced in an unsupervised way. Moreover, links to existing LRs, as used in ContrastMedium, provide us with a source of knowledge-based supervision that is leveraged to clean PCZs and turn them into full-fledged taxonomies. More precisely, our framework is fully unsupervised up to

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

the linking part. However, unsupervised linking to a knowledge base and using the knowledge base for taxonomy construction indeed requires the knowledge base itself. To this end, we use freely available resources like WordNet and BabelNet. Given the linking, we can exploit the knowledge from these LRs, together with a knowledge-based method (ContrastMedium), without the need for additional human effort or supervision. That is, the fact that these lexical knowledge resources already exist and are publicly available implies that we can apply our framework with no extra human intervention.

Experiments. We use the evaluation benchmark from the most recent edition of the TExEval challenge (SemEval 2016 – task 13) (Bordea *et al.* 2016). Our experimental setting consists of the following components:

- **Three gold-standard taxonomies**, namely the FOOD's sub hierarchy of the Google products taxonomy,[13] as well as the NASEM[14] and EuroVoc[15] taxonomies of SCIENCE.
- **The task baseline**, which induces the taxonomy structure only from relations between compound terms such as juice, apple juice by applying simple substring inclusion heuristics. This baseline approach does not leverage any external or background knowledge and only uses the input domain terminology.
- **The Cumulative Fowlkes&Mallows evaluation measure** (CF&M): this enables the comparison of a system taxonomy against a gold standard at different levels of depth of the taxonomy, as obtained by penalizing errors at the highest cuts of the hierarchy (Velardi *et al.* 2012).
- **The task participant's systems**: (1) the JUNLP system (Maitra and Das 2016) makes use of two string inclusion heuristics combined with information from BabelNet; (2) the NUIG–UNLP system (Pocostales 2016) implements a semi-supervised method that finds hypernym candidates by representing them as distributional vectors (Mikolov, Yih and Zweig 2013); (3) the QASSIT system (Cleuziou and Moreno 2016) is a semi-supervised methodology for the acquisition of lexical taxonomies based on genetic algorithms. It is based on the theory of pretopology (Gil-Lafuente and Aluja 2012) that offers a powerful formalism to model semantic relations; (4) our task-winning TAXI system (Panchenko *et al.* 2016) that relies on combining two sources of evidence: substring matching and Hearst-like patterns. Hypernymy relations are extracted from Wikipedia, GigaWord, ukWaC, a news corpus and the CommonCrawl, as well as from a set of focused crawls; (5) the USAAR system (Tan, Bond and van Genabith 2016) exploits the hypernym endo/exocentricity (Brugmann 1904) as a practical property for hypernym identification.

[13] http://www.google.com/basepages/producttype/taxonomy.en-US.txt
[14] http://sites.nationalacademies.org/PGA/Resdoc/PGA_0445
[15] http://eurovoc.europa.eu/drupal/

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Table 15. *Comparison based on the SemEval 2016 task 13 benchmark for* FOODS *and* SCIENCES *domains. We report the Cumulative Fowlkes&Mallows measure*

	Google	NASEM	EuroVoc
System	FOODS	SCIENCES	SCIENCES
Baseline	0.0019	0.0163	0.0056
JUNLP	0.2608	0.1774	0.1373
NUIG-UNLP	–	0.0090	0.1517
QASSIT	–	0.5757	0.3893
TAXI	0.2021	0.3634	0.3893
USAAR	0.0000	0.0020	0.0023
WordNet	0.5870	0.5760	0.6243
Our approach	**0.6862**	**0.7000**	**0.8157**

- **A reference taxonomy**: we use WordNet for evaluation purposes by treating it the same way as any other participant system's output.

