A Multitask Objective to Inject Lexical Contrast into Distributional Semantics

Nghia The Pham Angeliki Lazaridou Marco Baroni
Center for Mind/Brain Sciences
University of Trento
{thenghia.pham|angeliki.lazaridou|marco.baroni}@unitn.it

Abstract

Distributional semantic models have trouble distinguishing strongly contrasting words (such as antonyms) from highly compatible ones (such as synonyms), because both kinds tend to occur in similar contexts in corpora. We introduce the multitask Lexical Contrast Model (mLCM), an extension of the effective Skip-gram method that optimizes semantic vectors on the joint tasks of predicting corpus contexts and making the representations of WordNet synonyms closer than that of matching WordNet antonyms. mLCM outperforms Skip-gram both on general semantic tasks and on synonym/antonym discrimination, even when no direct lexical contrast information about the test words is provided during training. mLCM also shows promising results on the task of learning a compositional negation operator mapping adjectives to their antonyms.

1 Introduction

Distributional semantic models (DSMs) extract vectors representing word meaning by relying on the *distributional hypothesis*, that is, the idea that words that are related in meaning will tend to occur in similar contexts (Turney and Pantel, 2010). While extensive work has shown that contextual similarity is an excellent proxy to semantic similarity, a big problem for DSMs is that both words with very compatible meanings (e.g., near synonyms) and words with strongly contrasting meanings (e.g., antonyms) tend to occur in the same contexts. Indeed, Mohammad et al. (2013) have shown that synonyms and antonyms are indistinguishable in terms of their average degree of distributional similarity.

This is problematic for the application of DSMs to reasoning tasks such as entailment detection (*black* is very close to both *dark* and *white* in distributional semantic space, but it implies the former while contradicting the latter). Beyond word-level relations, the same difficulties make it challenging for compositional extensions of DSMs to capture the fundamental phenomenon of negation at the phrasal and sentential levels (the distributional vectors for *good* and *not good* are nearly identical) (Hermann et al., 2013; Preller and Sadrzadeh, 2011).

Mohammad and colleagues concluded that DSMs alone cannot detect semantic contrast, and proposed an approach that couples them with other resources. Pure-DSM solutions include isolating contexts that are expected to be more discriminative of contrast, tuning the similarity measure to make it more sensitive to contrast or training a supervised contrast classifier on DSM vectors (Adel and Schütze, 2014; Santus et al., 2014; Schulte im Walde and Köper, 2013; Turney, 2008). We propose instead to induce word vectors using a multitask cost function combining a traditional DSM context-prediction objective with a term forcing words to be closer to their WordNet synonyms than to their antonyms. In this way, we make the model aware that contrasting words such as *hot* and *cold*, while still semantically related, should not be nearest neighbours in the space.

In a similar spirit, Yih et al. (2012) devise a DSM in which the embeddings of the antonyms of a word are pushed to be the vectors that are farthest away from its representation. While their model is able to correctly pick the antonym of a target item from a list of candidates (since it is the most dissimilar element in the list), we conjecture that their radical strategy produces embeddings with poor performance on general semantic tasks.[1] Our method has instead a beneficial global

[1] Indeed, by simulating their strategy, we were able to inject lexical contrast into word embeddings, but performance on a general semantic relatedness task decreased dramati-

effect on semantic vectors, leading to state-of-the-art results in a challenging similarity task, and enabling better learning of a compositional negation function.

Our work is also closely related to Faruqui et al. (2015), who propose an algorithm to adapt pre-trained DSM representations using semantic resources such as WordNet. This post-processing approach, while extremely effective, has the disadvantage that changes only affect words that are present in the resource, without propagating to the whole lexicon. Other recent work has instead adopted multitask objectives similar to ours in order to directly plug in knowledge from structured resources at DSM induction time (Fried and Duh, 2015; Xu et al., 2014; Yu and Dredze, 2014). Our main novelties with respect to these proposals are the focus on capturing semantic contrast, and explicitly testing the hypothesis that the multitask objective is also beneficial to words that are not directly exposed to WordNet evidence during training.[2]

