Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces

Maximilian Köper, Christian Scheible, Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
{koepermn,scheibcn,schulte}@ims.uni-stuttgart.de

Abstract

While continuous word vector representations enjoy increasing popularity, it is still poorly understood (i) how reliable they are for other languages than English, and (ii) to what extent they encode deep semantic relatedness such as paradigmatic relations. This study presents experiments with continuous word vectors for English and German, a morphologically rich language. For evaluation, we use both published and newly created datasets of morpho-syntactic and semantic relations. Our results show that (i) morphological complexity causes a drop in accuracy, and (ii) continuous representations lack the ability to solve analogies of paradigmatic relations.

1 Introduction

Until recently, the majority of research on semantic spaces concentrated on vector spaces relying on context counts (*count vector spaces*). However, increasing attention is being devoted to low-dimensional continuous word vector representations. Unlike count vectors, these continuous vectors are the result of supervised training of context-predicting models (*predict vector spaces*).[1]

Mikolov et al. (2013) reported that a predict vector space trained with a simplified neural language model (cf. Bengio et al. (2003)) seemingly encodes syntactic and semantic properties, which can be recovered directly from the space through linear translations, to solve analogies such as

$$\overrightarrow{\text{king}} - \overrightarrow{\text{man}} = \overrightarrow{\text{queen}} - \overrightarrow{\text{woman}}.$$

Baroni et al. (2014) presented experiments where predict vectors outperform count vectors on several semantic benchmarks involving semantic relatedness, word clustering, and selectional preferences.

Several open questions regarding predict vectors remain. In this paper, we focus on two shortcomings of previous analyses. First, the analogies in the "syntactic" and "semantic" benchmark datasets by Mikolov et al. (2013) in fact cover mostly morpho-syntactic relations – even in the semantic category. Consequently, it is still unknown to what extent predict vector spaces encode deep semantic relatedness, such as paradigmatic relations. Rei and Briscoe (2014) offered some insight by testing hypernymy relations through similarity; Melamud et al. (2014) investigated synonymy, hypernymy, and co-hyponymy relations. However, no systematic evaluation of deep semantic analogies has been performed so far.

Second, it remains unclear whether comparable performance can be achieved for a wider range of relations in morphologically rich languages, as most previous work on predict vectors worked with English data. A notable exception is Zuanović et al. (2014), who achieved strong performance for superlative and country-capital analogies in Croatian. Wolf et al. (2013) learned mappings of predict vectors between English, Hebrew, and Arabic, but provided no deeper insight into the model's capabilities on a direct evaluation of semantic relations. Faruqui and Dyer (2014) trained predict vectors using two languages, but evaluated only in English.

We present a systematic exploration of morpho-syntactic and semantic relatedness in English and the morphologically richer language German. We show detailed results of the continuous bag-of-words model (CBOW) by Mikolov et al. (2013), which we apply to equivalent morpho-syntactic tasks for both

[1] The terminology follows Baroni et al. (2014).

Proceedings of the 11th International Conference on Computational Semantics, pages 40–45,
London, UK, April 15-17 2015.

languages. Pertaining to the question of deep semantic relatedness, we evaluate on existing benchmarks on general semantic relatedness, and on newly created paradigmatic semantic analogies. To make the models for the two languages as comparable as possible, they are trained on web corpora which were obtained with the same crawling technique, and which we subsample to comparable size.

We present evidence that – while general semantic relatedness is captured well by predict models – paradigmatic relations are problematic for count vector spaces. Moreover, our experiments on German show that its morphological richness does indeed make the prediction of analogies more difficult.

2 Data

2.1 Morpho-Syntactic and Semantic Tasks

We evaluate a variety of analogy and semantic relatedness tasks, 23 for English and 21 for German. They are in part taken from the literature and in part newly constructed.[2]

The **Google semantic/syntactic** analogy datasets *(Google-Sem/Syn)* were introduced in Mikolov et al. (2013). The datasets contain analogy questions of the form A:B::C:D, meaning A is to B as C is to D, where the fourth word (D) is unknown. We constructed German counterparts of the datasets through manual translation and subsequent cross-checking by three human judges. We omitted the relation type "adjective–adverb" for both languages, because it does not exist in German. The final task set contains five *Google-Sem* and eight *Google-Syn* relation types with 18 552 analogy tasks per language.