In Table 15, we report the results on the SemEval gold standards. Our approach significantly (χ^2 test, $p < 0.01$) outperforms all the other systems in all domains (i.e. Google FOOD, NASEM SCIENCE and EuroVoc SCIENCE), as well as the ground-truth taxonomy provided by WordNet. More importantly, the results indicate the overall robustness of our approach; that is, leveraging distributional semantics and symbolic knowledge (i.e. through linking to reference LRs) together is able to outperform not only the WordNet gold standard, which has limited coverage for fine-grained specific domains like these, but also the SemEval task participants, which all rely in some way or another on simple, yet powerful substring heuristics.

8 Conclusions

We have presented a framework for enriching lexical semantic resources, such as WordNet, with distributional information. Lexical semantic resources provide a well-defined semantic representation, but typically contain no corpus-based statistical information and are static in nature. Distributional semantic methods are well suited to address both of these problems, since models are induced from (in-domain) text and can be used to automatize the process of populating ontologies with new concepts. However, distributional semantic representations based on dense vectors have also major limitations in that they are uninterpretable on the symbolic level. By linking these representations to a reference LR, we can interpret them in an explicit way by means of the underlying relational knowledge model.

We provided a substantial investigation on enrichment of LRs with distributional semantics and evaluated the results in intrinsic and extrinsic ways showing that the resulting hybrid sense representations can be successfully applied to a variety of tasks that involve lexical-semantic knowledge. We tested the quality of the hybrid resources generated by our framework with a battery of intrinsic evaluations. Additionally,

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

we benchmarked the quality of our resource in a knowledge-based WSD setting, where we showed that our arguably low-resource approach – in that we rely only on a small LR like WordNet and raw, unlabeled text – reaches comparable quality with BabelNet, which in contrast is built on top of large amounts of high-quality collaborative content from Wikipedia. Finally, by combining distributional semantic vectors with links to a reference LRs, we are able to pave the way to the development of new algorithms to tackle hard, high-end tasks in knowledge acquisition like taxonomy cleaning and unsupervised, end-to-end taxonomy learning.

We believe that the hybrid LRs developed in our work will benefit high-end applications, e.g. ranging from entity-centric search (Lin *et al.* 2012; Schuhmacher, Dietz and Ponzetto 2015) all the way through full-fledged document understanding (Rospocher *et al.* 2016).

Downloads. We release all resources produced in this work under CC-BY 4.0 License[16]: (i) the PCZs resulting from our first experiment (Section 6.2); (ii) following the guidelines in McCrae, Fellbaum and Cimiano (2014), we created an RDF representation to share the mapping between our PCZs and lexical knowledge graphs (i.e. WordNet and BabelNet) (see Section 6.3) in the Linked Open Data Cloud; (iii) the types of the unmapped PCZ senses produced in the third experiment (see Section 6.4). All datasets, evaluation judgments, source code, and the demo can be accessed via `http://web.informatik.uni-mannheim.de/joint`.

References

Agirre, E., and Soroa, A. 2007. Semeval-2007 Task 02: evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*, Prague, Czech Republic.

Aprosio, A. P., Giuliano, C., and Lavelli, A. 2013. Extending the coverage of DBpedia properties using distant supervision over Wikipedia. In *Proceedings of the 2013 International Conference on NLP and DBpedia – Volume 1064 (NLP-DBPEDIA'13)*, Sydney, Australia.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India.

Baroni, M., Bernardi, R., Do, N.-Q., and Shan, C.-C. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France.

Baroni, M., Dinu, G., and Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, Baltimore, MD, USA.

Bartunov, S., Kondrashkin, D., Osokin, A., and Vetrov, D. P. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS'16)*, Cadiz, Spain.

Biemann, C. 2005. Ontology learning from text: a survey of methods. *LDV Forum* **20**(2): 75–93.

[16] `https://creativecommons.org/licenses/by/4.0/`

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

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, TextGraphs-1*, Brooklyn, NY, USA.

Biemann, C., and Riedl, M. 2013. Text: now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling* **1**(1): 55–95.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. 2009. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics* **7**(3): 154–65.

