Modelling Semantic Categories using Conceptual Neighborhood

Zied Bouraoui
CRIL - U. Artois - CNRS
zied.bouraoui@cril.fr

Jose Camacho-Collados
Cardiff University, UK
camachocolladosj@cardiff.ac.uk

Luis Espinosa-Anke
Cardiff University, UK
espinosa-ankel@cardiff.ac.uk

Steven Schockaert
Cardiff University, UK
schockaerts1@cardiff.ac.uk

Abstract

While many methods for learning vector space embeddings have been proposed in the field of Natural Language Processing, these methods typically do not distinguish between categories and individuals. Intuitively, if individuals are represented as vectors, we can think of categories as (soft) regions in the embedding space. Unfortunately, meaningful regions can be difficult to estimate, especially since we often have few examples of individuals that belong to a given category. To address this issue, we rely on the fact that different categories are often highly interdependent. In particular, categories often have conceptual neighbors, which are disjoint from but closely related to the given category (e.g. fruit and vegetable). Our hypothesis is that more accurate category representations can be learned by relying on the assumption that the regions representing such conceptual neighbors should be adjacent in the embedding space. We propose a simple method for identifying conceptual neighbors and then show that incorporating these conceptual neighbors indeed leads to more accurate region based representations.

1 Introduction

Vector space embeddings are commonly used to represent entities in fields such as machine learning (ML) (Bordes et al. 2013), natural language processing (NLP) (Camacho-Collados, Pilehvar, and Navigli 2016), information retrieval (IR) (Deerwester et al. 1990) and cognitive science (Gärdenfors 2000). An important point, however, is that such representations usually represent both individuals and categories as vectors (Ma, Cambria, and Gao 2016; Zheng et al. 2016; Boleda, Gupta, and Padó 2017). Note that in this paper, we use the term *category* to denote natural groupings of individuals, as it is used in cognitive science, with *individuals* referring to the objects from the considered domain of discourse. For example, the individuals *carrot* and *cucumber* belong to the *vegetable* category[1]. We use the term *entities* as an umbrella term covering both individuals and categories.

[1]Note that the same entity could be treated as an individual or a category depending on the context; e.g. *carrot* is a category of physical objects, but an instance of the *vegetable* category.

Given that a category corresponds to a set of individuals (i.e. its instances), modelling them as (possibly imprecise) regions in the embedding space seems more natural than using vectors. In fact, it has been shown that the vector representations of individuals that belong to the same category are indeed often clustered together in learned vector space embeddings (Gupta et al. 2015; Jameel, Bouraoui, and Schockaert 2017). The view of categories being regions is also common in cognitive science (Gärdenfors 2000). However, learning region representations of categories is a challenging problem, because we typically only have a handful of examples of individuals that belong to a given category. One common assumption is that natural categories can be modelled using *convex* regions (Gärdenfors 2000), which simplifies the estimation problem. For instance, based on this assumption, Bouraoui, Jameel, and Schockaert (2017) modelled categories using Gaussian distributions and showed that these distributions can be used for knowledge base completion. Unfortunately, this strategy still requires a relatively high number of training examples to be successful.

However, when learning categories, humans do not only rely on examples. For instance, there is evidence that when learning the meaning of nouns, children rely on the default assumption that these nouns denote mutually exclusive categories (Markman 1990). In this paper, we will in particular take advantage of the fact that many natural categories are organized into so-called *contrast sets* (Goldstone 1996). These are sets of closely related categories which exhaustively cover some sub-domain, and which are assumed to be mutually exclusive; e.g. the set of all common color names, the set $\{\text{fruit}, \text{vegetable}\}$ or the set $\{\text{NLP}, \text{IR}, \text{ML}\}$. Categories from the same contrast set often compete for coverage. For instance, we can think of the NLP domain as consisting of research topics that involve processing textual information *which are not covered by the IR and ML domains*. Categories which compete for coverage in this way are known as *conceptual neighbors* (Freksa 1991); e.g. NLP and IR, red and orange, fruit and vegetable. Note that the exact boundary between two conceptual neighbors may be vague (e.g. tomato can be classified as fruit or as vegetable).

In this paper, we propose a method for learning region representations of categories which takes advantage of con-

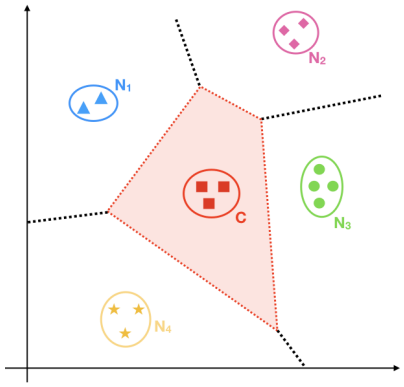

Figure 1: Using conceptual neighborhood for estimating category boundaries.

