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

Lukas Schmelzeisen\textsuperscript{1} and Steffen Staab\textsuperscript{1,2}
\textsuperscript{1}Institute for Web Science and Technologies (WeST), University of Koblenz-Landau, Germany
\textsuperscript{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), question 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 confute the different senses of a word and typically only learn the most predominant one.

This limitation is reinforced by context-free word representation\footnote{The NLP literature uses both distributed representations and distributional representations to refer to this concept. The intent is usually to distinguish between latent, dense vectors (low-dimensional) and sparse vectors (high-dimensional), respectively.} (Turney and Pantel 2010; Mikolov et al. 2013). A promising method to address this limitation. We propose to represent concepts by probability density estimates that approximate all points in the vector space. 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 sub-tasks 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-
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 (Sclavo 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; Santos, 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 taxonomic 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 (Klaptauf and Manandhar 2010), which is often error-prone. As a result, automatically created taxonomies confl ate 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 Pan-
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 \texttt{elmo.2x4096.S12.2048cnn.2xhighway} 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 it 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.

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 (Pereira-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 contextual-
ized 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 2015] 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. 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.

---

3Specifically, 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).
