

# Semantic Relatedness for All (Languages): A Comparative Analysis of Multilingual Semantic Relatedness using Machine Translation

André Freitas<sup>1</sup>, Siamak Barzegar<sup>2</sup>, Juliano Efon Sales<sup>1</sup>,  
Siegfried Handschuh<sup>1</sup> and Brian Davis<sup>1</sup>

<sup>1</sup> Department of Computer Science and Mathematics - University of Passau  
Innstrasse 43, ITZ-110, 94032 Passau, Germany  
{andre.freitas, juliano-sales, siegfried.handschuh}@uni-passau.de  
<sup>2</sup> Insight Centre for Data Analytics - National University of Ireland, Galway  
IDA Business Park, Lower Dangan, Galway, Ireland  
{siamak.barzegar, brian.davis}@insight-centre.org

**Abstract.** This paper provides a comparative analysis of the performance of four state-of-the-art distributional semantic models (DSMs) over 11 languages, contrasting the native language-specific models with the use of machine translation over English-based DSMs. The experimental results show that there is a significant improvement (average of 16.7% for the Spearman correlation) by using state-of-the-art machine translation approaches. The results also show that the benefit of using the most informative corpus outweighs the possible errors introduced by the machine translation. For all languages, the combination of machine translation over the Word2Vec English distributional model provided the best results consistently (average Spearman correlation of 0.68).

**Keywords:** multilingual distributional semantics, machine translation.

## 1 Introduction

Distributional Semantic Models (DSM) are consolidating themselves as fundamental components for supporting automatic semantic interpretation in different application scenarios in natural language processing. From *question answering systems*, to *semantic search* and *text entailment*, distributional semantic models support a scalable approach for representing the meaning of words, which can automatically capture comprehensive associative commonsense information by analysing word-context patterns in large-scale corpora in an unsupervised or semi-supervised fashion[8,19,18].

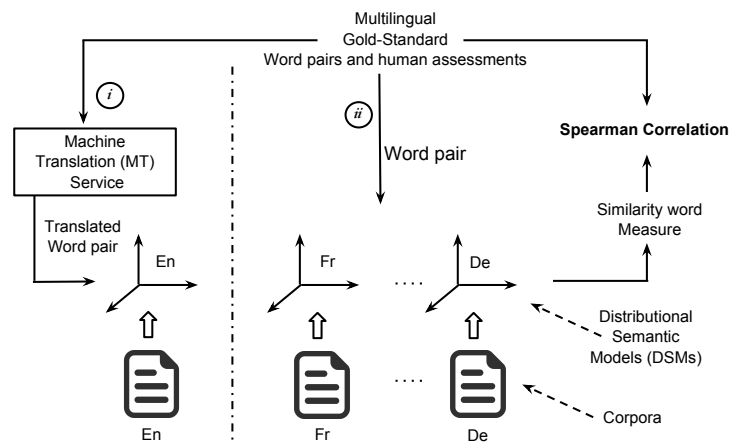
However, distributional semantic models are strongly dependent on the size and the quality of the reference corpora, which embeds the commonsense knowledge necessary to build comprehensive models. While high-quality texts containing large-scale commonsense information are present in English, such as Wikipedia, other languages may lack sufficient textual support to build distributional models.

To address this problem, this paper investigates how different distributional semantic models built from corpora in different languages and with different sizes perform

in computing semantic relatedness similarity and relatedness tasks. Additionally, we analyse the role of machine translation approaches to support the construction of better distributional vectors and for computing semantic similarity and relatedness measures for other languages. In other words, in the case that there is not enough information to create a DSM for a particular language, this work aims at evaluating whether the benefit of corpora volume for English outperforms the error introduced by machine translation.

Given a pair of words and a human judgement score that represents the semantic relatedness of these two words, the evaluation method aims at indicating how close distributional models score to humans. Three widely used word-pairs datasets are employed in this work: Miller & Charles (MC)[14], Rubenstein & Goodenough (RG)[17] and WordSimilarity 353 (WS-353)[7].

In the proposed model the word-pairs datasets are translated into English as a reference language and the distributional vectors are defined over the target end model (Figure 1). Despite the simplicity of the proposed method based on machine translation, there is a high relevance for the distributional semantics user/practitioner due to its simplicity of use and the significant improvement in the results.



