Learning Taxonomies of Concepts and not Words using Contextualized Word Representations: A Position Paper

Lukas Schmelzeisen[1] and **Steffen Staab**[1,2]
[1]Institute for Web Science and Technologies (WeST), University of Koblenz-Landau, Germany
[2]Web and Internet Science Group (WAIS), University of Southampton, United Kingdom
lukas@uni-koblenz.de, staab@uni-koblenz.de

Abstract

Taxonomies are semantic hierarchies of concepts. One limitation of current taxonomy learning systems is that they define concepts as single words. This position paper argues that contextualized word representations, which recently achieved state-of-the-art results on many competitive NLP tasks, are a promising method to address this limitation. We outline a novel approach for taxonomy learning that (1) defines concepts as synsets, (2) learns density-based approximations of contextualized word representations, and (3) can measure similarity and hypernymy among them.

1 Introduction

A *taxonomy* is a hierarchical representation of semantic knowledge as a set of *concepts* (or *senses*) and a directed acyclic graph of hyponym–hypernym (is-a) relations among them. They have been beneficial in a variety of tasks that require semantic understanding, such as document clustering (Hotho, Staab, and Stumme 2003), query understanding (Hua et al. 2017), questions answering (Yang et al. 2017), and situation recognition (Yatskar et al. 2017). Manually curated taxonomies, such as the one contained in WordNet (Miller 1995), are highly precise, but limited in coverage of both rare concepts and specialized domains. Therefore, one line of research has focused on creating taxonomies automatically from natural language corpora, which are widely available for many domains and languages.

One major limitation of almost all taxonomy learning system is that they do not distinguish between words and concepts. In general, however, a many-to-many relationship holds between them. For example, the word "Venus" may either refer to the concept of a particular planet or to a Roman goddess, and "morning star" can either refer to the same planet or to a type of weapon. Instead, most automatically constructed taxonomies conflate the different senses of a word and typically only learn the most predominant one.

This limitation is reinforced by context-free word representations[1] (Turney and Pantel 2010; Mikolov et al. 2013; Pennington, Socher, and Manning 2014): methods that encode each word as one point in a vector space of meaning and are thereby unable to account for multiple senses of a word. Such representations are widely used in taxonomy learning and many other disciplines of natural language processing (NLP). In 2018, an emerging trend in NLP have been task-independent deep neural network architectures based on language model pre-training, which have achieved state-of-the-art results in a number of competitive disciplines, such as questions answering or natural language inference (Peters et al. 2018; Radford et al. 2018; Devlin et al. 2018). One quality that is common to all of these systems is that they allow for contextualized word representation: depending on their contexts, occurrences of the same word can be mapped to the same, similar or very different points in the vector space.

This position paper argues that contextualized representations, besides being a powerful input representations for machine learning, provide a promising approach for distinguishing between the different concepts a word can refer to. We propose to represent concepts by probability density estimates that approximate all points in the vector space that correspond to occurrences of the respective word sense, and to detect whether two concepts are in a taxonomic relation via divergence between their distributions. Additionally, our concept representations make it possible to transfer the strength of contextualized word representations to scenarios where no context information is available.

The remainder of this paper reviews the related work in taxonomy learning and word representation (Section 2), presents a brief explorative analysis of the lexical semantics encoded by contextualized word representations (Section 3), details our idea of concept representations for taxonomy learning (Section 4), and concludes (Section 5).

2 Related Work

2.1 Taxonomy Learning

Taxonomy learning typically consists of at least the subtasks term extraction, hypernym detection, and taxonomy construction (Maedche and Staab 2001). A recent survey is provided by Wang, He, and Zhou (2017).

Term Extraction The goal of this first subtask is to automatically find seed words that are specific to the do-

[1]The NLP literature uses both *distributed representations* and *distributional representations* to refer to this concept. The indent is usually to distinguish between latent, dense vectors (low-dimensional) and sparse vectors (high-dimensional), respectively.

main over which the taxonomy should be constructed. The domain is usually specified latently via a set of domain-specific documents. Most extraction approaches find candidate (multi-) words via part-of-speech patterns and filter them using statistical measures that aim to estimate theoretical principles such as salience, relevance, or cohesion (Sclano and Velardi 2007; Periñán-Pascual 2018).

