Lexical Substitution for Evaluating Compositional Distributional Models

Maja Buljan* **Sebastian Padó*** **Jan Šnajder†**
* Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
{maja.buljan,pado}@ims.uni-stuttgart.de
† TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb
jan.snajder@fer.hr

Abstract

Compositional Distributional Semantic Models (CDSMs) model the meaning of phrases and sentences in vector space. They have been predominantly evaluated on limited, artificial tasks such as semantic sentence similarity on hand-constructed datasets. This paper argues for lexical substitution as a means to evaluate CDSMs. Lexical substitution is a more natural task, enables us to evaluate meaning composition at the level of individual words, and provides a common ground to compare CDSMs with dedicated lexical substitution models. We create a lexical substitution dataset for CDSM evaluation from an English-language corpus with manual "all-words" lexical substitution annotation. Our experiments indicate that the Practical Lexical Function CDSM outperforms simple component-wise CDSMs and performs on par with the context2vec lexical substitution model using the same context.

1 Introduction

Compositional Distributional Semantics Models (CDSMs) compute phrase meaning in semantic space as a function of the meanings of the phrase constituents (Baroni et al., 2014). The most basic CDSMs represent words as vectors and compose phrase vectors by component-wise operations of the constituent vectors (Mitchell and Lapata, 2008). More complex models represent predicates with matrices and tensors (Baroni and Zamparelli, 2010; Grefenstette, 2013; Paperno et al., 2014).

Given the large number of different CDSMs proposed in the literature (Erk, 2012), meaningful evaluation becomes crucial. The dominant evaluation method, adopted by the majority of CDSM studies, is pairwise phrase similarity (Mitchell and Lapata, 2008; Guevara, 2010; Grefenstette et al., 2012; Grefenstette, 2013; Paperno et al., 2014). Only a handful of studies pursued other evaluation tasks, such as textual entailment (Marelli et al., 2014a,b) or sentiment analysis (Socher et al., 2013).

Arguably, phrase similarity evaluation has three major problems. First, the task is affected by the general limitations of rating scales, such as inconsistencies in annotations, scale region bias, and fixed granularity (Schuman and Presser, 1996). Phrase similarity datasets used for CDSM evaluation demonstrate slight to fair inter-annotator agreement, as well as overlap between groups of items rated as *low* and *high* in similarity (Mitchell and Lapata, 2008).

Secondly, phrase similarity is a task that is rather difficult to put down precisely, especially for long phrases. Generally, phrases can be (dis)similar in any number of ways. Annotators commonly agree that some sentence pairs are semantically highly similar (*private company files annual account* and *private company registers annual account*, Pickering and Frisson 2001), and others are semantically unrelated (*man waves hand* vs. *employee leaves company*). In contrast, their assessments become less confident for cases like *delegate buys land* and *agent sells property* (Kartsaklis et al., 2013), where there is a semantic relation other than synonymy. Similarity is also arguably not a useful measure when sentences are semantically deviant, as it is often the case in the datasets: how similar are *private company files annual account* and *private company smooths annual account*?

The third problem is that the most widely used phrase similarity datasets are constructed in a balanced fashion along psycholinguistic principles. For instance, the adjective-noun-verb-adjective-noun ("ANVAN") dataset (Pickering and Frisson, 2001; Kartsaklis et al., 2013), from which the examples above are drawn, was constructed from a set of particularly ambiguous verbs paired with strongly disambiguating contexts. Such setups often do not

Proceedings of NAACL-HLT 2018, pages 206–211
New Orleans, Louisiana, June 1 - 6, 2018.

correlate well with usefulness on more natural data. In a previous study on lexical substitution, Kremer et al. (2014) found that the advantage of machine learning-based models over simple baselines was much harder to show on a real-world corpus than on the previously used manually constructed benchmark dataset (McCarthy and Navigli, 2009).