**Fig. 1.** Depiction of the experimental setup of the experiment.

This work presents a systematic study involving 11 languages and four distributional semantic models (DSMs), providing a comparative quantitative analysis of the performance of the distributional models and the impact of machine translation approaches for different models.

In summary, this paper answers the following research questions:

1. Does machine translation to English perform better than the word vectors in the original language (for which languages and for which distributional semantic models)?
2. Which DSMs and languages benefit more and less from the translation?

3. What is the quality of state-of-the-art machine translation approaches for word pairs (for each language)?

Moreover, this paper contributes with two resources which can be used by the community to evaluate multi-lingual semantic similarity and relatedness models: *(i)* a high quality manual translation of the three word-pairs datasets - Miller & Charles (MC)[14], Rubenstein & Goodenough (RG)[17] and WordSimilarity 353 (WS-353)[7] - for 10 languages and *(ii)* the 44 pre-computed distributional models (four distributional models for each one of the 11 languages) which can be accessed as a service<sup>3</sup>, together with the multi-lingual approaches mediated by machine translation.

This paper is organised as follows: Section 2 describes the related work, Section 3 describes the experimental setting; while Section 4 analyses the results and provides the comparative analysis from different models and languages, Finally, Section 5 provides the conclusion.

## 2 Related Work

Most of related work has concentrated on leveraging joint multilingual information to improve the performance of the models.

Faruqui & Dyer[6] use the distributional invariance across languages and propose a technique based on canonical correlation analysis (CCA) for merging multilingual evidence into vectors generated monolingually. They evaluate the resulting word representations on semantic similarity/relatedness evaluation tasks, showing the improvement of multi-lingual over the monolingual scenario.

Utt & Pado[20], develop methods that take advantage of the availability of annotated corpora in English using a translation-based approach to transport the word-link-word co-occurrences to support the creation of syntax-based DSMs.

Navigli & Ponzetto[15] propose an approach to compute semantic relatedness exploiting the joint contribution of different languages mediated by lexical and semantic knowledge bases. The proposed model uses a graph-based approach of joint multi-lingual disambiguated senses which outperforms the monolingual scenario and achieves competitive results for both resource-rich and resource-poor languages.

Zou et al.[21] describe an unsupervised semantic embedding (bilingual embedding) for words across two languages that represent semantic information of monolingual words, but also semantic relationships across different languages. The motivation of their works was based on the fact that it is hard to identify semantic similarities across languages, specially when co-occurrences words are rare in the training parallel text. Al-Rfou et al.[1] produced multilingual word embeddings for about 100 languages using Wikipedia as the reference corpora.

Comparatively, this work aims at providing a comparative analysis of existing state-of-the-art distributional semantic models for different languages as well as analyzing the impact of a machine translation over an English DSM.

---

<sup>3</sup> The service is available at <http://rebrand.ly/dinfra>.

### 3 Experimental Setup

The experimental setup consists of the instantiation of four distributional semantic models (Explicit Semantic Analysis (ESA)[9], Latent Semantic Analysis (LSA)[12], Word2Vec (W2V)[13] and Global Vectors (GloVe)[16]) in 11 different languages - English, German, French, Italian, Spanish, Portuguese, Dutch, Russian, Swedish, Arabic and Farsi.

The DSMs were generated from Wikipedia dumps (January 2015), which were pre-processed by lowercasing, stemming and removing stopwords. For LSA and ESA, the models were generated using the SSpace Package[11], while W2V and GloVe were generated using the code shared by the respective authors. For the experiment the vector dimensions for LSA, W2V and GloVe were set to 300 while ESA was defined with 1500 dimensions. The difference of size occurs because ESA is composed of sparse vectors. All models used in the generation process the default parameters defined in each implementation.

Each distributional model was evaluated for the task of computing semantic similarity and relatedness measures using three human-annotated gold standard datasets: Miller & Charles (MC)[14], Rubenstein & Goodenough (RG)[17] and WordSimilarity 353 (WS-353)[7]. As these word-pairs datasets were originally in English, except for those language available in previous works ([5,4]), the word pairs were translated and reviewed with the help of professional translators, skilled in data localisation tasks. The datasets are available at <http://rebrand.ly/multilingual-pairs>.