The **paradigmatic semantic relation** dataset *(Sem-Para)* also contains analogy tasks. Here, the paradigmatic relation between A and B is the same as between C and D. The dataset was constructed from antonymy, synonymy, and hypernymy relation pairs collected by Lenci & Benotto for English and by Scheible & Schulte im Walde for German, using the methodology described in Scheible and Schulte im Walde (2014): Relying on a random selection of target nouns, verbs and adjectives from WordNet/GermaNet – balanced for semantic class, degree of polysemy, and frequency according to the WaCKy corpora (Baroni et al., 2009) –, antonyms, synonyms, and hypernyms were collected in an experiment hosted on Amazon Mechanical Turk. We constructed analogy questions by selecting only those target-response pairs that were submitted by at least four out of ten turkers. Then, we exhaustively combined all pairs for each word class and relation type.[3] The resulting English dataset contains 7 516 analogies; the German dataset contains 2 462 analogies.

In the same way, we created an analogy dataset with 10 000 unique analogy questions from the hypernymy and meronymy relations in *BLESS* (Baroni and Lenci, 2011), by randomly picking semantic relation pairs. BLESS is available only for English, but we included it in *Sem-Para* as it is a popular semantic benchmark.

Overall, the *Sem-Para* dataset constitutes a deep semantic challenge, containing very specific, domain-related and potentially low-frequent semantic details that are difficult to solve even for humans. For example, the tasks include antonyms such as *biblical:secular::deaf:hearing* or *screech:whisper::ink:erase*; hypernyms such as *groove:dance::maze:puzzle*; and synonyms such as *skyline:horizon::rumor:gossip*.

The **general semantic** dataset *(Sem-Gen)* does not require to solve analogies but to predict the degree of semantic relatedness between word pairs. It contains three semantic benchmarks:

1. *RG* (Rubenstein and Goodenough, 1965) and its German equivalent *Gur65* (Gurevych, 2005).
2. *WordSim353* (Finkelstein et al., 2001) and its translation into German *WordSim280* by Schmidt et al. (2011): As Schmidt et al. did not re-rate the German relation pairs after translation (which we considered necessary due to potential meaning shifts), we collected new ratings for the German pairs from 10 subjects, applying the same conditions as the original WordSim353 collection task. To ensure identical size for both languages, we reduced the English data to the common 280 pairs.

[2]The new datasets are available at `http://www.ims.uni-stuttgart.de/data/analogies/`.

[3]Regarding hypernymy and meronymy (see BLESS below), we restricted the pair combination such that the word to be predicted is always the hypernym or holonym, respectively. The reason for this restriction is that there are too many correct choices for the corresponding hyponyms and meronyms.

	Google-Sem			Google-Syn			Sem-Gen			Sem-Para w/o BLESS			TOEFL		
	CBOW	SKIP	BOW	CBOW	SKIP	BOW	CBOW	SKIP	BOW	CBOW	SKIP	BOW	CBOW	SKIP	BOW
EN W	68.8	**71.8**	39.5	**81.9**	80.5	57.9	**77.9**	77.8	77.8	**19.3**	16.4	15.6	**96.2**	**96.2**	72.2
EN L	68.3	**71.8**	40.3	47.1	**47.4**	29.3	**80.5**	78.6	66.4	18.4	**15.9**	15.8	**90.0**	87.5	66.2
DE W	42.4	**45.9**	27.3	**48.4**	47.1	31.0	**75.6**	73.3	58.9	14.7	14.4	**14.8**	**69.0**	68.3	54.4
DE L	43.5	**45.9**	28.9	**31.8**	31.5	23.7	73.3	**75.7**	64.7	**15.1**	13.8	14.9	**69.4**	68.5	55.8

Table 1: Results (ρ for Sem-Gen, accuracy for others) by task category across models.