ceptual neighborhood, especially in scenarios where the number of available training examples is small. The main idea is illustrated in Figure 1, which depicts a situation where we are given some examples of a target category C as well as some related categories N_1, N_2, N_3, N_4. If we have to estimate a region from the examples of C alone, the small elliptical region shown in red would be a reasonable choice. More generally, a standard approach would be to estimate a Gaussian distribution from the given examples. However, vector space embeddings typically have hundreds of dimensions, while the number of known examples of the target category is often far lower (e.g. 2 or 3). In such settings we will almost inevitably underestimate the coverage of the category[2]. However, in the example from Figure 1, if we take into account the knowledge that N_1, N_2, N_3, N_4 are conceptual neighbors of C, the much larger, shaded region becomes a more natural choice for representing C. Indeed, the fact that e.g. C and N_1 are conceptual neighbors suggests that any point in between the examples of these categories needs to be contained either in the region representing C or the region representing N_1. In the spirit of prototype approaches to categorization (Rosch 1973), without any further information it makes sense to assume that their boundary is more or less half-way in between the known examples.

The contribution of this paper is two-fold. First, we propose a method for identifying conceptual neighbors from text corpora. We essentially treat this problem as a standard text classification problem, by relying on categories with large numbers of training examples to generate a suitable distant supervision signal. Second, we show that the predicted conceptual neighbors can effectively be used to learn better category representations.

2 Related Work

In distributional semantics, categories are frequently modelled as vectors. For example, Gupta, Boleda, and Padó (2018) study the problem of deciding for a word pair (i, c) whether i denotes an instance of the category c, which they refer to as *instantiation*. They treat this problem as a binary classification problem, where e.g. the pair (AAAI, conference) would be a positive example, while (conference, AAAI) and (New York, conference) would be negative examples. Different from our setting, their aim is thus essentially to model the instantiation relation itself, similar in spirit to how hypernymy has been modelled in NLP (Weeds et al. 2014; Roller, Erk, and Boleda 2014). To predict instantiation, they use a simple neural network model which takes as input the word vectors of the input pair (i, c). They also experiment with an approach that instead models a given category as the average of the word vectors of its known instances and found that this led to better results.

A few authors have already considered the problem of learning region representations of categories. Most closely related, Bouraoui and Schockaert (2018) model ontology concepts using Gaussian distributions. In Jameel and Schockaert (2016), a model is presented which embeds Wikipedia entities such that entities which have the same WikiData type are characterized by some region within a low-dimensional subspace of the embedding. Within the context of knowledge graph embedding, several approaches have been proposed that essentially model semantic types as regions (Neelakantan and Chang 2015; Guo et al. 2015). A few approaches have also been proposed for modelling word meaning using regions (Erk 2009; Jameel and Schockaert 2017) or Gaussian distributions (Vilnis and McCallum 2015). Along similar lines, several authors have proposed approaches inspired by probabilistic topic modelling, which model latent topics using Gaussians (Das, Zaheer, and Dyer 2015) or related distributions (Batmanghelich et al. 2016).

On the other hand, the notion of conceptual neighborhood has been covered in most detail in the field of spatial cognition, starting with the influential work of Freksa (1991). In computational linguistics, moreover, this representation framework aligns with lexical semantics traditions where word meaning is constructed in terms of *semantic decomposition*, i.e. lexical items being minimally decomposed into structured forms (or templates) rather than sets of features (Pustejovsky 1991), effectively mimicking a sort of conceptual neighbourhood. In Pustejovsky's *generative lexicon*, a set of "semantic devices" is proposed such that they behave in semantics similarly as grammars do in syntax. Specifically, this framework considers the *qualia* structure of a lexical unit as a set of expressive semantic distinctions, the most relevant for our purposes being the so-called *formal role*, which is defined as "that which distinguishes the object within a larger domain", e.g. shape or color. This semantic interplay between cognitive science and computational linguistics gave way to the term *lexical coherence*, which has been used for contextualizing the meaning of words in terms of how they relate to their conceptual neighbors (Wellner et al. 2006), or by providing expressive lexical semantic resources in the form of ontologies (Pustejovsky et al. 2006).

3 Model Description

Our aim is to introduce a model for learning region-based category representations which can take advantage of

[2]Note that k examples span a subspace of at most $k-1$ dimensions, and can thus not provide us with any information about the variance along directions which are orthogonal to that subspace.

knowledge about the conceptual neighborhood of that category. Throughout the paper, we focus in particular on modelling categories from the BabelNet taxonomy (Navigli and Ponzetto 2012), although the proposed method can be applied to any resource which (i) organizes categories in a taxonomy and (ii) provides examples of individuals that belong to these categories. Selecting BabelNet as our use case is a natural choice, however, given its large scale and the fact that it integrates many lexical and ontological resources.