Two automatic machine translation approaches were evaluated: the Google Translate Service and the Microsoft Bing Translation Service. As Google Translate Service performed 16% better for overall word-pairs translations, this was set as the main machine translation model.

The DInfra platform [2] provided the DSMs used in the work. To support experimental reproducibility, both experimental data and software are available at <http://rebrand.ly/dinfra>.

## 4 Evaluation & Results

### 4.1 Spearman Correlation and Corpus Size

Table 1 shows the correlation between the average Spearman correlation values for each DSM and two indicators of corpus size: # of tokens and # of unique tokens.

ESA is consistently more robust (on average) than the other models in relation to the corpus size due the fact that ESA has larger context windows in opposition to the other distributional models. While ESA considers the whole document as its context window, the other models are restricted to five (LSA) and ten (Word2Vec and GloVe) words.

Another observation is that the evaluation of the WS-353 dataset is more dependent on the corpus size, which can be explained by the broader number of semantic relations expressed under the semantic relatedness umbrella.

Table 2 shows the size of each corpus in different languages regarding the number of unique tokens and the number of tokens.

Gold standard	MC		RG		WS353	
	unique tokens	tokens	unique tokens	tokens	unique tokens	tokens
ESA	0.39	0.48	0.67	<b>0.73</b>	0.33	0.39
LSA	<b>0.74</b>	<b>0.75</b>	<b>0.82</b>	0.68	<b>0.64</b>	0.66
W2V	0.43	0.58	0.71	0.72	0.57	<b>0.79</b>
Glove	0.34	0.51	0.51	0.61	0.59	0.63

**Table 1.** Correlation between corpus size and different models.

## 4.2 Word-pair Machine Translation Quality

The second step evaluates the accuracy of state-of-the-art machine translation approaches for word-pairs (Table 3). The accuracy of the translation for the WS-353 word pairs significantly outperforms the other datasets. This shows that the higher semantic distance between word pairs (semantic relatedness) has the benefit of increasing the contextual information during the machine translation process, subsequently improving the mutual disambiguation process.

lang	unique tokens	tokens
en	4.238	902.044
de	4.233	312.380
fr	1.749	247.492
ru	1.766	202.163
it	1.411	178.378
nl	2.021	105.224
pt	0.873	96.712
sv	1.730	82.376
es	0.829	76.587
ar	1.653	46.481
fa	0.925	32.557

**Table 2.** The sizes of the corpora in terms of the number of unique tokens and tokens (scale of  $10^6$ ).

dataset/lang	de	fr	ru	it	nl	pt	sv	es	ar	fa
MC	0.48	0.47	0.58	0.42	0.57	0.60	0.55	0.60	0.53	0.38
RG	0.45	0.65	0.53	0.41	0.59	0.51	0.58	0.59	0.43	0.36
WS353	<b>0.78</b>	<b>0.85</b>	<b>0.76</b>	<b>0.76</b>	<b>0.85</b>	<b>0.81</b>	<b>0.78</b>	<b>0.79</b>	<b>0.57</b>	<b>0.43</b>

**Table 3.** Translation accuracy.

For WS-353 the set of best-performing translations has an average accuracy of 80% (with maximum 85% and minimum 76%). This value dropped significantly for Arabic and Farsi (average 50%).