3. 80 *TOEFL* (Test of English as a Foreign Language) questions by Landauer and Dumais (1997) for English, and 426 questions from a similar collection by Mohammad et al. (2007) for German. Each semantic similarity question is multiple choice, with four alternatives for a given stem. Unlike the original English TOEFL data, the German dataset also contains phrases, which we disregarded.

2.2 Corpora

We obtain vectors using the *COW* web corpora *ENCOW14* for English and *DECOW12* for German (Schäfer and Bildhauer, 2012). The corpora contain lemma and part-of-speech annotations. In addition, we applied some basic pre-processing: we removed non-alphanumeric tokens and sentences with fewer than four words, and we lowercased all tokens. In order to limit effects of corpus size, we subsampled the English corpus to contain approximately the same number of tokens as the German corpus, 7.9 billion.

3 Experiments

3.1 Setup and Evaluation

Our setups vary model type (two predict models and one count model), language (English and German), and word forms vs. lemmas in the training data – leading to a total of $3\times2\times2$ models. Our predict models are the standard CBOW and SKIP-gram models, trained with the `word2vec` toolkit (Mikolov et al., 2013). We use negative sampling with 15 negative samples, 400 dimensions, a symmetrical window of size 2, subsampling with $p = 10^{-5}$, and a frequency threshold of 50 to filter out rare words.

Our count model is a standard bag-of-words model with positive point-wise mutual information weighting and dimensionality reduction through singular value decomposition. The dimensionality and the window size were set identical to the predict vectors.

We solve analogy tasks with the 3COSMUL method (Levy and Goldberg, 2014), and similarity tasks with cosine similarity. For the *Google*, *TOEFL*, and *Sem-Para* tasks, we report accuracy; for *RG* and *WordSim* we report Spearman's rank-order correlation coefficient ρ.

3.2 Results

Table 1 compares the word-based (W) and lemma-based (L) results of the English (EN) and the German (DE) ***predict vs. count models***. We first confirm previous insight (Baroni et al., 2014) that the predict models (CBOW; SKIP) in most cases outperform the count models (BOW). Second, we also confirm that the SKIP-gram model outperforms CBOW only on *Google-Sem* (Mikolov et al., 2013). Third, we find that lemmatized models generally perform slightly better on semantic tasks, whereas full word forms are necessary for morpho-syntactic tasks. Table 2 presents a breakdown by task for the overall best model (CBOW). Based on this, we will now discuss our two main questions.

(i) Morphological richness of target language: For the *Google-Sem/Syn* analogies, the level of performance is generally higher in English than in German. The only exceptions are the tasks *nationality-adjective* (L), and *plural-verbs* (both W+L). Our experiments demonstrate that, compared to English, the *Google* analogies are more difficult to solve for the morphologically richer language German. Using full

	Google-Sem (Acc)					Google-Syn (Acc)								Sem-Para (Acc)							Sem-Gen (ρ)		Sem-Gen (Acc)
	common-countries	capital-world	currency	city-in-state	family	opposite	nationality-adjective	comparative	superlative	plural-nouns	plural-verbs	present-participle	past-tense	adj-ant	verb-ant	noun-ant	noun-syn	noun-hyp	BLESS-hyp	BLESS-mer	RG/Gur65	WordSim280	TOEFL
EN W	94.0	74.6	19.5	67.5	83.3	50.2	85.5	95.4	94.8	92.4	80.9	77.6	68.6	11.6	1.3	0.4	4.3	0.8	1.0	0.0	82.3	73.6	96.2
EN L	92.6	73.1	21.4	70.0	71.9	49.5	84.8	56.0	46.3	67.3	15.8	39.1	6.6	12.5	3.8	0.0	7.3	1.3	1.9	0.0	84.3	76.7	90.0
DE W	82.0	55.8	14.9	17.7	60.5	23.1	40.3	69.8	37.9	73.8	83.9	15.4	53.5	4.2	0.0	5.0	5.9	1.1	–	–	75.1	76.1	69.0
DE L	81.8	58.8	17.5	17.5	60.7	21.4	85.1	14.8	7.9	63.0	37.7	17.7	1.3	5.3	0.3	3.6	8.6	1.1	–	–	79.1	76.7	69.4