As the possible conceptual neighbors of a given BabelNet category C, we consider all its siblings in the taxonomy, i.e. all categories $C_1, ..., C_k$ which share a direct parent with C. To select which of these siblings are most likely to be conceptual neighbors, we look at mentions of these categories in a text corpus. As an illustrative example, consider the pair (hamlet,village) and the following sentence[3]:

In British geography, a hamlet is considered smaller than a village and ...

From this sentence, we can derive that *hamlet* and *village* are disjoint but closely related categories, thus suggesting that they are conceptual neighbors. However, training a classifier that can identify conceptual neighbors from such sentences is complicated by the fact that conceptual neighborhood is not covered in any existing lexical resource, to the best of our knowledge, which means that large sets of training examples are not readily available. To address this lack of training data, we rely on a distant supervision strategy. The central insight is that for categories with a large number of known instances, we can use the embeddings of these instances to check whether two categories are conceptual neighbors. In particular, our approach involves the following three steps:

1. Identify pairs of categories that are likely to be conceptual neighbors, based on the vector representations of their known instances.
2. Use the pairs from Step 1 to train a classifier that can recognize sentences which indicate that two categories are conceptual neighbors.
3. Use the classifier from Step 2 to predict which pairs of BabelNet categories are conceptual neighbors and use these predictions to learn category representations.

Note that in Step 1 we can only consider BabelNet categories with a large number of instances, while the end result in Step 3 is that we can predict conceptual neighborhood for categories with only few known instances. We now discuss the three aforementioned steps one by one.

3.1 Step 1: Predicting Conceptual Neighborhood from Embeddings

Our aim here is to generate distant supervision labels for pairs of categories, indicating whether they are likely to be conceptual neighbors. These labels will then be used in Section 3.2 to train a classifier for predicting conceptual neighborhood from text.

Let A and B be siblings in the BabelNet taxonomy. If enough examples of individuals belonging to these categories are provided in BabelNet, we can use these instances to estimate high-quality representations of A and B, and thus estimate whether they are likely to be conceptual neighbors. In particular, we split the known instances of A into a training set I^A_{train} and test set I^A_{test}, and similar for B. We then train two types of classifiers. The first classifier estimates a Gaussian distribution for each category, using the training instances in I^A_{train} and I^B_{train} respectively. This should provide us with a reasonable representation of A and B regardless of whether they are conceptual neighbors. In the second approach, we first learn a Gaussian distribution from the joint set of training examples $I^A_{train} \cup I^B_{train}$ and then train a logistic regression classifier to separate instances from A and B. In particular, note that in this way, we directly impose the requirement that the regions modelling A and B are adjacent in the embedding space (intuitively corresponding to two halves of a Gaussian distribution). We can thus expect that the second approach should lead to better predictions than the first approach if A and B are conceptual neighbors and to worse predictions if they are not. In particular, we propose to use the relative performance of the two classifiers as the required distant supervision signal for predicting conceptual neighborhood.

We now describe the two classification models in more detail, after which we explain how these models are used to generate the distant supervision labels.

1. **Gaussian Classifier** The first classifier follows the basic approach from Bouraoui and Schockaert (2018), where Gaussian distributions were similarly used to model WikiData categories. In particular, we estimate the probability that an individual e with vector representation $\mathbf{e}$ is an instance of the category A as follows:

$$P(A|\mathbf{e}) = \lambda_A \cdot \frac{f(\mathbf{e}|A)}{f(\mathbf{e})}$$

where λ_A is the prior probability of belonging to category A, the likelihood $f(\mathbf{e}|A)$ is modelled as a Gaussian distribution and $f(\mathbf{e})$ will also be modelled as a Gaussian distribution. Intuitively, we think of the Gaussian $f(.|A)$ as defining a soft region, modelling the category A. Given the high-dimensional nature of typical vector space embeddings, we use a mean field approximation:

$$f(\mathbf{e}|A) = \prod_{i=1}^{d} f_i(e_i|A)$$

Where d is the number of dimensions in the vector space embedding, e_i is the i^{th} coordinate of $\mathbf{e}$, and $f_i(.|A)$ is a univariate Gaussian. To estimate the parameters μ_i and σ_i^2 of this Gaussian, we use a Bayesian approach with a flat prior:

$$f_i(e_i|A) = \int G(e_i; \mu_i, \sigma_i^2) NI\chi^2(\mu, \sigma^2) d\mu d\sigma$$

where $G(e_i; \mu_i, \sigma_i^2)$ represents the Gaussian distribution with mean μ_i and variance σ_i^2 and $NI\chi^2$ is the normal inverse-χ^2 distribution. In other words, instead of using a single estimate of the mean μ and variance σ_2 we average over all plausible choices of these parameters. The use

[3] https://en.wikipedia.org/wiki/Hamlet_(place)

of the normal inverse-χ^2 distribution for the prior on μ_i and σ_i^2 is a common choice, which has the advantage that the above integral simplifies to a Student-t distribution. In particular, we have:

$$f_i(e_i|A) = t_{n-1}\left(\overline{x_i}, \frac{(n+1)\sum_{j=1}^{n}(a_i^j - \overline{x_i})^2}{n(n-1)}\right)$$

where we assume $I^A_{train} = \{a_1, ..., a_n\}$, a_i^j denotes the i^{th} coordinate of the vector embedding of a_j, $\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} a_i^j$ and t_{n-1} is the Student t-distribution with $n-1$ degrees of freedom. The probability $f(\mathbf{e})$ is estimated in a similar way, but using all BabelNet instances. The prior λ_A is tuned based on a validation set. Finally, we classify e as a positive example if $P(A|\mathbf{e}) > 0.5$.