Hypernym Detection This next task consists of the identification of hypernym–hyponym pairs over the seed vocabulary (and possibly additional words). Pattern-based approaches rely on word pairs occurring in specific lexicosyntactic patterns in a corpus (Roller, Kiela, and Nickel 2018). They achieve high precision but suffer from low recall because they rely on both words occurring together in context. A second type of approach is based on vector representations of words, and is thus not reliant on local context. In the unsupervised case, usually a score is assigned to each possible word pair, which is expected to be higher for hypernym pairs than for negative instances. Many competing methods exist of which none is clearly superior to others; they differ both in the construction of the vector space as well as the employed detection measure (Santus, Shwartz, and Schlechtweg 2017; Chang et al. 2018). Supervised approaches typically achieve better results on existing benchmarks, but have been shown to not truly detect hypernymy, but rather which words are prototypical hypernyms, making them unreliable for real-world applications (Levy et al. 2015; Santus, Shwartz, and Schlechtweg 2017). Pattern- and vector-based approaches have been integrated successfully by Shwartz, Goldberg, and Dagan (2016) using LSTM neural networks.

Taxonomy Construction In this subtask the final taxonomic hierarchy is constructed from identified hypernym pairs. This is non-trivial because identified hypernym pairs are noisy and basically never induce a connected directed acyclic graph. Approaches can be divided into clustering-based (de Knijff, Frasincar, and Hogenboom 2013) and graph-based ones (Velardi, Faralli, and Navigli 2013; Gupta et al. 2017). Afterwards, clean-up operations like cycle elimination are sometimes performed (Liang et al. 2017a).

Discussion To the best of our knowledge, there is no taxonomy learning system providing a principled way to model multiple word senses. At most, existing system consider in which grammatical role a word occurs in (Santus, Shwartz, and Schlechtweg 2017), which is often a bad signal for separating word senses, or explicitly try to word sense disambiguate the input corpus in a preprocessing stage (Klapaftis and Manandhar 2010), which is often error-prone. As a result, automatically created taxonomies conflate word senses and important semantic properties like transitivity of the hypernymy relation do not usually hold (Liang et al. 2017b).

2.2 Word Representations

Context-free Word Representations Most *context-free word representations* map each word to a single point in an often latent vector space of meaning—regardless of the context the word is used in—with the goal to place words with similar meaning close to each other (Turney and Pantel 2010). After groundbreaking results achieved by factorizing word co-occurrence matrices (Mikolov et al. 2013; Pennington, Socher, and Manning 2014), such representations have been used to encode words in most state-of-the-art NLP architectures since. These representations are only able to represent one sense per word as they map each word to exactly one vector. One strand of research has thus focused on mapping each word to multiple vectors that should each represent a different word sense (Camacho-Collados and Pilehvar 2018). The number of senses per word have been either learned by clustering contexts, in a preceding step (Huang et al. 2012) and jointly during model training (Neelakantan et al. 2014), or specified via a given sense inventory (Chen, Liu, and Sun 2014). A related line of work criticizes the assumption that all meanings of a word can be discretely separated, and instead represent each word as probability densities in the vector space—usually as Gaussian distributions (Vilnis and McCallum 2015; Athiwaratkun and Wilson 2017). The general meaning of a word is then characterized by the mean vector of the distribution analogous to before, while additionally the covariance can be interpreted as spread of meaning or uncertainty. We still classify these advanced representation approaches as context-free, because even though their main training signal is what contexts a word occurs in, they are unable to adapt or select the representation for a word given a specific context. Possibly as a consequence of this, they have seen almost no adoption in practice. Specifically, we are not aware of any taxonomy learning approaches building on these representations.

Contextualized Word Representations Until very recently, almost all state-of-the-art solutions in NLP were highly specialized task-specific architectures. In contrast, in 2018 three related task-agnostic architectures have been published that achieved state-of-the-art results across a wide range of competitive tasks, suggesting some generalizable language understanding for the first time. All of these systems built on the language modeling objective: training a model to predict a word given its surrounding context. Because such training examples can be built from unlabeled corpora, much training data is available. ELMo (Peters et al. 2018) trains representations with stacked bidirectional LSTMs, but still employs task-specific architectures on top of them. OpenAI GPT (Radford et al. 2018) and BERT (Devlin et al. 2018) do away with this and instead train task-agnostic transformer stacks that are only fine-tuned together with a single dense layer for each downstream task. The latter mainly improves upon the former by joint conditioning on both preceding and following contexts. Critically, all systems allow for *contextualized word representation*: they map each word occurrence to a vector specifically considering the surrounding context. Much of their success is attributed to the ability to better disambiguate polysemous words in a given sentence. This representation approach is easily applicable for many NLP tasks, where inputs are usually sentences and context information is thus available. However, due to the hierarchical nature of taxonomies, there is no straightforward way to utilize the generalization power of contextualized representations for our task.