Table 2: Results by task for the English and German CBOW models.

word forms, these differences are consequently the strongest for the *Google-Syn* morpho-syntactic tasks[4] *opposite, comparative, superlative, plural-nouns, present-participle*, and *past-tense*, where considerably more word forms per lemma exist in German than in English. As a consequence, the German search space is larger, and it becomes more difficult to predict the correct form. For example, while English only uses three adjective word forms per lemma, i.e., positive, comparative and superlative (e.g., *fast, faster, fastest*), German inflects adjectives for case, gender and number (e.g., *schneller(e|en|er|es)* are all valid translations of *faster*). The results for *nationality-adjective* confirm this insight, because the lemma-based (L) German data with a reduced search space (i.e., only offering one adjective lemma instead of the various inflected forms) clearly improves over the word-based German version (40.3% → 85.1%). Regarding *plural-verbs*, we assume that the German task is not more difficult than the English task, because even though German verbs are also inflected, written language predominantly uses two verb forms (third person singular and plural), as in English.

(ii) Deep semantic tasks: First, we contrast the *Google* tasks with varying morpho-syntactic and light semantic content against the semantic relation tasks *Sem-Gen* and the deep semantic tasks *Sem-Para*. We observe that performance across models and languages is still high when addressing semantic relatedness on a coarse-grained level (*Sem-Gen*): This is true when the number of related pairs is comparably low, and the relation types differ more strongly (*RG* and *WordSim*), or when the search space is very restricted (*TOEFL*, which is a multiple choice task). However, accuracy is dramatically low when deep semantic knowledge is required, as in *Sem-Para*. Only *adj-ant* and *noun-syn* achieve accuracy scores of over 5.0% for both languages. In most cases, lemmatization slightly helps by reducing the search space, because distinguishing between word forms is not required by the tasks. Yet, the gain is lower than we had expected due to lemmatization errors on the web data, which led to a considerable set of full inflected forms still being part of the search spaces.

Data analysis reveals the following major error types in the *Sem-Para* task category: Next to a minority of clearly wrong solutions, the CBOW model suggested wrong words/lemmas that are nevertheless related to the requested solution, either morphologically or semantically. An example for a wrong but morphologically similar solution is *Freiheit (freedom)* instead of *gefangen (caught)* as the prediction for *unfruchtbar:fruchtbar::frei:? (sterile:fertile::free:?)*. Examples for wrong but semantically similar solutions are the hyponym *Holzstuhl (wooden chair)* instead of the hypernym *Möbel (furniture)* for *Atomwaffe:Waffe::Stuhl:? (atomic weapon:weapon::chair:?)*; the synonym *erhöhen (increase)* instead of the antonym *abfallen (decrease)* for *verbieten:erlauben::ansteigen:? (forbid:allow::increase:?)*; and the synonym *undetermined* instead of the antonym *known* for *manual:automatic::unknown:?*. Overall, wrong semantic suggestions are most often synonyms (instead of hypernyms or antonyms).

Morphological variation is again a more serious problem for the German data, not only regarding inflection but also regarding composition: many wrong solutions are compounds suggested for their heads (as in the *Stuhl–Holzstuhl* example above). Further examples of this type of error are *Cayennepfeffer (cayenne pepper)* instead of *Salz (salt)* as the antonym of *Pfeffer (pepper)*; and *Lufttemperatur (air temperature)* instead of *Wärme (warmth)* as the synonym of *Temperatur (temperature)*.

[4]The performance gap on the *Google-Sem* tasks is smaller. An exception is *city-in-state*, where this gap may be attributed to better coverage of American cities in English.