2. **GLR Classifier.** We first train a Gaussian classifier as in Section 1, but now using the training instances of both A and B. Let us denote the probability predicted by this classifier as $P(A \cup B|\mathbf{e})$. The intuition is that entities for which this probability is high should either be instances of A or of B, provided that A and B are conceptual neighbors. If, on the other hand, A and B are not conceptual neighbors, relying on this assumption is likely to lead to errors (i.e. there may be individuals whose representation is in between A and B which are not instances of either), which is what we need for generating the distant supervision labels. If $P(A \cup B|\mathbf{e}) > 0.5$, we assume that e either belongs to A or to B. To distinguish between these two cases, we train a logistic regression classifier, using the instances from I^A_{train} as positive examples and those from I^B_{train} as negative examples. Putting everything together, we thus classify e as a positive example for A if $P(A \cup B|\mathbf{e}) > 0.5$ and e is classified as a positive example by the logistic regression classifier. Similarly, we classfiy e as a positive example for B if $P(A \cup B|\mathbf{e}) > 0.5$ and e is classified as a negative example by the logistic regression classifier. We will refer to this classification model as GLR (Gaussian Logistic Regression).

Generating Distant Supervision Labels To generate the distant supervision labels, we consider a ternary classification problem for each pair of siblings A and B. In particular, the task is to decide for a given individual e whether it is an instance of A, an instance of B, or an instance of neither (where only disjoint pairs A and B are considered). For the Gaussian classifier, we predict A iff $P(A|\mathbf{e}) > 0.5$ and $P(A|\mathbf{e}) > P(B|\mathbf{e})$. For the GLR classifier, we predict A if $P(A \cup B|\mathbf{e}) > 0.5$ and the associated logistic regression classifier predicts A. The condition for predicting B is analogous. The test examples for this ternary classification problem consist of the elements from I^A_{test} and I^B_{test}, as well as some negative examples (i.e. individuals that are neither instances of A nor B). To select these negative examples, we first sample instances from categories that have the same parent as A and B, choosing as many such negative examples as we have positive examples. Second, we also sample the same number of negative examples from randomly selected categories in the taxonomy.

Let F^1_{AB} be the F1 score achieved by the Gaussian classifier and F^2_{AB} the F1 score of the GLR classifier. Our hypothesis is that $F^1_{AB} \ll F^2_{AB}$ suggests that A and B are conceptual neighbors, while $F^1_{AB} \gg F^2_{AB}$ suggests that they are not. This intuition is captured in the following score:

$$s_{AB} = \frac{F^2_{AB}}{F^1_{AB} + F^2_{AB}}$$

where we consider A and B to be conceptual neighbors if $s_{AB} \gg 0.5$.

3.2 Step 2: Predicting Conceptual Neighborhood from Text

We now consider the following problem: given two BabelNet categories A and B, predict whether they are likely to be conceptual neighbors based on the sentences from a text corpus in which they are both mentioned. To train such a classifier, we use the distant supervision labels from Section 3.1 as training data. Once this classifier has been trained, we can then use it to predict conceptual neighborhood for categories for which only few instances are known.

To find sentences in which both A and B are mentioned, we rely on a disambiguated text corpus in which mentions of BabelNet categories are explicitly tagged. Such a disambiguated corpus can be automatically constructed, using methods such as the one proposed by Mancini et al. (2017), for instance. For each pair of candidate categories, we thus retrieve all sentences where they co-occur. Next, we represent each extracted sentence as a vector. To this end, we considered two possible strategies:

1. **Word embedding averaging:** We compute a sentence embedding by simply averaging the word embeddings of each word within the sentence. Despite its simplicity, this approach has been shown to provide competitive results (Arora, Liang, and Ma 2017), in line with more expensive and sophisticated methods e.g. based on LSTMs.
2. **Contextualized word embeddings:** The recently proposed contextualized embeddings (Peters et al. 2018a; Devlin et al. 2019) have already proven successful in a wide range of NLP tasks. Instead of providing a single vector representation for all words irrespective of the context, contextualized embeddings predict a representation for each word occurrence which depends on its context. These representations are usually based on pre-trained language models. In our setting, we extract the contextualized embeddings for the two candidate categories within the sentence. To obtain this contextualized embedding, we used the last layer of the pre-trained language model, which has been shown to be most suitable for capturing semantic information (Peters et al. 2018b; Tenney, Das, and Pavlick 2019). We then use the concatenation of these two contextualized embeddings as the representation of the sentence.