2 The multitask Lexical Contrast Model

Skip-gram model The multitask Lexical Contrast Model (mLCM) extends the Skip-gram model (Mikolov et al., 2013). Given an input text corpus, Skip-gram optimizes word vectors on the task of approximating, for each word, the probability of other words to occur in its context. More specifically, its objective function is:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)\right) \quad (1)$$

where $w_1, w_2, ..., w_T$ is the training corpus, consisting of a list of target words w_t, for which we want to learn the vector representations (and serving as contexts of each other), and c is the window size determining the span of context words to be considered. $p(w_{t+j}|w_t)$, the probability of a context word given the target word is computed using softmax:

$$p(w_{t+j}|w_t) = \frac{e^{{v'_{w_{t+j}}}^T v_{w_t}}}{\sum_{w'=1}^{W} e^{{v'_{w'}}^T v_{w_t}}} \quad (2)$$

where v_w and v'_w are respectively the target and context vector representations of word w, and W is the number of words in the vocabulary. To avoid the $O(|W|)$ time complexity of the normalization term in Equation (2), Mikolov et al. (2013) use either hierarchical softmax or negative sampling. Here, we adopt the negative sampling method.

Injecting lexical contrast information We account for lexical contrast by implementing a 2-task strategy, combining the Skip-gram context prediction objective with a new term:

$$\frac{1}{T}\sum_{t=1}^{T}\left(J_{skipgram}(w_t) + J_{lc}(w_t)\right) \quad (3)$$

The *lexical contrast* objective $J_{lc}(w_t)$ tries to enforce the constraint that contrasting pairs should have lower similarity than compatible ones within a max-margin framework. Our formulation is inspired by Lazaridou et al. (2015), who use a similar multitask strategy to induce multimodal embeddings. Given a target word w, with sets of antonyms $A(w)$ and synonyms $S(w)$, the max-margin objective for lexical contrast is:

$$-\sum_{s\in S(w), a\in A(w)} \max(0, \Delta - cos(v_w, v_s) + cos(v_w, v_a)) \quad (4)$$

where Δ is the margin and $cos(x, y)$ stands for cosine similarity between vectors x and y. Note that, by equation (3), the $J_{lc}(w_t)$ term is evaluated each time a word is encountered in the corpus. We extract antonym and synonym sets from WordNet (Miller, 1995). If a word w_t is not associated to synonym/antonym information in WordNet, we set $J_{lc}(w_t) = 0$.

3 Experimental setup

We compare the performance of mLCM against Skip-gram. Both models' parameters are estimated by backpropagation of error via stochastic gradient descent. Our text corpus is a Wikipedia[3] 2009 dump comprising approximately 800M tokens and 200K distinct word types.[4] Other hyper-parameters, selected without tuning, include: vector size (300), window size (5), negative samples (10), sub-sampling to disfavor frequent words (10^{-3}). For mLCM, we use 7500 antonym pairs

cally, with a 25% drop in terms of Spearman correlation.

[2] After submitting this work, we became aware of Ono et al. (2015), that implement very similar ideas. However, one major difference between their work and ours is that their strategy is in the same direction of (Yih et al., 2012), which might result in poor performance on general semantic tasks.

[3] https://en.wikipedia.org

[4] We only consider words that occur more than 50 times in the corpus

	MEN	SimLex
Skip-gram	0.73	0.39
mLCM	**0.74**	**0.52**

Table 1: Relatedness/similarity tasks

	AUC
Skip-gram	0.62
mLCM	**0.78**
mLCM-propagate	0.66

Table 2: Synonym vs antonym task

and 15000 synonym pairs; on average, 2.5 pairs per word and 9000 words are covered.

Both models are evaluated in four tasks: two lexical tasks testing the general quality of the learned embeddings and one focusing on antonymy, and a negation task which verifies the positive influence of lexical contrast in a compositional setting.

4 Lexical tasks

4.1 Relatedness and similarity

In classic semantic relatedness/similarity tasks, the models provide cosine scores between pairs of word vectors that are then compared to human ratings for the same pairs. Performance is evaluated by Spearman correlation between system and human scores. For general relatedness, we use the **MEN** dataset of Bruni et al. (2014), which consists of 3,000 word pairs comprising 656 nouns, 57 adjectives and 38 verbs. The **SimLex** dataset from Hill et al. (2014b), comprising 999 word pairs (666 noun, 222 verb and 111 adjective pairs) was explicitly built to test a tighter notion of strict "semantic" similarity.

Table 1 reports model performance. On MEN, mLCM outperforms Skip-gram by a small margin, which shows that the new information, at the very least, does not have any negative effect on general semantic relatedness. On the other hand, lexical contrast information has a strong positive effect on measuring strict semantic similarity, leading mLCM to achieve state-of-the-art SimLex performance (Hill et al., 2014a).