	Sem-Para (Rec10)						
	adj-ant	verb-ant	noun-ant	noun-syn	noun-hyp	BLESS-hyp	BLESS-mer
	CBOW						
EN W	25.5	7.7	3.9	**29.1**	7.9	4.6	0.6
EN L	23.6	**9.1**	4.3	26.8	9.0	5.6	0.6
DE W	14.4	4.2	17.5	27.3	4.9	–	–
DE L	15.1	7.1	16.1	27.1	6.2	–	–
	SKIP						
EN W	**25.7**	7.2	3.4	21.6	5.0	4.0	0.6
EN L	23.7	6.7	3.2	21.9	5.7	5.4	0.8
DE W	15.2	2.9	17.1	24.2	5.8	–	–
DE L	**15.5**	2.9	16.8	22.3	3.9	–	–
	BOW						
EN W	24.9	7.1	6.1	21.0	18.6	6.1	1.9
EN L	16.4	6.4	**6.7**	20.3	**19.6**	**8.5**	**2.4**
DE W	6.3	**7.8**	**28.3**	26.8	4.9	–	–
DE L	8.5	5.8	22.8	**31.0**	**6.8**	–	–

Table 3: *Sem-Para* results across models, for recall at ten.

Table 3 compares the *Sem-Para* results across models, now relying on recall of the target being in the top 10 (Rec10). We consider this measure a fairer choice than accuracy because (a) the *Sem-Para* dataset contains considerably more difficult tasks, and (b) the higher proportions allow a better comparison across conditions. Bold font indicates the best results per column and language. Similar to before, the best results are reached for *adj-ant* and *noun-syn*, as well as for *noun-ant*, with Rec10 between 25.7% and 31.0%. Performance on *noun-hyp* reaches $> 15\%$ in only two cases, and the *verb-ant* and *BLESS* results are always $< 10.0\%$ for both languages and W/L conditions. Furthermore, there is no clear tendency for one of the languages or W vs. L to outperform the other. It is clear, however, that the superiority of the CBOW model in comparison to BOW vanished: in most cases, the BOW models outperform the CBOW (and SKIP) models, most impressively for *noun-ant* and *noun-hyp*.

4 Conclusion

We presented a systematic cross-lingual investigation of predict vectors on morpho-syntactic and semantic tasks. First, we showed that their overall performance in German, a morphologically richer language, is lower than in English. Second, we found that none of the vector spaces encodes deep semantic information reliably: In both languages, they lack the ability to solve analogies of paradigmatic relations.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580 (Sabine Schulte im Walde).

References

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43*(3), 209–226.

Baroni, M., G. Dinu, and G. Kruszewski (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238–247.

Baroni, M. and A. Lenci (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP Workshop on Geometrical Models for Natural Language Semantics*, pp. 1–10.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin (2003). A neural probabilistic language model. *Journal of Machine Learning Research 3*, 1137–1155.

Faruqui, M. and C. Dyer (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on the World Wide Web*, pp. 406–414.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pp. 767–778.

Landauer, T. K. and S. 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), 211–240.

Levy, O. and Y. Goldberg (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, pp. 171–180.

Melamud, O., I. Dagan, J. Goldberger, I. Szpektor, and D. Yuret (2014). Probabilistic modeling of joint-context in distributional similarity. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, pp. 181–190.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119.

Mikolov, T., W.-t. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751.

Mohammad, S., I. Gurevych, G. Hirst, and T. Zesch (2007). Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 571–580.

Rei, M. and T. Briscoe (2014). Looking for hyponyms in vector space. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, pp. 68–77.

Rubenstein, H. and J. B. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM 8*(10), 627–633.

Schäfer, R. and F. Bildhauer (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 486–493.

Scheible, S. and S. Schulte im Walde (2014). A database of paradigmatic semantic relation pairs for German nouns, verbs, and adjectives. In *Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, pp. 111–119.

Schmidt, S., P. Scholl, C. Rensing, and R. Steinmetz (2011). Cross-lingual recommendations in a resource-based learning scenario. In *Towards Ubiquitious Learning, Proceedings of the 6th European Conference on Technology Enhanced Learning*, pp. 356–369.

Wolf, L., Y. Hanani, K. Bar, and N. Dershowitz (2013). Joint `word2vec` networks for bilingual semantic representations. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*.

Zuanović, L., M. Karan, and J. Šnajder (2014). Experiments with neural word embeddings for Croatian. In *Proceedings of the 9th Language Technologies Conference*, pp. 69–72.