For both strategies, we average their corresponding sentence-level representations across all sentences in which the same two candidate categories are mentioned. Finally, we train an SVM classifier on the resulting vectors to predict for the pair of siblings (A, B) whether $s_{AB} > 0.5$ holds.

3.3 Step 3: Category Induction

Let C be a category and assume that $N_1, ..., N_k$ are conceptual neighbors of this category. Then we can model C by generalizing the idea underpinning the GLR classifier. In particular, we first learn a Gaussian distribution from all the instances of C and $N_1, ..., N_k$. This Gaussian model allows us to estimate the probability $P(C \cup N_1 \cup ... \cup N_k \mid \mathbf{e})$ that e belongs to one of $C, N_1, ..., N_k$. If this probability is sufficiently high (i.e. higher than 0.5), we use a multinomial logistic regression classifier to decide which of these categories e is most likely to belong to. Geometrically, we can think of the Gaussian model as capturing the relevant local domain, while the multinomial logistic regression model carves up this local domain, similar as in Figure 1.

In practice, we do not know with certainty which categories are conceptual neighbors of C. Instead, we select the k categories (for some fixed constant k), among all the siblings of C, which are most likely to be conceptual neighbors, according to the text classifier from Section 3.2.

4 Experiments

The central problem we consider is category induction: given some instances of a category, predict which other individuals are likely to be instances of that category. When enough instances are given, standard approaches such as the Gaussian classifier from Section 1, or even a simple SVM classifier, can perform well on this task. For many categories, however, we only have access to a few instances, either because the considered ontology is highly incomplete or because the considered category only has few actual instances. The main research question which we want to analyze is whether (predicted) conceptual neighborhood can help to obtain better category induction models in such cases. In Section 4.1, we first provide more details about the experimental setting that we followed. Section 4.2 then discusses our main quantitative results. Finally, in Section 4.3 we present a qualitative analysis.

4.1 Experimental setting

Taxonomy As explained in Section 3, we used BabelNet (Navigli and Ponzetto 2012) as our reference taxonomy. BabelNet is a large-scale full-fledged taxonomy consisting of heterogeneous sources such as WordNet (Fellbaum 1998), Wikidata (Vrandečić and Krötzsch 2014) and WiBi (Flati et al. 2016), making it suitable to test our hypothesis in a general setting.

Vector space embeddings. Both the distant labelling method from Section 3.1 and the category induction model itself need access to vector representations of the considered instances. To this end, we used the NASARI vectors[4], which have been learned from Wikipedia and are already linked to BabelNet (Camacho-Collados, Pilehvar, and Navigli 2016).

BabelNet category selection. To test our proposed category induction model, we consider all BabelNet categories with fewer than 50 known instances. This is motivated by the view that conceptual neighborhood is mostly useful in cases where the number of known instances is small. For each of these categories, we split the set of known instances into 90% for training and 10% for testing. To tune the prior probability λ_A for these categories, we hold out 10% from the training set as a validation set.

The conceptual neighbors among the considered test categories are predicted using the classifier from Section 3.2. To obtain the distant supervision labels needed to train that classifier, we consider all BabelNet categories with at least 50 instances. This ensures that the distant supervision labels are sufficiently accurate and that there is no overlap with the categories which are used for evaluating the model.

Text classifier training. As the text corpus to extract sentences for category pairs we used the English Wikipedia. In particular, we used the dump of November 2014, for which a disambiguated version is available online[5]. This disambiguated version was constructed using the shallow disambiguation algorithm of Mancini et al. (2017). As explained in Section 3.2, for each pair of categories we extracted all the sentences where they co-occur, including a maximum window size of 10 tokens between their occurrences, and 10 tokens to the left and right of the first and second category within the sentence, respectively. For the averaging-based sentence representations we used the 300-dimensional pre-trained GloVe word embeddings (Pennington, Socher, and Manning 2014).[6] To obtain the contextualized representations we used the pre-trained 768-dimensional BERT-base model (Devlin et al. 2019).[7].

The text classifier is trained on 3,552 categories which co-occur at least once in the same sentence in the Wikipedia corpus, using the corresponding scores s_{AB} as the supervision signal (see Section 3.2). To inspect how well conceptual neighborhood can be predicted from text, we performed a 10-fold cross validation over the training data, removing for this experiment the *unclear* cases (i.e., those category pairs with s_{AB} scores between 0.4 and 0.6). We also considered a simple baselineWE based on the number of co-occurring sentences for each pairs, which we might expect to be a reasonably strong indicator of conceptual neighborhood, i.e. the more often two categories are mentiond in the same sentence, the more likely that they are conceptual neighbors. The results for this cross-validation experiment are summarized in Table 1. Surprisingly, perhaps, the word vector averaging method seems more robust overall, while being considerably faster than the method using BERT. The results also confirm the intuition that the number of co-occurring sentences is positively correlated with conceptual neighborhood, although the results for this baseline are clearly weaker than those for the proposed classifiers.