GS	Models	en	de	fr	ru	it	nl	pt	sv	es	ar	fa	Model AVG.	GS AVG.
MC	ESA	0.69	0.67	0.54	0.66	0.37	0.54	<b>0.67</b>	0.37	0.58	0.37	0.56	0.53	<b>0.56</b>
	LSA	0.79	<b>0.70</b>	0.55	0.63	0.58	0.55	0.41	<b>0.58</b>	0.66	<b>0.46</b>	0.45	0.56	
	W2V	<b>0.84</b>	<b>0.70</b>	0.55	0.64	<b>0.74</b>	<b>0.57</b>	0.37	0.40	<b>0.74</b>	0.38	<b>0.68</b>	<b>0.58</b>	
	Glove	0.69	0.64	<b>0.64</b>	<b>0.76</b>	0.51	0.55	0.62	0.40	0.65	0.38	0.45	0.56	
RG	ESA	0.80	0.68	0.45	0.63	0.50	0.58	0.51	0.50	0.59	0.36	0.57	0.54	0.53
	LSA	0.72	0.65	0.30	0.51	0.48	0.52	0.30	0.53	0.35	0.35	0.46	0.45	
	W2V	<b>0.85</b>	<b>0.78</b>	<b>0.57</b>	0.64	<b>0.69</b>	<b>0.63</b>	0.42	<b>0.57</b>	<b>0.64</b>	<b>0.36</b>	0.55	<b>0.58</b>	
	Glove	0.74	0.69	0.50	<b>0.70</b>	0.59	0.54	<b>0.52</b>	0.49	0.61	0.32	<b>0.59</b>	0.56	
WS353	ESA	0.50	0.39	0.32	<b>0.44</b>	0.34	0.53	0.44	0.43	0.37	0.26	0.37	0.39	0.41
	LSA	0.54	0.45	0.35	0.40	0.33	0.47	0.39	0.40	0.36	0.28	0.43	0.39	
	W2V	<b>0.69</b>	<b>0.54</b>	<b>0.50</b>	<b>0.53</b>	<b>0.50</b>	<b>0.58</b>	<b>0.53</b>	<b>0.45</b>	<b>0.53</b>	<b>0.44</b>	<b>0.53</b>	<b>0.51</b>	
	Glove	0.49	0.41	0.34	0.42	0.30	0.46	0.38	0.33	0.32	0.26	0.36	0.36	
	Lang AVG.	0.70	0.61	0.47	0.58	0.49	0.54	0.46	0.45	0.53	0.35	0.50	0.50	

**Table 4.** Spearman correlation for the language-specific models.

GS	Models	de	fr	ru	it	nl	pt	sv	es	ar	fa	Model AVG.	Diff.
MC	ESA-MT	0.55	0.53	0.42	0.38	0.45	0.38	0.48	0.39	0.31	0.58	0.45	-0.08 (-15.1%)
	LSA-MT	0.61	0.72	0.65	0.67	0.66	0.70	0.74	0.78	0.69	0.75	0.70	0.14 (25.0%)
	W2V-MT	<b>0.68</b>	<b>0.79</b>	<b>0.68</b>	<b>0.77</b>	<b>0.69</b>	<b>0.76</b>	<b>0.81</b>	<b>0.83</b>	<b>0.71</b>	0.74	<b>0.75</b>	<b>0.17 (29.3%)</b>
	GloVe-MT	0.45	0.78	0.67	0.64	0.63	0.56	0.61	0.82	0.69	<b>0.79</b>	0.66	0.10 (17.9%)
RG	ESA-MT	0.62	0.53	0.52	0.61	0.63	0.57	0.56	0.47	0.38	0.71	0.56	0.02 (3.7%)
	LSA-MT	0.63	0.62	0.59	0.74	0.67	0.64	0.67	0.62	0.55	0.70	0.64	<b>0.19 (42.2%)</b>
	W2V-MT	<b>0.69</b>	<b>0.79</b>	0.69	<b>0.78</b>	0.74	<b>0.75</b>	<b>0.71</b>	<b>0.73</b>	0.57	0.79	<b>0.72</b>	0.14 (24.1%)
	GloVe-MT	0.62	0.77	<b>0.71</b>	0.77	<b>0.78</b>	0.66	0.66	0.72	<b>0.65</b>	<b>0.80</b>	0.71	0.15 (26.8%)
WS353	ESA-MT	0.42	0.45	0.41	0.41	0.44	0.43	0.40	0.35	0.42	0.32	0.40	0.01 (2.6%)
	LSA-MT	0.51	0.51	0.47	0.48	0.51	0.39	0.51	0.44	0.37	0.43	0.46	<b>0.07 (17.9%)</b>
	W2V-MT	<b>0.62</b>	<b>0.59</b>	<b>0.57</b>	<b>0.57</b>	<b>0.63</b>	<b>0.51</b>	<b>0.59</b>	<b>0.55</b>	<b>0.50</b>	<b>0.52</b>	<b>0.57</b>	0.06 (11.8%)
	GloVe-MT	0.45	0.48	0.42	0.43	0.46	0.33	0.42	0.41	0.33	0.37	0.41	0.05 (13.9%)
	Lang AVG.	0.57	0.63	0.57	0.60	0.61	0.56	0.60	0.59	0.52	0.63	0.56	

**Table 5.** Spearman correlation for the machine translation models over the English corpora. *Diff.* represents the difference of machine translation score minus the language specific.