Bordea, G., Buitelaar, P., Faralli, S., and Navigli, R. 2015. SemEval-2015 task 17: taxonomy extraction evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'15)*, Denver, CO, USA.

Bordea, G., Lefever, E., and Buitelaar, P. 2016. SemEval-2016 task 13: taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

Bordes, A., Weston, J., Collobert, R., and Bengio, Y. 2011. Learning structured embeddings of knowledge bases. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, San Francisco, CA, USA.

Brugmann, K. 1904. *Kurze vergleichende Grammatik der indogermanischen Sprachen*. Strassburg, France: Karl J. Trubner.

Bryl, V., and Bizer, C. 2014. Learning conflict resolution strategies for cross-language Wikipedia data fusion. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14 Companion)*, Seoul, South Korea.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. 2015a. A unified multilingual semantic representation of concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Volume 1: Long Papers (ACL'15)*, Beijing, China.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. 2015b. NASARI: a novel approach to a semantically aware representation of items. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Denver, CO, USA.

Carlson, A., Betteridge, J., Kisiel, R., Settles, B., Hruschka, Jr., E. R., and Mitchell, T. M. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, GA, USA.

Cecchini, F. M., Riedl, M., and Biemann, C. 2017. Using pseudowords for algorithm comparison: an evaluation framework for graph-based word sense induction. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NODALIDA'17)*, Gothenburg, Sweden.

Chen, X., Liu, Z., and Sun, M. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, A meeting of SIGDAT, a Special Interest Group of the ACL (EMNLP'14)*, Doha, Qatar.

Chiarcos, C., Hellmann, S., and Nordhoff, S. 2012. Linking linguistic resources: examples from the open linguistics working group. In C. Chiarcos, S. Nordhoff, and S. Hellmann (eds.), *Linked Data in Linguistics – Representing and Connecting Language Data and Language Metadata*, pp. 201–16. Heidelberg, Germany: Springer.

Clark, S. 2015. Vector space models of lexical meaning. In S. Lappin, and C. Fox (eds.), *Handbook of Contemporary Semantics*, 2nd edition, pp. 493–522. New York, NY, USA: Wiley-Blackwell.

Cleuziou, G., and Moreno, J. G. 2016. QASSIT at SemEval-2016 Task 13: on the integration of semantic vectors in pretopological spaces for lexical taxonomy acquisition. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Cuadros, M., and Rigau, G. 2007. SemEval-2007 task 16: evaluation of wide coverage knowledge resources. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*, Prague, Czech Republic.

Cuadros, M., and Rigau, G. 2008. KnowNet: building a large net of knowledge from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1 (COLING'08)*, Manchester, UK.

Curran, J. R. 2002. Ensemble methods for automatic thesaurus extraction. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing – Volume 10 (EMNLP'02)*, Philadelphia, PA, USA.

De Marneffe, M.-C., MacCartney, B., and Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

Di Marco, A., and Navigli, R. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics* **39**(3): 709–54.

Evert, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis. Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Germany.

Fader, A., Soderland, S., and Etzioni, O. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, Edinburgh, UK.

Faralli, S., and Navigli, R. 2012. A new minimally supervised framework for domain word sense disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, Jeju Island, South Korea.

Faralli, S., and Navigli, R. 2013. Growing multi-domain glossaries from a few seeds using probabilistic topic models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, A meeting of SIGDAT, a Special Interest Group of the ACL (EMNLP'13)*, Seattle, WA, USA.

Faralli, S., Panchenko, A., Biemann, C., and Paolo Ponzetto, S. 2017. The contrast medium algorithm: taxonomy induction from noisy knowledge graphs with just a few links. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1: Long Papers (EACL'17)*, Valencia, Spain.

Faralli, S., Stilo, G., and Velardi, P. 2015. Large scale homophily analysis in twitter using a twixonomy. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, Buenos Aires, Argentina.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., and Smith, N. A. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Denver, CO, USA.