In this paper, we pursue the idea that lexical substitution (McCarthy and Navigli, 2009) is a more suitable evaluation task for CDSMs. Lexical substitution is the task of finding meaning-preserving substitutes for a target word in context: e.g., the word *submits* is a legitimate substitute for *files* in *private company files annual account*, but not in *office clerk files old papers*. Lexical substitution provides a frame for comparing synonyms in context, and disambiguating the context-appropriate sense of polysemous words. Since this is a problem CDSMs can account for, lexical substitution seems a suitable task for testing and comparing different CDSMs. Additionally, lexical substitution is a more natural task than similarity ratings, it makes it possible to evaluate meaning composition at the level of individual words, and provides a common ground to compare CDSMs with dedicated lexical substitution models.

2 The ANVAN-LS Dataset

To perform more realistic evaluations for CDSMs, we would like to test them on a sample of actual human utterances rather than hand-selected examples. That being said, for the evaluation to be useful, the structures on which we evaluate should be relatively uniform and limited, at least initially while moving from artificial to natural datasets. We thus construct a dataset[1] for English that strikes a balance between these factors: we maintain the adjective-noun-verb-adjective-noun (ANVAN) format, but our phrases are based on corpus sentences, and some or all words in the ANVAN can form the targets of lexical substitution ranking tasks.

Sampling ANVAN Queries. The starting point of our corpus construction is the English CoInCo corpus (Kremer et al., 2014), which consists of roughly 2,500 sentences taken from the *news* and *fiction* section of the MASC corpus and provides manually annotated lexical substitutes for all content words (nouns, adjectives, verbs, and adverbs).

[1]The corpus is freely available at http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/ANVAN-LS.html

We extracted all clauses from CoInCo that met the following requirements: (1) the dependency structure of the clause includes an ANVAN structure; (2) at least one constituent word of the ANVAN sub-clause has at least two human-provided single-word substitutes that exceeded a minimal frequency threshold (more than 5 occurrences in a large corpus, cf. Section 3); (3) the POS tags were correct. This resulted in 165 ANVAN phrases, which contained an average of 4.4 (out of 5) target words with substitutes, for a total of 732 target words for lexical substitution.

An issue that required additional consideration was the adjective positions of the ANVANs. In the CoInCo, we found a substantial number of noun-noun compounds whose modifiers were tagged as adjectives. Conversely, many adjective modifiers in the corpus were substituted with nouns by human annotators. To account for this variability, we extended the ANVAN schema conservatively by allowing nouns to fill the A position if it was observed in the large corpus robustly as a modifier (i.e., as part of at least 100 N-N bigram types, each of which occurred at least 300 times).

Building Lexical Substitution Tasks. For each of the 732 target words, we constructed a lexical substitution query by pairing the target with two correct substitutes and two confounders (see Table 1). Since substitutes provided by more annotators in CoInCo tend to be more reliable, we picked the two most frequently given substitutes for each target. In the case of ties, we chose the lemma with the highest corpus frequency.

To acquire challenging confounders, we retrieved the 20 most similar lemmas with the same part of speech for each target (according to the unigram space; cf. Section 3) and then eliminated all annotator-provided substitutes for this target. From the remainder, we chose the two most closely matched by corpus frequency to the frequencies of the two chosen annotator-provided substitutes. Given the relatively high number of human substitutes in CoInCo, this results in highly similar, but contextually inappropriate confounders. Finally, the acquired confounders were manually checked to make sure that the automatic selection process did not yield a context-appropriate substitute or a semantically unrelated term. In such cases, the next best candidate (by lemma similarity and frequency) was chosen.

target	construction	**arm**	build	large	airfield
substitute$_1$	construction	*branch*	build	large	airfield
substitute$_2$	construction	*part*	build	large	airfield
confounder$_1$	construction	*back*	build	large	airfield
confounder$_2$	construction	*hand*	build	large	airfield

Table 1: ANVAN-LS lexical substitution example

3 Experimental Setup

Task and Evaluation. We evaluate models on ANVAN-LS in the form of a ranking task for each query: models are supposed to rank the correct substitutes for each target higher than its confounders. Our evaluation measure is the mean average precision (MAP) of all queries. In our case,

$$\text{MAP} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{4} P_{l_i}(k)\Delta r_{l_i}(k)$$

where k is the rank in the 4-item list l of substitution candidates, $P_l(k)$ is the precision at cut-off point k in list l, the $\Delta r_l(k)$ is the change in recall from items $k-1$ to k in l, and N is the total number of ANVAN queries. We calculate MAP both overall, and by target positions in ANVAN, to obtain more detailed insights into performance depending on part of speech and word position.