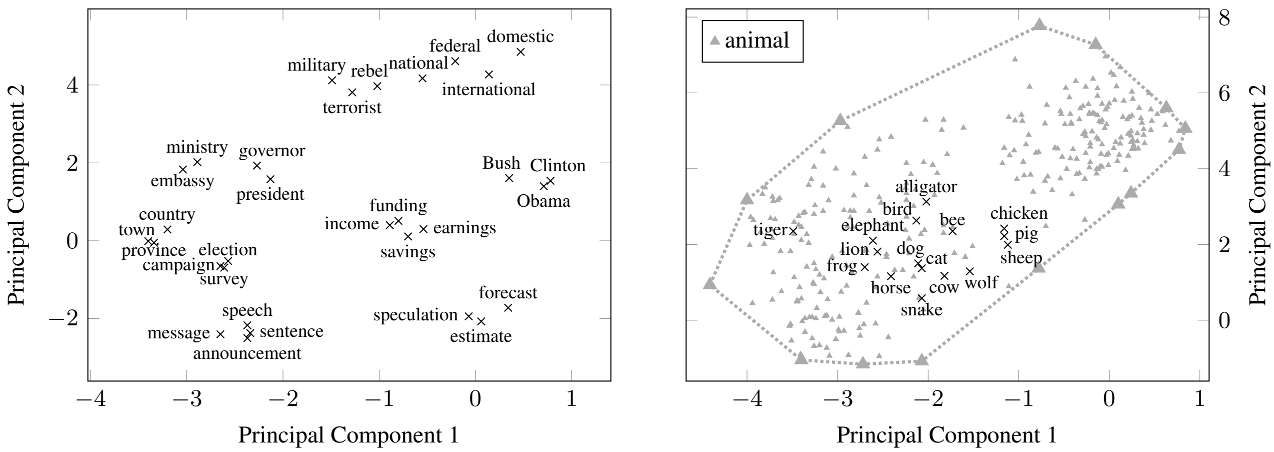

Figure 1: PCA projections of 1024-dimensional ELMo vectors (all biLM layers averaged). Shown words were selected manually. X's indicate the projection of the centroid of all representations for their respective label. Additionally, in the right panel, triangles are the projections of all representations of the word "animal" together with their convex hull as a dotted line.

3 Explorative Analysis

We present a brief explorative analysis of the lexical semantics encoded by contextualized word representations here, to prepare the argument in Section 4.

Figure 1 shows two visualizations of ELMo representations. For this, the pre-trained model `elmo_2x4096_512_2048cnn_2xhighway`[2] was used to compute all contextualized word representations on the first percent (roughly 200 000 different words and 7 500 000 word occurrences) of the corpus by Chelba et al. (2014), a corpus of news paper articles on which the model was originally trained on by Peters et al. (2018).

We observe a number of interesting properties:

- The centroids of all ELMo representations of a word tend to show strong local clustering behavior based on word similarity. Specifically, in the left panel of Figure 1 this is visible for a number of words from the political domain. Such behavior is not immediately relevant to taxonomy learning, but since encoding similarity is the main goal of context-free word representations (Turney and Pantel 2010) this suggest a strong representational power for ELMo, that is preserved when only considering centroids. Clustering was even better visible when visualizing with t-SNE (van der Maaten and Hinton 2008), but we only show PCA projections here for consistently with the next point.
- The representations of hypernym words tend to exhibit a larger spread in the vector space than the representations of their hyponyms. This motivates our approach for detecting hypernymy using contextualized representations, which we detail in the next section. Specifically, we visualize this behavior in the right panel of Figure 1, where we show that the centroids of words for different animal species lie inside the convex hull of the representations of "animal". Surprisingly, this behavior was not visible when visualizing with t-SNE. We suspect that this it the case because of an implementation detail of ELMo, namely the final contextualized representations being derived from context-free ones, and that thereby the contextualized representations are still separable in high-dimensional space, which is the optimization objective of t-SNE.
- For words with multiple senses, the representations of individual senses tend to lie close to clusters of other words that are similar in meaning to the respective sense. This suggests their suitability for discovering synsets of words, as we propose in the next section. Because of limited space, we do not show this behavior here, but Stanovsky and Hopkins (2018) made similar observations.