Faruqui, M., and Kumar, S. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Denver, CO, USA.

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

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5): 378.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. 2002. Sweetening ontologies with DOLCE. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web: 13th International Conference (EKAW 2002)*, Sigüenza, Spain, Berlin, Heidelberg.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. 2013. PPDB: the paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, Atlanta, GA, USA.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Gil-Lafuente, A. M., and Aluja, J. G. 2012. *Towards an Advanced Modelling of Complex Economic Phenomena: Pretopological and Topological Uncertainty Research Tools*. Berlin, Heidelberg: Springer.

Glavaš, G., and Ponzetto, S.-P. 2017. Dual tensor model for detecting asymmetric lexico-semantic relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, Copenhagen, Denmark.

Goikoetxea, J., Soroa, A., and Agirre, E. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Denver, CO, USA.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. 2012. UBY – a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics – Volume 2 (COLING'92)*, Nantes, France.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. 2013. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* **194**: 28–61.

Hovy, E., Navigli, R., and Ponzetto, S.-P. 2013. Collaboratively built semi-structured content and artificial intelligence: the story so far. *Artificial Intelligence* **194**: 2–27.

Jauhar, S. K., Dyer, C., and Hovy, E. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Denver, CO, USA.

Jenatton, R., Roux, N. L., Bordes, A., and Obozinski, G. 2012. A latent factor model for highly multi-relational data. In *Proceedings of Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems (NIPS'12)*, Lake Tahoe, NV, USA.

Jurgens, D., and Pilehvar, M. T. 2016. SemEval-2016 task 14: semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

Kapanipathi, P., Jain, P., Venkataramani, C., and Sheth, A. 2014. User interests identification on Twitter using a hierarchical knowledge base. In *Proceedings of the Semantic Web: Trends and Challenges: 11th International Conference (ESWC'14*), Cham.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics – Volume 1 (ACL'03)*, Sapporo, Japan.

Klein, P., Ponzetto, S. P., and Glavaš, G. 2017. Improving neural knowledge base completion with cross-lingual projections. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (EACL'17)*, Valencia, Spain.

Kozareva, Z., and Hovy, E. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, Cambridge, MA, USA.

Krause, S., Hennig, L., Gabryszak, A., Xu, F., and Uszkoreit, H. 2015. Sar-graphs: a linked linguistic knowledge resource connecting facts with language. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, Beijing, China.

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

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Levy, O., and Goldberg, Y. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, Baltimore, MD, USA.

Li, J., and Jurafsky, D. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, Lisbon, Portugal.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics – Volume 2 (COLING'98)*, Montreal, Quebec, Canada.

Lin, T., Pantel, P., Gamon, M., Kannan, A., and Fuxman, A. 2012. Active objects: actions for entity-centric search. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*, Lyon, France.

Maitra, P., and Das, D. 2016. JUNLP at SemEval-2016 Task 13: a language independent approach for hypernym identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

McCrae, J. P., Fellbaum, C., and Cimiano, P. 2014. Publishing and linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, Reykjavik, Iceland.

Mihalcea, R., Chklovski, T., and Kilgarriff, A. 2004. The Senseval-3 English lexical sample task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, Barcelona, Spain.

Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, Lisbon, Portugal.

Mihalcea, R., and Moldovan, D. I. 2001. eXtended WordNet: progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems (NIPS'13)*, Lake Tahoe, NV, USA.

Mikolov, T., Yih, W.-T., and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Atlanta, GA, USA.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, Princeton, NJ, USA.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. 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: Vol. 2 (ACL'09)*, Suntec, Singapore.

Nastase, V., and Strube, M. 2013. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence* **194**: 62–85.

Navigli, R. 2009. Word sense disambiguation: a survey. *ACM CSUR* **41**(2): 1–69.

Navigli, R., and Ponzetto, S. P. 2012a. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* **193**: 217–50.

