Capturing Discriminative Attributes in a Distributional Space: Task Proposal

Alicia Krebs and **Denis Paperno**
a.m.krebs@student.rug.nl | denis.paperno@unitn.it
Center for Mind and Brain Sciences (CIMeC), University of Trento, Rovereto, Italy

Abstract

If lexical similarity is not enough to reliably assess how word vectors would perform on various specific tasks, we need other ways of evaluating semantic representations. We propose a new task, which consists in extracting semantic *differences* using distributional models: given two words, what is the difference between their meanings? We present two proof of concept datasets for this task and outline how it may be performed.

1 Introduction

All similar pairs of words are similar in the same way: they share a substantial number of semantic properties (although properties themselves may belong to different groups, i.e. visual, functional, etc.). Cosine of two feature vectors in a distributional semantic space is a formalization of this idea, standardly used as a measure of semantic similarity for the evaluation of distributional models (Baroni et al., 2014a; Landauer and Dumais, 1997). While similarity tasks have become the standard in the evaluation of distributional models, the validity of those tasks has been put into question: inter-annotator agreement tends to be low, the small size of some of the most popular datasets is a concern, and subjective similarity scores have limitations when it comes to task-specific applications (Faruqui et al., 2016; Batchkarov et al., 2016). In contrast to similarity, the nature of semantic *difference* between two (related) words can vary greatly. Modeling difference can help capture individual aspects of meaning; similarity alone may be too simple a task to assess semantic representations in all their complexity, and therefore insufficient for driving the progress of computational models. Our project is related to previous work that attempts to predict the discriminative features of referents, using natural images to represent the input objects (Lazaridou et al., 2016). Attributes have also been used to simulate similarity judgements and concept categorization (Silberer and Lapata, 2014). On a more abstract level, our work is related to previous attempts at using offset vectors to capture lexical relations without explicit supervision (Mikolov et al., 2013), which have been shown to be able to generalise well to a range of relations (Vylomova et al., 2015).

We created two proof of concept datasets for the difference task: a small dataset of differences as feature oppositions and a bigger one with differences as presence vs. absence of a feature.

2 The Small Dataset

We used a random sample of seed words from the BLESS dataset (Baroni and Lenci, 2011) along with their semantic neighbors to create word pairs that were in some ways similar and denoted concrete objects. For each word pair, one or more pair(s) of discriminating attributes were assigned manually. For example, the word pair `[scooter, moped]` received two pairs of attributes: `[big, small]` and `[fast, slow]`. Some word pairs were also added manually to further exemplify specific differences, such as `[horse, foal]` for the age properties. The resulting dataset contains 91 items. To get a simple unsupervised baseline on the detection of difference direction, we calculated a similarity score for each item, using the cooccurrence counts of the best count-based configuration presented in Baroni et al. (2014b), which were extracted from the concatenation of the web-crawled ukWack corpus (Baroni et al., 2009), Wikipedia, and the BNC, for a total of 2.8 billion tokens. This similarity score calculates whether the attribute is closer to the first

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 51–54,
Berlin, Germany, August 12, 2016.

or second word. We found that 67% of items had positive scores. The most successful types of attributes were $color$ (34 out 51), age (9 out of 9) and $diet$ (4 out of 5).

$$Score = (CosSim(w_1, a_1) \cdot CosSim(w_1, a_2)) \\ -(CosSim(w_2, a_2) \cdot CosSim(w_2, a_1))$$

The dataset is too small for training supervised models; our attempts (logistic regression on pairwise cosines with cross-validation) showed negligibly low results.

3 Feature Norms Dataset

Only some differences can be expressed in the format assumed above, i.e. as the opposition of two attributes, such as *yellow* vs. *red* being the difference between bananas and apples. Other differences are better expressed as the presence or absence of a feature. For instance, the difference between a narwhal and a dolphin is the presence of a horn. For natural salient features of word concepts, we turned to property norms.

We used the set of feature norms collected by McRae et al. (2005), which includes features for 541 concepts (living and non-living entities), collected by asking 725 participants to produce features they found important for each concept. Production frequencies of these features indicate how salient they are. Feature norms of concepts are able to encode semantic knowledge because they tap into the representations that the participants have acquired through repeated exposure to those concepts. McRae et al. divided disjunctive features, so that if a participant produced the feature `is_green_or_red` the concept will be associated with both the feature `is_green` and the feature `is_red`. Concepts that have different meanings had been disambiguated before being shown to participants. For example, there are two entries for *bow*, `bow_(weapon)` and `bow_(ribbon)`. Because the word vector for *bow* encodes the properties of both senses, we did not differentiate between entries that have multiple senses. In our dataset, the concept `bow` has the features of both the weapon and the ribbon.