4.2 Distinguishing antonyms and synonyms

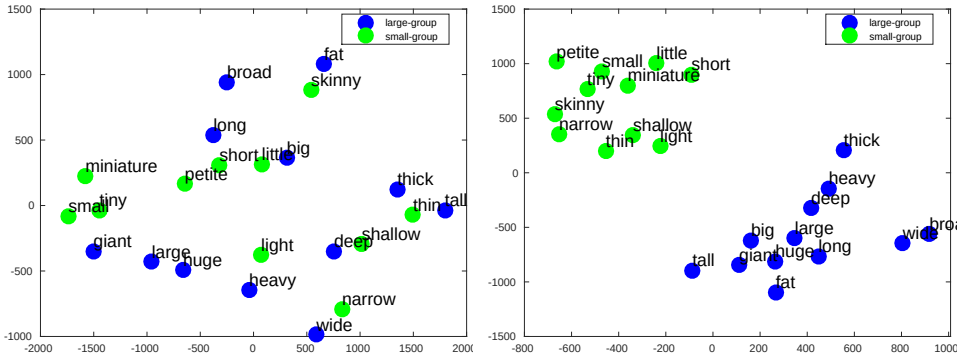
Having shown that capturing lexical contrast information results in higher-quality representations for general purposes, we focus next on the specific task of distinguishing contrasting words from highly compatible ones. We use the adjective part of dataset of Santus et al. (2014), that contains 262 antonym and 364 synonym pairs. We compute cosine similarity of all pairs and use the area under the ROC curve (AUC) to measure model performance. Moreover, we directly test mLCM's ability to propagate lexical contrast across the vocabulary by retraining it without using WordNet information for any of the words in the dataset, i.e. the words in the dataset are removed from the synonym or antonym sets of all the adjectives used in training (**mLCM-propagate** in the results table).

The results, in Table 2, show that mLCM can successfully learn to distinguish contrasting words from synonyms. The performance of the mLCM model trained without explicit contrast information about the dataset words proves moreover that lexical contrast information is indeed propagated through the lexical network.

4.3 Vector space structure

To further investigate the effect of lexical contrast information, we perform a qualitative analysis of how it affects the space structure. We pick 20 scalar adjectives denoting spatial or weight-related aspects of objects and living beings, where 10 indicate the presence of the relevant property to a great degree (*big, long, heavy...*), whereas the remaining 10 suggest that the property is present in little amounts (*little, short, light...*). We project the 300-dimensional vectors of these adjectives onto a 2-dimensional plane using the t-SNE toolkit,[5] which attempts to preserve the structure of the original high-dimensional word neighborhoods. Figure 1 shows that, in Skip-gram space, pairs at the extreme of the same scale (*light* vs *heavy*, *narrow* vs *wide*, *fat* vs *skinny*) are very close to each other compared to other words; whereas for mLCM the extremes are farther apart from each other, as expected. Moreover, the adjectives at the two ends of the scales are grouped together. This is a very nice property, since many adjectives in one group will tend to characterize the same objects. Within the two clusters, words that are more similar (e.g., *wide* and *broad*) are still closer to each other, just as we would expect them to be.

[5] http://lvdmaaten.github.io/tsne/



(a) Skip-gram space (b) mLCM space

Figure 1: Arrangement of some scalar adjectives in Skip-gram vs mLCM spaces

5 Learning Negation

Having shown that injecting lexical contrast information into word embeddings is beneficial for lexical tasks, we further explore if it can also help composition. Since mLCM makes contrasting and compatible words more distinguishable from each other, we conjecture that it would be easier for compositional DSMs to capture negation in mLCM space. We perform a proof-of-concept experiment where we represent *not* as a function that is trained to map an adjective to its antonym (*good* to *bad*). That is, by adopting the framework of Baroni et al. (2014), we take *not* to be a matrix that, when multiplied with an adjective-representing vector, returns the vector of an adjective with the opposite meaning. We realize that this is capturing only a tiny fraction of the linguistic uses of negation, but it is at least a concrete starting point.

First, we select a list of adjectives and antonyms from WordNet; for each adjective, we only pick the antonym of its first sense. This yields a total of around 4,000 antonym pairs. Then, we induce the *not* matrix with least-squares regression on training pairs. Finally, we assess the learned negation function by applying it to an adjective and computing accuracy in the task of retrieving the correct antonym as nearest neighbour of the *not*-composed vector, searching across all WordNet adjectives (10K items). The results in Table 3 are obtained by using 10-fold cross-validation on the 4,000 pairs. We see that mLCM outperforms Skip-gram by a large margin.