Navigli, R., and Ponzetto, S. P. 2012b. Joining forces pays off: multilingual joint Word Sense Disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, Jeju Island, South Korea.

Navigli, R., and Velardi, P. 2010. Learning Word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10*), Uppsala, Sweden.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Neelakantan, A., Shankar, J., Passos, A., and McCallum A. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, Doha, Qatar.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* **104**(1): 11–33.

Niemann, E., and 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 (IWCS'11)*, Oxford, UK.

Nieto Piña, L., and Johansson, R. 2016. Embedding senses for efficient graph-based word sense disambiguation. In *Proceedings of the 2016 Workshop on Graph-based Methods for Natural Language Processing (Textgraphs'16)*, San Diego, CA, USA.

Norvig, P. 2016. The semantic web and the semantics of the web: where does meaning come from? In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*, Montreal, Quebec, Canada.

Padó, S., and Lapata M. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* **33**(2): 161–99.

Panchenko, A. 2016. Best of both worlds: making word sense embeddings interpretable. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Panchenko, A., Faralli, S., Ponzetto, S. P., and Biemann, C. 2017a. Using linked disambiguated distributional networks for word sense disambiguation. In *Proceedings of the 1st EACL Workshop on Sense, Concept and Entity Representations and their Applications*, Valencia, Spain.

Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C., Ponzetto, S. P., and Biemann, C. 2016. TAXI at SemEval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

Panchenko, A., Marten, F., Ruppert, E., Faralli, S., Ustalov, D., Ponzetto, S. P., and Biemann, C. 2017b. Unsupervised, knowledge-free, and interpretable word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP'17)*, Copenhagen, Denmark.

Panchenko, A., and Morozova, O. 2012. A study of hybrid similarity measures for semantic relation extraction. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, Avignon, France.

Panchenko, A., Romanov, P., Morozova, O., Naets, H., Philippovich, A., Romanov, A., and Fairon, C. 2013. Serelex: search and visualization of semantically related words. In *Proceedings of the European Conference on Information Retrieval (ECIR'13)*, Moscow, Russia.

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., and Biemann C. 2017c. Unsupervised does not mean uninterpretable: the case for word sense induction and disambiguation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, Valencia, Spain.

Panchenko, A., Simon, J., Riedl, M., and Biemann, C. 2016. Noun sense induction and disambiguation using graph-based distributional semantics. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS'16)*, Bochum, Germany.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. 2011. *English Gigaword,* 5th ed. Philadelphia: Linguistic Data Consortium.

Pavel, S., and Euzenat, J. 2013. Ontology matching: state of the art and future challenges. *IEEE Transaction on Knowledge and Data Engineering* **25**(1): 158–76.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Pedersen, T., Patwardhan, S., and Michelizzi, J. 2004. WordNet::similarity – measuring the relatedness of concepts. In *Proceedings of the HLT-NAACL 2004: Demonstration Papers (HLT-NAACL'04 Demos)*, Boston, MA, USA.

Pelevina, M., Arefiev, N., Biemann, C., and Panchenko, A. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, Berlin, Germany.

Pennington, J., Socher, R., and Manning, C. 2014. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, Doha, Qatar.

Pham, N. T., Lazaridou, A., and Baroni, M. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Volume 2: Short Papers (ACL'15)*, Beijing, China.

Pocostales, J. 2016. NUIG-UNLP at SemEval-2016 task 13: a simple word embedding-based approach for taxonomy extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

Ponzetto, S. P., and Navigli, R. 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*, Pasadena, CA, USA.

Ponzetto, S. P., and Navigli, R. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, Uppsala, Sweden.

Ponzetto, S. P., and Strube, M. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence* **175**: 1737–56.

Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. 2007. SemEval-2007 task 17: english lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*, Prague, Czech Republic.

Reisinger, J., and Mooney, R. J. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10)*, Los Angeles, CA, USA.