Corpus and Semantic Space. Following Baroni and Zamparelli (2010) and Paperno et al. (2014), we use a concatenation of the ukWaC, BNC, and English Wikipedia corpora (around 2.8G words). We build a square co-occurrence matrix, using the complete vocabulary of our ANVAN sentence dataset, and a set of the most frequent content words in our corpus, for a total of 30K words. In addition, we built two co-occurrence matrices of bigrams, composed of all predicates (adjectives and verbs) in our vocabulary and their most frequent noun co-occurrents, as observed in the corpus, with a threshold of 300. The bigrams were observed in co-occurrence with the aforementioned 30K vocabulary and most frequent corpus terms, resulting in a 650K-by-30K matrix for A-N and N-N bigrams and a 620K-by-30K matrix for V-N bigrams.

For all three matrices, the 3-word-window co-occurrence counts were transformed with PPMI and reduced to 300 dimensions with SVD. All models were built with DISSECT (Dinu et al., 2013).

CDSMs. We consider two component-wise CDSMs: the simple additive and multiplicative models (Mitchell and Lapata, 2008), defined as

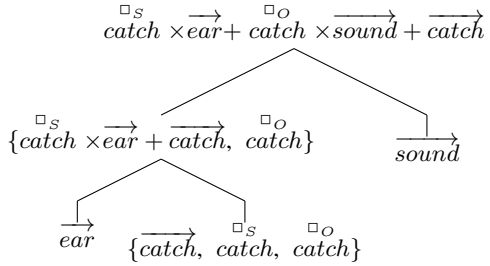

Figure 1: Example for Practical Lexical Function model composition for *ear catch sound*

$\mathbf{p} = \mathbf{u} \oplus \mathbf{v}$, where $\oplus$ is either component-wise addition or multiplication. We also work with two variants of the Practical Lexical Function model (PLF), which are derived from the Lexical Function Model (LF, Baroni and Zamparelli (2010)), which represents predicates (verbs and adjectives in our case) as matrices to be multiplied with vector representations of their nominal arguments. More specifically, when applied to an ANVAN sentence (such as *pointed ear catch sharp sound*), the PLF model incorporates vector representations for each of the five constituent words, along with an adjective matrix for *pointed* and *sharp*, as well as $\text{verb}_{\text{subject}}$ and $\text{verb}_{\text{object}}$ matrices for the verb and its subject and object arguments. An example of composition for a verb with its subject and object arguments is given in Figure 1.

Formally, the standard PLF ($\text{PLF}_{\text{Paperno}}$) defines the composition for ANVAN-style sentences as: $V_S^{\square} \cdot \overrightarrow{S} + V_O^{\square} \cdot \overrightarrow{O} + \overrightarrow{V}$, where $\overrightarrow{S}$ and $\overrightarrow{O}$ are the composed A-N bigrams, $A^{\square} \cdot \overrightarrow{N}$. The required matrices are learned with ridge regression from unigram and bigram vectors.

Gupta et al. (2015) pointed out that the standard PLF "overcounts" the predicate by adding it explicitly, and proposed a rectified variant which simply leaves out the function word vector $\overrightarrow{V}$. We also experiment with this model, $\text{PLF}_{\text{Gupta}}$.

All CDSMs rank the four candidates for each target (cf. Table 1) by comparing the vector for the original sentence against four sentences in which the target is replaced by the two correct and two incorrect substitutes. We use the raw dot product as similarity measure, following Roller and Erk (2016), to boost frequent candidates.