Baselines. To put the performance of our model in perspective, we consider three baseline methods for category induction. First, we consider the performance of the Gaussian classifier from Section 1, as a representative example

[4] Downloaded from http://lcl.uniroma1.it/nasari/.

[5] Available at http://lcl.uniroma1.it/sw2v

[6] Pre-trained embeddings downloaded from https://nlp.stanford.edu/projects/glove/

[7] We used the implementation available at https://github.com/huggingface/pytorch-pretrained-BERT

	Acc	F1	Pr	Rec
Avg.	**70.6**	**69.0**	**69.4**	**69.0**
BERT	66.9	65.8	65.9	66.2
#sents	61.6	46.6	43.3	54.3

Table 1: Cross-validation results on the training split of the text classifier (accuracy and macro-average F1, precision and recall).

	Pr	Rec	F1
Gauss	23.0	27.4	22.3
Multi	37.7	75.2	44.2
Similarity$_1$	28.7	69.2	33.8
Similarity$_2$	30.0	68.1	34.0
Similarity$_3$	31.6	67.2	34.3
Similarity$_4$	32.8	78.5	38.2
Similarity$_5$	37.2	80.6	42.8
SECOND-WEA$_1$	32.7	**90.1**	41.9
SECOND-WEA$_2$	42.2	82.6	49.3
SECOND-WEA$_3$	43.4	83.1	50.4
SECOND-WEA$_4$	**47.7**	84.2	**54.2**
SECOND-WEA$_5$	44.0	82.6	51.1
SECOND-BERT$_1$	38.5	87.1	47.0
SECOND-BERT$_2$	43.9	84.1	50.8
SECOND-BERT$_3$	44.9	84.4	52.2
SECOND-BERT$_4$	46.2	85.4	53.3
SECOND-BERT$_5$	43.8	84.7	51.3

Table 2: Results (%) of the category induction experiments

of how well we can model each category when only considering their given instances; this model will be referred to as *Gauss*. Second, we consider a variant of the proposed model in which we assume that all siblings of the category are conceptual neighbors; this model will be referred to as *Multi*. Third, we consider a variant of our model in which the neighbors are selected based on similarity. To this end, we represent each BabelNet as their vector from the NASARI space. From the set of siblings of the target category C, we then select the k categories whose vector representation is most similar to that of C, in terms of cosine similarity. This baseline will be referred to as *Similarity*$_k$, with k the number of selected neighbors.

We refer to our model as *SECOND-WEA*$_k$ or *SECOND-BERT*$_k$ (SEmantic categories with COnceptual NeighborhooD), depending on whether the word embedding averaging strategy is used or the method using BERT.

4.2 Quantitative Results

Our main results for the category induction task are summarized in Table 2. In this table, we show results for different choices of the number of selected conceptual neighbors k, ranging from 1 to 5. As can be seen from the table, our approach substantially outperforms all baselines, with *Multi* being the most competitive baseline. Interestingly, for the

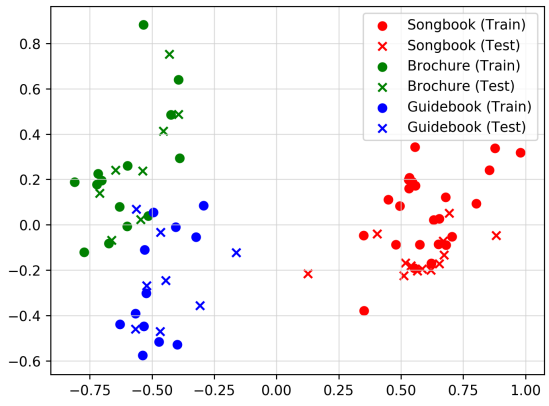

Figure 2: Instances of three BabelNet categories which intuitively can be seen as conceptual neighbors.

Similarity baseline, the higher the number of neighbors, the more the performance approaches that of *Multi*. The relatively strong performance of *Multi* shows that using the siblings of a category in the BabelNet taxonomy is in general useful. However, as our results show, better results can be obtained by focusing on the predicted conceptual neighbors only. It is interesting to see that even selecting a single conceptual neighbor is already sufficient to substantially outperform the Gaussian model, although the best results are obtained for $k = 4$. Comparing the *WEA* and *BERT* variants, it is notable that *BERT* is more successful at selecting the single best conceptual neighbor (reflected in an F1 score of 47.0 compared to 41.9). However, for $k \geq 2$, the results of the *WEA* and *BERT* are largely comparable.