[2] Available at: `https://allennlp.org/elmo`

4 Concept Representations

In this section, we outline our idea for concepts representations from contextualized word representations. Their aim is to address the current limitations of current taxonomy learning systems, namely that they do not distinguish between words and concepts, which we believe to be a major hindering factor for learning accurate taxonomies. Our approach mainly consists of changes in the subtasks of term extraction and hypernymy detection whereas for constructing the final taxonomies existing techniques can be used without major modification.

For term extraction, our idea reuses existing work for finding domain-specific seed words by estimating the relevance of words to the target domain (Periñán-Pascual 2018). Going beyond previous work though, we propose to learn synsets of word senses for defining concepts instead of just defining them by single words. A *synset* is a set of different words each sharing a common sense interpretation. They have been popularized by WordNet (Miller 1995), and are a psychologically plausible definition for concepts, as humans are usually able to infer the specific sense meant as the intersection of all words in a synset (Stanovsky and Hopkins 2018). For finding synsets, we propose to (1) calculate the contextualized word representation vectors for all word occurrences in the given corpus, (2) group vectors using a clustering algorithm, and, for retaining only domain-specific concepts, to optionally (3) filter out all clusters not containing at least one seed word as previously determined via term extraction. Each resulting cluster of word vectors then constitutes one synset. We leave open what clustering algorithm should be chosen for this, although simple ones like k-means clustering might already be sufficient. Here, the parameter k would control into how many concepts the vector space should be separated. Indeed, k-means clustering of ELMo representations has been shown to work well by Stanovsky and Hopkins (2018). This is also the only work in this direction that we are aware of: the authors cluster all ELMo vectors belonging to a specific word to determine the number of senses of that word, but do not investigate properties of clustering all ELMo vectors of a corpus. Alternatively, to avoid choosing a parameter k Dirichlet processes could be used.

The result of the previous step is a set of clusters (of contextualized word representation vectors) each characterizing one synset/concept. For determining similarity and hypernymy among these concepts as well as for a more parameter efficient representation, we propose to learn what we call *concept representations*: probability density estimates that approximate all vectors belonging to one concept. In continuation of current research (Vilnis and McCallum 2015; Athiwaratkun and Wilson 2017), we specifically suggest to use Gaussian distributions although others are conceivable, such as Student's t-distribution. For Gaussians, the mean vector is usually interpreted as characterizing the general meaning of a word whereas the covariance signifies the generality/unspecificity of the word. Following Vilnis and McCallum (2015), the mean and covariance of the density can be found by averaging all vectors of a synset or calculating the empirical covariance among them, respectively. Athiwaratkun and Wilson (2017) suggest to learn multimodal Gaussian distributions, but we deem this unnecessary in our case, since the input vectors for each distribution should already be comparatively close together as a result of the preceding clustering step.

For context-free representations, similarity of words is usually estimated using the dot product of the respective vectors. In an analogous way, similarity for density-based representations can be estimated via the inner product (probability product kernel) of the respective densities (Vilnis and McCallum 2015). Our method for determining hypernymy is motivated by the *distributional inclusion hypothesis* (Geffet and Dagan 2005), which states that a hypernym–hyponym relation holds among word senses exactly when the hypernym sense can occur in all the contexts the hyponym sense can occur in. Since contextualized word representations characterize exactly which semantic context holds for a given word occurrence, our concept representations should characterize which semantic contexts a concept can occur in. There, we formulate the problem of detecting hypernymy among two concepts as to what degree one concept's density is "included" in the other one's. Athiwaratkun and Wilson (2018) survey a number of measures for this, the most well-known being the Kullback-Leibler divergence. One advantage of this approach over conventional ones is that it naturally allows to characterize the strength of the hypernymy relation, as motivated by Vulic et al. (2017).