Richter, M., Quasthoff, U., Hallsteinsdóttir, E., and Biemann, C. 2006. Exploiting the Leipzig corpora collection. In *Proceedings of IS-LTC'06*, Ljubljana, Slovenia.

Riedl, M. 2016. *Unsupervised Methods for Learning and Using Semantics of Natural Language*. Ph.D. thesis. Germany: TU Darmstadt. http://tuprints.ulb.tu-darmstadt.de/5435/.

Riedl, M., and Biemann, C. 2013. Scaling to large^3 data: an efficient and effective method to compute distributional thesauri. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, a Meeting of SIGDAT, a Special Interest Group of the ACL (EMNLP'13)*, Seattle, WA, USA.

Riedl, M., and Biemann, C. 2015. A single word is not enough: ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, Lisbon, Portugal.

Roller, S., Erk, K., and Boleda, G. 2014. Inclusive yet selective: supervised distributional hypernymy detection. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING'14)*, Dublin, Ireland.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics* **37**–**38**: 132–51.

Rothe, S., and Schütze, H. 2015. AutoExtend: extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

Processing of the Asian Federation of Natural Language Processing, Volume 1: Long Papers (ACL'15), Beijing, China.

Ruppert, E., Kaufmann, M., Riedl, M., and Biemann, C. 2015. JoBimViz: a web-based visualization for graph-based distributional semantic models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, System Demonstrations (ACL'15)*, Beijing, China.

Schuhmacher, M., Dietz, L., and Ponzetto, S. P. 2015. Ranking entities for web queries through text and knowledge. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM'15)*, Melbourne, Australia.

Shwartz, V., Goldberg, Y., and Dagan, I. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers (ACL'16)*, Berlin, Germany.

Snow, R., Jurafsky, D., and Ng, A. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL'06)*, Sydney, Australia.

Snow, R., Jurafsky, D., and Ng, A. Y. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of the Advances in Neural Information Processing Systems 17 (NIPS'04*), Vancouver, BC, Canada.

Socher, R., Chen, D., Manning, C. D., and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems (NIPS'13)*, Lake Tahoe, NV, USA.

Suchanek, F. M., Kasneci, G., and Weikum, G. 2008. YAGO: a large ontology from Wikipedia and WordNet. *Journal of Web Semantics* **6**(3): 203–17.

Tan, L., Bond, F., and van Genabith, J. 2016. USAAR at SemEval-2016 task 13: hyponym endocentricity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT'16)*, San Diego, CA, USA.

Tarjan, R. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing* **1**(2): 146–60.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**: 141–88.

Van de Cruys, T. 2010. Mining for meaning: the extraction of lexico-semantic knowledge from text. *Groningen Dissertations in Linguistics* **82**.

Van Dongen, S. 2008. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications* **30**(1): 121–41.

Velardi, P., Faralli, S., and Navigli, R. 2013. OntoLearn reloaded: a graph-based algorithm for taxonomy induction. *Computational Linguistics* **39**(3): 665–707.

Velardi, P., Navigli, R., Faralli, S., and Ruiz-Martínez, J. M. 2012. A new method for evaluating automatically learned terminological taxonomies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Véronis, J. 2004. HyperLex: lexical cartography for information retrieval. *Computer Speech and Language* **18**: 223–52.

Wang, Z., Li, J., Wang, Z., and Tang, J. 2012. Cross-lingual knowledge linking across Wiki knowledge bases. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*, Lyon, France.

Weeds, J., Weir, D., and McCarthy, D. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, Geneva, Switzerland.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press

West, R., Gabrilovich, E., Murphy, K., Sun, S., Gupta, R., and Lin, D. 2014. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, Seoul, South Korea.

Wu, W., Li, H., Wang, H., and Zhu, K. Q. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD'12)*, Scottsdale, AZ, USA.

Yu, M., and Dredze, M. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers (ACL'14)*, Baltimore, MD, USA.

https://doi.org/10.1017/S135132491700047X Published online by Cambridge University Press