4.3 Qualitative Analysis

To illustrate how conceptual neighborhood can improve classification results, Fig. 2 shows the two first principal components of the embeddings of the instances of three BabelNet categories: *Songbook*, *Brochure* and *Guidebook*. All three categories can be considered to be conceptual neighbors. Brochure and Guidebook are closely related categories, and we may expect there to exist borderline cases between them. This can be clearly seen in the figure, where some instances are located almost exactly on the boundary between the two categories. On the other hand, Songbook is slightly more separated in the space. Let us now consider the left-most data point from the Songbook test set, which is essentially an outlier, being more similar to instances of Guidebook than typical Songbook instances. When using a Gaussian model, this data point would not be recognised as a plausible instance. When incorporating the fact that Brochure and Guidebook are conceptual neighbors of Songbook, however, it is more likely to be classified correctly.

To illustrate the notion of conceptual neighborhood itself, Table 3 displays some selected category pairs from the training set (i.e. the category pairs that were used to train the text classifier), which intuitively correspond to conceptual

High confidence	Medium confidence
Actor – Comedian	Cruise ship – Ocean liner
Journal – Newspaper	Synagogue – Temple
Club – Company	Mountain range – Ridge
Novel – Short story	Child – Man
Tutor – Professor	Monastery – Palace
Museum – Public aquarium	Fairy tale – Short story
Lake – River	Guitarist – Harpsichordist

Table 3: Selected examples of siblings A–B for which the conceptual neighborhood score s_{AB} is higher than 0.9 (left column) and around 0.5 (right column).

Concept	Top neighbor	F1
Bachelor's degree	Undergraduate degree	34
Episodic video game	Multiplayer gamer	34
501(c) organization	Not-for-profit arts organization	29
Heavy bomber	Triplane	41
Ministry	United States government	33

Table 4: Top conceptual neighbors selected for categories associated with a low F1 score.

neighbors. The left column contains some selected examples of category pairs with a high s_{AB} score of at least 0.9. As these examples illustrate, we found that a high s_{AB} score was indeed often predictive of conceptual neighborhood. As the right column of this table illustrates, there are several category pairs with a lower s_{AB} score of around 0.5 which intuitively still seem to correspond to conceptual neighbors. When looking at category pairs with even lower scores, however, conceptual neighborhood becomes rare. Moreover, while there are several pairs with high scores which are not actually conceptual neighbors (e.g. the pair *Actor – Makup Artist*), they tend to be categories which are still closely related. This means that the impact of incorrectly treating them as conceptual neighbors on the performance of our method is likely to be limited. On the other hand, when looking at category pairs with a very low confidence score we find many unrelated pairs, which we can expect to be more harmful when considered as conceptual neighbors, as the combined Gaussian will then cover a much larger part of the space. Some examples of such pairs include *Primary school – Financial institution*, *Movie theatre – Housing estate*, *Corporate title – Pharaoh* and *Fraternity – Headquarters*.

Finally, in Tables 4 and 5, we show examples of the top conceptual neighbors that were selected for some categories from the test set. Table 4 shows examples of BabelNet categories for which the F1 score of our SECOND-WEA$_1$ classifier was rather low. As can be seen, the conceptual neighbors that were chosen in these cases are not suitable. For instance, *Bachelor's degree* is a near-synonym of *Undergraduate degree*, hence assuming them to be conceptual neighbors would clearly be detrimental. In contrast, when looking at the examples in Table 5, where categories are shown with a higher F1 score, we find examples of conceptual neighbors that are intuitively much more meaningful.

Concept	Top neighbor	F1
Amphitheater	Velodrome	67
Proxy server	Application server	61
Ketch	Cutter	74
Quintet	Brass band	67
Sand dune	Drumlin	71

Table 5: Top conceptual neighbors selected for categories associated with a high F1 score.

5 Conclusions

We have studied the role of conceptual neighborhood for modelling categories, focusing especially on categories with a relatively small number of instances, for which standard modelling approaches are challenging. To this end, we have first introduced a method for predicting conceptual neighborhood from text, by taking advantage of BabelNet to implement a distant supervision strategy. We then used the resulting classifier to identify the most likely conceptual neighbors of a given target category, and empirically showed that incorporating these conceptual neighbors leads to a better performance in a category induction task.

In terms of future work, it would be interesting to look at other types of lexical relations that can be predicted from text. One possible strategy would be to predict conceptual betweenness, where a category B is said to be between A and C if B has all the properties that A and C have in common (Schockaert and Li 2018) (e.g. we can think of *wine* as being conceptually between *beer* and *rum*). In particular, if B is predicted to be conceptually between A and C then we would also expect the region modelling B to be between the regions modelling A and C.

Acknowledgments. Jose Camacho-Collados, Luis Espinosa-Anke and Steven Schockaert were funded by ERC Starting Grant 637277. Zied Bouraoui was supported by CNRS PEPS INS2I MODERN.