The McRae dataset uses the brain region taxonomy (Cree and McRae, 2003) to classify features into different types, such as *function*, *sound* or *taxonomic*. We decided to only work with visual features, which exist for all concrete concepts, while features such as *sound* or *taste* are only relevant for some concepts. This classification distinguishes between three types of visual features: *motion*, *color* and *form and surface*. We first selected words that had at least one visual feature of any type. We then created word pairs by selecting the 50 closest neighbours of every word in the dataset.

For each word pair, if there was a feature that the first word had but the second didn't, that word pair and feature item was added to our dataset. The set was built in such a way that the feature of each item always refers to an attribute of the first word. For example, in Table 2, *wings* is an attribute of *airplane*. The word pair `[airplane,helicopter]` will only be included in the order `[helicopter,airplane]` if *helicopter* has a feature that *airplane* doesn't have. The relations are thus asymmetric and have fixed directionality. For simplicity, multi-word features were processed so that only the final word is taken into account (e.g. `has_wings` becomes `wings`). In total, our dataset contains 528 concepts, 24 963 word pairs, and 128 515 items.

$word_1$	$word_2$	$feature$
airplane	helicopter	wings
bagpipe	accordion	pipes
canoe	sailboat	fibreglass
dolphin	seal	fins
gorilla	crocodile	bananas
oak	pine	leaves
octopus	lobster	tentacles
pajamas	necklace	silk
skirt	jacket	pleats
subway	train	dirty

Table 2: Examples of word pairs and their features

We computed a simple unsupervised baseline for direction of difference (e.g. is *subway* or *train* dirty?), choosing the first word iff $\cos(w_1 w_f) > \cos(w_2, w_f)$, and achieved 69% accuracy. Ultimately, this dataset could be used to build a model that can predict an exhaustive list of distinctive attributes for any pair of words. This could be done in a binary set-up where the dataset has been supplemented with negative examples: for a given triple, predict whether the attribute is a difference between $word_1$ and $word_2$.

type	w_1	w_2	a_1	a_2
color	tomato	spinach	red	green
color	banana	carrot	yellow	orange
color	tiger	panther	orange	black
age	cat	kitten	old	young
age	dog	pup	old	young
age	horse	foal	old	young
diet	deer	fox	herbivorous	carnivorous
diet	cow	lion	herbivorous	carnivorous
sex	pig	sow	male	female
sex	tiger	tigress	male	female

Table 1: Small Dataset: Examples of distinctive attribute pairs.

4 Conclusion

A system for basic language understanding should be able to detect when concepts are similar to each other, but also in what way concepts differ from each other. We've demonstrated how an evaluation set that captures differences between concepts can be built.

The baselines we computed show that the difference task we propose is a non-trivial semantic task. Even with the simplest evaluation setting where the difference was given and only the direction of the difference was to be established (e.g. where the task was to establish if tomato is red and spinach green or vice versa), the baseline methods achieved less than 70% accuracy. A more realistic evaluation setup would challenge models to produce a set of differences between two given concepts.

The dataset versions described in this paper are proof of concept realizations, and we keep working on improving the test sets. For instance, to counter the inherent noise of feature norms, we plan on using human annotation to confirm the validity of the test partition of the dataset.

In the future, solving the difference task could help in various applications, for example automatized lexicography (automatically generating features to include in dictionary definitions), conversational agents (choosing lexical items with contextually relevant differential features can help create more pragmatically appropriate, human-like dialogs), machine translation (where explicitly taking into account semantic differences between translation variants can improve the quality of the output), etc.

References

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014a. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014b. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *First Workshop on Evaluating Vector Space Representations for NLP (RepEval 2016)*.

George S Cree and Ken McRae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *First Workshop on Evaluating Vector Space Representations for NLP (RepEval 2016)*.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic

analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2016. The red one!: On learning to refer to things based on their discriminative properties. *arXiv preprint arXiv:1603.02618*.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL (1)*, pages 721–732.

Ekaterina Vylomova, Laura Rimmel, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *arXiv preprint arXiv:1509.01692*.