Lexical Substitution Model. As competitor, we consider a dedicated lexical substitution model, namely context2vec (Melamud et al., 2016). Since it has demonstrated state-of-the-art performance on lexical substitution and word sense disambiguation

tasks, it is a suitable competitor model for CDSMs on a similar problem. Context2vec uses word embeddings to compute a set of viable substitutes given a context, using a bidirectional LSTM recurrent neural network to build a sentential context representation. We work with two instantiations: first, using only ANVAN as context ($C2V_{ANVAN}$), and second, using the full CoInCo sentence from which the ANVAN was extracted ($C2V_{Sent}$). The first model is directly comparable to the CDSMs in that it uses the same context information, while the second one enables us to gauge to what extent the models can benefit from a richer context. We let context2vec generate 1000 substitutions for each target word. If any substitution candidates were not included in this list, the missing items were defined to be tied at the last position, and the AP was defined as the MAP of all permutations of the missing items with respect to their ranking.

Baselines. Two baselines are defined to compare our models against. The first is the random baseline (Random): the MAP of all possible rankings of two relevant and two irrelevant items, equal to 0.680. The second baseline (LemmaSim) ranks the lemma-level similarities of the target word and its substitutes candidates without context.

4 Results

Table 2 lists the performance of our experiments on ANVAN-LS, first evaluated for each target word position (top rows) and then averaged across all target positions (bottom row). The rightmost column shows the average performance of all tested models. Even though there is some variance in the numbers between subject-position targets and object-position targets, we see consistent patterns by part of speech: adjectives are easiest to substitute, followed by nouns, while verbs are substantially more difficult. The difficulty with verbs corresponds to the findings of Medić et al. (2017) for Croatian, who proposed a correlation between the syntactic valence of words and their difficulty (but see below).

The worst model by far is LemmaSim. This is surprising, given that Kremer et al. (2014) found lemma-level similarity to be very competitive. Further analysis showed that its bad performance is due to our choice of confounders as highly similar lemmas (cf. Section 2). An example where high similarity indicates syntagmatic rather than paradigmatic relatedness is *ohio democrat embark overland* ***trip****/*itinerary/*sightseeing/journey/travel*,[2] where the two confounders can form noun compounds with the target, like *sightseeing trip*. The simple Add and Mult models also perform worse than random, in contrast to many other studies, underscoring the difficulty of performing well on our dataset. They do relatively well for adjectives, but worse on nouns, and Add struggles with verbs.

The $PLF_{Paperno}$ model performs at baseline level overall, but shows particularly clear differences among parts of speech. It does very well for all adjectives and nouns, but very poorly on verbs, where it is in fact the worst model overall, both measured absolutely, and relatively to the overall performance of the model. An analysis showed that this is indeed, as Gupta et al. (2015) claim, due to the overpowering effect of the predicate vector which is added in the final step of the composition and thus tends to dominate the composed phrase. Consequently, $PLF_{Paperno}$ essentially falls back to verb lemma similarity (cf. the example *russian team* ***win****/earn/*clinch/get/*succeed gold medal*).

PLF_{Gupta} performs comparable to $PLF_{Paperno}$ in all positions, but demonstrates a significant improvement on verbs, for example *russian team* ***win****/earn/get/*clinch/*succeed gold medal*. The PLF_{Gupta} model also outperforms the baselines, overall and for all individual positions, it emerges as the best performing CDSM.

Finally, the two context2vec models beat the baseline both on all positions and overall. Interestingly, $C2V_{ANVAN}$, which uses the same information provided to the CDSMs, performs roughly on par with PLF_{Gupta}, the best-scoring CDSM. Evidently, Context2Vec (as a lexical substitution model) and PLF (as a CDSM) perform comparably – it is not the case that one of the two model families shows a clear advantage over the other, given the same context. $C2V_{Sent}$, which has access to richer context, shows another substantial improvement. An interesting difference between PLF and Context2Vec is that the lexical substitution models – which do not take syntactic structure into account – show less variance among positions than the PLF, which does take syntactic structure into account.

In their analysis of CDSM performance on a Croatian language ANVAN dataset, Medić et al. (2017) found a superior performance of simple

[2]The substitute candidates are ordered by decreasing similarity, and the confounders marked with asterisks.