4.1 Caveats

Our approach requires computing and storing the contextualized word representations of all word occurrences for a training corpus with a given pre-trained model, which is already resource intensive[3]. Training a new model from scratch is orders of magnitude more expensive.

Additionally, being able to handle multiword expressions (MWEs), such as "morning star", is of critical importance to taxonomy learning techniques, since most non-trivial concepts do not have single word identifiers. There has been no research so far on finding a single vector encoding for a given MWE from the contextualized word representation of its constituent words, which is hard because the semantics of most MWEs are non-compositional. However, this is not a fundamental problem to our approach since MWEs can just be mapped to single words, such as "morning_star" in a preprocessing step before training the contextualized word representations. The same technique is commonly performed for word2vec (Mikolov et al. 2013). Though this means, that no pre-trained model could be used.

5 Conclusion

We have sketched an approach for learning taxonomies using contextualized word representations that distinguishes words and concepts. Our outlined idea is novel in that we are the first to propose (1) defining concepts in automatically constructed taxonomies as synsets, which we plan to learn by clustering contextualized word representations, (2) approximating a set of related contextualized word representations via probability density estimates, which we call concept representations, and (3) using such representations to determine similarity and hypernymy among concepts.

Further, our concept representations provide the opportunity to study other interesting semantic relations among concepts, for example, modeling union and intersection of concepts via the union and intersection of their densities, respectively. Additionally, they could find usage in a number of NLP tasks as substitutes for the contextualized word representations that they approximate, which take comparatively long times to compute even on powerful GPUs, since our representations are far more parameter efficient.

Beyond our proposed idea of using contextualized word representations for taxonomy learning, which is in continuation of existing distributional hypernymy detection techniques, we foresee contextualized word representations also being used to advance pattern-based approaches. Here, they would allow to learn not only lexicosyntactic but, for the first time, also semantic patterns that indicate hypernymy.

[3]Specifically, we measured that calculating ELMo representation with the pre-trained model took about 5 ms per word token (roughly 58 days per billion tokens) on an AMD Ryzen Threadripper 1950X CPU and about 0.2 ms per token (2.5 days per billion tokens) on a NVIDIA Titan V GPU. Storing the calculated representations took about 8 kB per token (8 TB per billion tokens).

References

[Athiwaratkun and Wilson 2017] Athiwaratkun, B., and Wilson, A. G. 2017. Multimodal word distributions. In *ACL (1)*, 1645–1656. ACL.

[Athiwaratkun and Wilson 2018] Athiwaratkun, B., and Wilson, A. G. 2018. Hierachical denstity order embeddings. In *ICLR*.

[Camacho-Collados and Pilehvar 2018] Camacho-Collados, J., and Pilehvar, M. T. 2018. From word to sense embeddings: A survey on vector representations of meaning. *CoRR* abs/1805.04032.

[Chang et al. 2018] Chang, H.; Wang, Z.; Vilnis, L.; and McCallum, A. 2018. Distributional inclusion vector embedding for unsupervised hypernymy detection. In *NAACL-HLT*, 485–495. ACL.

[Chelba et al. 2014] Chelba, C.; Mikolov, T.; Schuster, M.; Ge, Q.; Brants, T.; Koehn, P.; and Robinson, T. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *INTERSPEECH*, 2635–2639. ISCA.

[Chen, Liu, and Sun 2014] Chen, X.; Liu, Z.; and Sun, M. 2014. A unified model for word sense representation and disambiguation. In *EMNLP*, 1025–1035. ACL.

[de Knijff, Frasincar, and Hogenboom 2013] de Knijff, J.; Frasincar, F.; and Hogenboom, F. 2013. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data Knowl. Eng.* 83:54–69.

[Devlin et al. 2018] Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

[Geffet and Dagan 2005] Geffet, M., and Dagan, I. 2005. The distributional inclusion hypotheses and lexical entailment. In *ACL*, 107–114. ACL.

[Gupta et al. 2017] Gupta, A.; Lebret, R.; Harkous, H.; and Aberer, K. 2017. Taxonomy induction using hypernym subsequences. In *CIKM*, 1329–1338. ACM.