References

Arora, S.; Liang, Y.; and Ma, T. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proc. ICLR*.

Batmanghelich, K.; Saeedi, A.; Narasimhan, K.; and Gershman, S. 2016. Nonparametric spherical topic modeling with word embeddings. In *Proc. ACL*, 537–542.

Boleda, G.; Gupta, A.; and Padó, S. 2017. Instances and concepts in distributional space. In *Proc. EACL*, 79–85.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings NIPS*. 2787–2795.

Bouraoui, Z., and Schockaert, S. 2018. Learning conceptual space representations of interrelated concepts. In *Proceedings IJCAI*, 1760–1766.

Bouraoui, Z.; Jameel, S.; and Schockaert, S. 2017. Inductive reasoning about ontologies using conceptual spaces. In *Proc. AAAI*, 4364–4370.

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

Das, R.; Zaheer, M.; and Dyer, C. 2015. Gaussian LDA for topic models with word embeddings. In *Proc. ACL*, 795–804.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*.

Erk, K. 2009. Representing words as regions in vector space. In *Proc. CoNLL*, 57–65.

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

Flati, T.; Vannella, D.; Pasini, T.; and Navigli, R. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artificial Intelligence* 241:66–102.

Freksa, C. 1991. Conceptual neighborhood and its role in temporal and spatial reasoning. In Singh, M., and Travé-Massuyès, L., eds., *Decision Support Systems and Qualitative Reasoning*. North-Holland, Amsterdam. 181–187.

Gärdenfors, P. 2000. *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Goldstone, R. L. 1996. Isolated and interrelated concepts. *Memory & Cognition* 24(5):608–628.

Guo, S.; Wang, Q.; Wang, B.; Wang, L.; and Guo, L. 2015. Semantically smooth knowledge graph embedding. In *Proc. ACL*, 84–94.

Gupta, A.; Boleda, G.; Baroni, M.; and Padó, S. 2015. Distributional vectors encode referential attributes. In *Proc. EMNLP*, 12–21.

Gupta, A.; Boleda, G.; and Padó, S. 2018. Instantiation. *CoRR* abs/1808.01662.

Jameel, S., and Schockaert, S. 2016. Entity embeddings with conceptual subspaces as a basis for plausible reasoning. In *Proc. ECAI*, 1353–1361.

Jameel, S., and Schockaert, S. 2017. Modeling context words as regions: An ordinal regression approach to word embedding. In *Proc. CoNLL*, 123–133.

Jameel, S.; Bouraoui, Z.; and Schockaert, S. 2017. MEmbER: Max-margin based embeddings for entity retrieval. In *Proc. SIGIR*, 783–792.

Ma, Y.; Cambria, E.; and Gao, S. 2016. Label embedding for zero-shot fine-grained named entity typing. In *Proc. COLING*, 171–180.

Mancini, M.; Camacho-Collados, J.; Iacobacci, I.; and Navigli, R. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proc. CoNLL*, 100–111.

Markman, E. M. 1990. Constraints children place on word meanings. *Cognitive Science* 14:57–77.

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

Neelakantan, A., and Chang, M. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. In *Proc. NAACL*, 515–525.

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

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018a. Deep contextualized word representations. In *Proc. NAACL-HLT*, 2227–2237.

Peters, M.; Neumann, M.; Zettlemoyer, L.; and Yih, W.-t. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proc. EMNLP*, 1499–1509.

Pustejovsky, J.; Havasi, C.; Littman, J.; Rumshisky, A.; and Verhagen, M. 2006. Towards a generative lexical resource: The brandeis semantic ontology. In *Proc. LREC*, 1702–1705.

Pustejovsky, J. 1991. The generative lexicon. *Computational Linguistics* 17(4):409–441.

Roller, S.; Erk, K.; and Boleda, G. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proc. COLING*, 1025–1036.

Rosch, E. H. 1973. Natural categories. *Cognitive Psychology* 4(3):328–350.

Schockaert, S., and Li, S. 2018. Reasoning about betweenness and RCC8 constraints in qualitative conceptual spaces. In *Proc. IJCAI*, 1963–1969.

Tenney, I.; Das, D.; and Pavlick, E. 2019. Bert rediscovers the classical nlp pipeline. In *Proc. ACL*.

Vilnis, L., and McCallum, A. 2015. Word representations via Gaussian embedding. In *Proc. ICLR*.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10):78–85.

Weeds, J.; Clarke, D.; Reffin, J.; Weir, D.; and Keller, B. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proc. COLING*, 2249–2259.

Wellner, B.; Pustejovsky, J.; Havasi, C.; Rumshisky, A.; and Saurí, R. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 117–125.

Zheng, R.; Tian, T.; Hu, Z.; Iyer, R.; Sycara, K.; et al. 2016. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In *Proc. COLING*, 2678–2688.