	Baselines		CDSMs				Lexical substitution		All
Position	Random	LemmaSim	Add	Mult	$PLF_{Paperno}$	PLF_{Gupta}	$C2V_{ANVAN}$	$C2V_{Sent}$	Average
ANVAN	.680	.680	.716	.715	.730	.727	.694	.707	.706
A**N**VAN	.680	.575	.652	.633	.695	.688	.708	.744	.672
AN**V**AN	.680	.537	.618	.670	.536	.680	.697	.723	.643
ANV**A**N	.680	.625	.668	.668	.721	.715	.690	.710	.685
ANVA**N**	.680	.580	.633	.666	.725	.723	.723	.772	.688
Average	.680	.599	.656	.669	.681	.706	.702	.731	.678

Table 2: Evaluation results on ANVAN-LS (mean average precision)

CDSMs (such as addition) for nouns while the PLF performed better on verbs. They attributed this to the role of valence, arguing that the functional role of the verbs, and the disambiguation potential of its argument positions, is better captured by the PLF. In contrast, the variance in performance between word positions that we find for the different models on the ANVAN-LS dataset indicates that the difficulty of substituting verbs might not be due to the intrinsic factor of valence, but due to remaining shortcomings in all CDSM models to properly model predicate-argument combination.

5 Conclusion

This paper presented the case for using lexical substitution as a better evaluation setup for compositional distributional semantic models (CDSMs). We created a new corpus, ANVAN-LS, on the basis of a corpus with manually annotated lexical substitution, and evaluated a battery of models. Our evaluation on this corpus (1) uses a corpus-based, rather than manually constructed, dataset, and should be more indicative for the performance of models on "real-world data" than previous ANVAN-based evaluations; (2) is challenging, with a high baseline, which simple CDSMs like component-wise addition and multiplication were indeed not able to beat; (3) enables a detailed evaluation at a per-position level; (4) makes it possible, to our knowledge for the first time, to compare CDSMs with dedicated lexical substitution models on par, and shows that the two model families perform comparably when using the same context, but differ in their performance by position.

The last result in particular opens up a new line of research, namely an investigation of similarities and differences between the two model families. The improvement we observe for $C2V_{Sent}$ over $C2V_{ANVAN}$ in particular calls for a move from ANVAN-style sentences to more complex and varied sentence structures. It remains to be researched how capable CDSMs are to model meaning modulation that extends beyond the immediate predicate-argument structure (Kremer et al., 2014).

Acknowledgments. We acknowledge partial funding by Deutsche Forschungsgemeinschaft through SFB 732 (project D10) and by the Croatian Science Foundation under the project UIP-2014-09-7312. We would also like to thank Domagoj Alagić for his support.

References

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

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*. Cambridge, MA, pages 1183–1193.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT – distributional semantics composition toolkit. In *Proceedings of ACL*. Sofia, Bulgaria, pages 31–36.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass* 6(10):635–653.

Edward Grefenstette. 2013. Category-theoretic quantitative compositional distributional models of natural language semantics. *arXiv preprint arXiv:1311.1539* .

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2012. Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS 2012*. Potsdam, Germany.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*. Uppsala, Sweden, pages 33–37.

Abhijeet Gupta, Jason Utt, and Sebastian Padó. 2015. Dissecting the practical lexical function model for compositional distributional semantics. In *Proceedings of STARSEM*. Denver, Colorado, pages 153–158.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of CoNLL*. Sofia, Bulgaria, pages 114–123.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us-analysis of an "all-words" lexical substitution corpus. In *Proceedings of EACL*. Gothenburg, Sweden, pages 540–549.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of SemEval*. Dublin, Ireland, pages 1–8.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*. Reykjavík, Iceland, pages 216–223.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation* 43(2):139–159.

Zoran Medić, Jan Šnajder, and Sebastian Padó. 2017. Does free word order hurt? Assessing the Practical Lexical Function model for Croatian. In *Proceedings of STARSEM*. Vancouver, BC, pages 115–120.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of CONLL*. Berlin, Germany, pages 51–61.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*. Columbus, OH, pages 236–244.

Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of ACL*. Baltimore, MD, pages 90–99.

Martin Pickering and Steven Frisson. 2001. Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(2).

Stephen Roller and Katrin Erk. 2016. Pic a different word: A simple model for lexical substitution in context. In *Proceedings of NAACL/HLT*. San Diego, CA, pages 1121–1126.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*. Seattle, WA, pages 1631–1642.