Figure 2 shows heatmaps of the weight matrices learnt for *not* by the two models. Intriguingly, for mLCM, the *not* matrix has negative values on the diagonal, that is, it will tend to flip the values in the input vector, not unlike what arithmetic negation would do. On the other hand, the Skip-gram-based *not* matrix is remarkably identity-like, with large positive values concentrated on the diagonal. Thus, under this approach, an adjective will be almost identical to its antonym, which explains why it fails completely on the test set data: the nearest neighbour of *not-X* will typically be X itself.

	train	test
Skip-gram	0.44	0.02
mLCM	**0.87**	**0.27**

Table 3: Average accuracy in retrieving antonym as nearest neighbour when applying the *not* composition function to 4,000 adjectives.

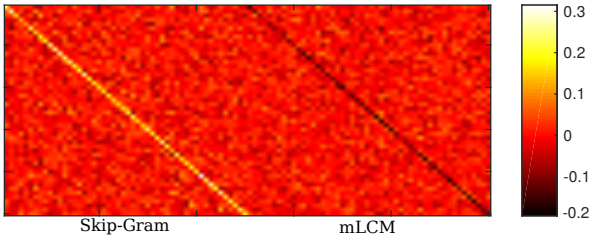


Figure 2: Heatmaps of *not*-composition matrices.

6 Conclusion

Given the promise shown by mLCM in the experiments reported here, we plan to test it next on a range of linguistically interesting phenomena that are challenging for DSMs and where lexical contrast information might help. These include modeling a broader range of negation types (de Swart, 2010), capturing lexical and phrasal inference (Levy et al., 2015), deriving adjectival scales (Kim and de Marneffe, 2013) and distinguishing semantic similarity from referential compatibility

(Kruszewski and Baroni, 2015).

7 Acknowledgments

This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Heike Adel and Hinrich Schütze. 2014. Using mined coreference chains as a resource for a semantic task. In *Proceedings of EMNLP*, pages 1447–1452, Doha, Qatar.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9(6):5–110.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Henriette de Swart. 2010. *Expression and Interpretation of Negation: an OT Typology*. Springer, Dordrecht, Netherlands.

Manaal Faruqui, Jesse Dodge, Sujay Jauhar, Chris Dyer, Ed Hovy, and Noah Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, Denver, CO. In press.

Daniel Fried and Kevin Duh. 2015. Incorporating both distributional and relational semantics in word representations. In *Proceedings of ICLR Workshop Track*, San Diego, CA. Published online: `http://www.iclr.cc/doku.php?id=iclr2015:main#accepted_papers`.

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. 2013. "Not not bad" is not "bad": A distributional account of negation. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pages 74–82, Sofia, Bulgaria.

Felix Hill, KyungHyun Cho, Sebastien Jean, Coline Devin, and Yoshua Bengio. 2014a. Not all neural embeddings are born equal. *arXiv preprint arXiv:1410.0718*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of EMNLP*, pages 1625–1630, Seattle, WA.

Germán Kruszewski and Marco Baroni. 2015. So similar and yet incompatible: Toward automated identification of semantically compatible words. In *Proceedings of NAACL*, pages 64–969, Denver, CO.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL*, pages 153–163, Denver, CO.

Omer Levy, Steffen Remus, Chris Biemann, , and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of NAACL*, Denver, CO. In press.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*, pages 746–751, Atlanta, Georgia.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad, Bonnie Dorr, Graeme Hirst, and Peter Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989, Denver, Colorado, May–June. Association for Computational Linguistics.

Anne Preller and Mehrnoosh Sadrzadeh. 2011. Bell states and negative sentences in the distributed model of meaning. *Electr. Notes Theor. Comput. Sci.*, 270(2):141–153.

Enrico Santus, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014. Taking antonymy mask off in vector space. In *Proceedings of PACLIC*, pages 135–144, Phuket,Thailand.

Sabine Schulte im Walde and Maximilian Köper. 2013. Pattern-based distinction of paradigmatic relations for German nouns, verbs, adjectives. In *Proceedings of GSCL*, pages 184–198, Darmstadt, Germany.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Peter Turney. 2008. A uniform approach to analogies, synonyms, antonyms and associations. In *Proceedings of COLING*, pages 905–912, Manchester, UK.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A general framework for incorporating knowledge into word representations. In *Proceedings of CIKM*, pages 1219–1228, Shanghai, China.

Wen-tau Yih, Geoffrey Zweig, and John Platt. 2012. Polarity inducing latent semantic analysis. In *Proceedings of EMNLP-CONLL*, pages 1212–1222.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550, Baltimore, MD.