[Hotho, Staab, and Stumme 2003] Hotho, A.; Staab, S.; and Stumme, G. 2003. Ontologies improve text document clustering. In *ICDM*, 541–544. IEEE Computer Society.

[Hua et al. 2017] Hua, W.; Wang, Z.; Wang, H.; Zheng, K.; and Zhou, X. 2017. Understand short texts by harvesting and analyzing semantic knowledge. *IEEE Trans. Knowl. Data Eng.* 29(3):499–512.

[Huang et al. 2012] Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *ACL (1)*, 873–882. ACL.

[Klapaftis and Manandhar 2010] Klapaftis, I. P., and Manandhar, S. 2010. Taxonomy learning using word sense induction. In *HLT-NAACL*, 82–90. ACL.

[Levy et al. 2015] Levy, O.; Remus, S.; Biemann, C.; and Dagan, I. 2015. Do supervised distributional methods really learn lexical inference relations? In *HLT-NAACL*, 970–976. ACL.

[Liang et al. 2017a] Liang, J.; Xiao, Y.; Zhang, Y.; Hwang, S.; and Wang, H. 2017a. Graph-based wrong isa relation detection in a large-scale lexical taxonomy. In *AAAI*, 1178–1184. AAAI Press.

[Liang et al. 2017b] Liang, J.; Zhang, Y.; Xiao, Y.; Wang, H.; Wang, W.; and Zhu, P. 2017b. On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies. In *AAAI*, 1185–1191. AAAI Press.

[Maedche and Staab 2001] Maedche, A., and Staab, S. 2001. Ontology learning for the semantic web. *IEEE Intelligent Systems* 16(2):72–79.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

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

[Neelakantan et al. 2014] Neelakantan, A.; Shankar, J.; Passos, A.; and McCallum, A. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*, 1059–1069. ACL.

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global vectors for word representation. In *EMNLP*, 1532–1543. ACL.

[Periñán-Pascual 2018] Periñán-Pascual, C. 2018. DEXTER: A workbench for automatic term extraction with specialized corpora. *Nat. Lang. Eng.* 24(2):163–198.

[Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL-HLT*, 2227–2237. ACL.

[Radford et al. 2018] Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

[Roller, Kiela, and Nickel 2018] Roller, S.; Kiela, D.; and Nickel, M. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *ACL (2)*, 358–363. ACL.

[Santus, Shwartz, and Schlechtweg 2017] Santus, E.; Shwartz, V.; and Schlechtweg, D. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *EACL (1)*, 65–75. ACL.

[Sclano and Velardi 2007] Sclano, F., and Velardi, P. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. In *IESA*, 287–290. Springer.

[Shwartz, Goldberg, and Dagan 2016] Shwartz, V.; Goldberg, Y.; and Dagan, I. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *ACL (1)*, 2389–2398. ACL.

[Stanovsky and Hopkins 2018] Stanovsky, G., and Hopkins, M. 2018. Spot the odd man out: Exploring the associative power of lexical resources. In *EMNLP*, 1533–1542. ACL.

[Turney and Pantel 2010] Turney, P. D., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.* 37:141–188.

[van der Maaten and Hinton 2008] van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR* 9(Nov):2579–2605.

[Velardi, Faralli, and Navigli 2013] Velardi, P.; Faralli, S.; and Navigli, R. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Comput. Linguist.* 39(3):665–707.

[Vilnis and McCallum 2015] Vilnis, L., and McCallum, A. 2015. Word representations via gaussian embedding. In *ICLR*.

[Vulic et al. 2017] Vulic, I.; Gerz, D.; Kiela, D.; Hill, F.; and Korhonen, A. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics* 43(4):781–835.

[Wang, He, and Zhou 2017] Wang, C.; He, X.; and Zhou, A. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *EMNLP*, 1190–1203. ACL.

[Yang et al. 2017] Yang, S.; Zou, L.; Wang, Z.; Yan, J.; and Wen, J. 2017. Efficiently answering technical questions - A knowledge graph approach. In *AAAI*, 3111–3118. AAAI Press.

[Yatskar et al. 2017] Yatskar, M.; Ordonez, V.; Zettlemoyer, L.; and Farhadi, A. 2017. Commonly uncommon: Semantic sparsity in situation recognition. In *CVPR*, 6335–6344. IEEE Computer Society.