For MC and RG, the average translation accuracy for the semantic similarity pairs is 51.5%. This difference may be a result of a deficit of contextual information during the machine translation process. For these word-pairs datasets, the difference between best translation performers and lower performers (across languages) is smaller. Additionally, the final translation accuracy for all languages and all word-pairs datasets is 59%. French, Dutch and Spanish are the languages with best automatic translations.

### 4.3 Language-Specific DSMs

In the first part of the experiment, the Spearman correlations ( $\rho$ ) between the human assessments and the computation of the semantic similarity and relatedness for all DSMs instantiated for all languages were evaluated (Figure 1 (ii)). Table 4.2 shows the Spearman correlation for each DSM using language-specific corpora (without machine translation), for the three word-pairs datasets.

GS	Models	de	fr	ru	it	nl	pt	sv	es	ar	fa	Model AVG.	GS AVG.
MC	ESA	-0.18	-0.03	-0.36	0.03	-0.16	-0.44	0.31	-0.32	-0.16	0.03	-0.13	<b>0.41</b>
	LSA	-0.13	0.31	0.04	0.16	0.20	0.70	0.27	0.17	0.50	0.68	0.29	
	W2V	-0.02	0.43	0.07	0.05	0.21	<b>1.04</b>	1.00	0.13	<b>0.88</b>	0.09	0.39	
	GloVe	-0.31	0.22	-0.11	0.25	0.14	-0.10	0.51	<b>0.26</b>	0.85	0.75	0.25	
RG	ESA	-0.09	0.19	-0.18	0.21	0.08	0.11	0.12	-0.19	0.06	0.25	0.06	<b>0.41</b>
	LSA	-0.03	1.04	0.14	0.52	0.30	<b>1.15</b>	0.26	<b>0.77</b>	0.57	0.52	0.52	
	W2V	-0.11	0.39	0.08	0.14	0.18	0.76	0.23	0.14	0.59	0.44	0.28	
	GloVe	-0.11	0.55	0.01	0.31	<b>0.43</b>	0.28	0.35	0.17	<b>1.04</b>	0.36	0.34	
WS353	ESA	0.08	0.40	-0.07	0.18	-0.18	-0.02	-0.07	-0.07	0.60	-0.13	0.07	0.36
	LSA	0.12	0.43	0.19	0.45	0.09	-0.01	0.27	0.21	0.34	0.01	0.21	
	W2V	0.14	0.19	0.09	0.14	0.08	-0.04	0.33	0.04	0.12	0.00	0.11	
	GloVe	0.10	0.41	0.00	0.41	0.00	-0.14	0.28	0.30	0.28	0.04	0.17	
	Lang AVG.	0.06	0.52	0.13	0.36	0.23	0.29	0.70	0.22	0.59	0.82		

**Table 6.** Difference between the language-specific and the machine translation model.

The comparative language-specific analysis indicates that English is the best-performing language (0.70), followed by German (0.61). The lowest Spearman correlation was observed in Arabic (0.35). From the tested DSMs, W2V is consistently the best-performing DSM (0.56). The language-specific DSMs achieved higher correlations for MC and RG (0.56 and 0.53, respectively), in comparison to 0.41 for WS-353.

The results for the language-specific DSMs were contrasted to the machine translation (MT) approach, according to the diagram depicted in Figure 1 (i). The Spearman correlation for the MT-mediated approach are shown in Table 5.

#### 4.4 Machine Translation based Semantic Relatedness

Using the MT models, W2V is consistently the best performing DSM (average 0.68), while ESA is consistently the worst performing model (0.47). We can interpret this result by stating that the benefit of using machine translation for ESA does not introduce significant performance improvements in comparison to the language-specific baselines.

The best performing languages are French and Farsi ( $\rho = 0.63$ ). The Spearman correlation variance across languages in the MT models is low, as the impact of the use of the English corpus on the DSM model has a higher positive impact on the results in comparison to the variation of the quality of the machine translation. The results for all languages achieve very similar correlation values.

The impact of the MT model can be better interpreted by examining the difference between the machine translation and the domain-specific models (depicted in Table 6). LSA accounts for the largest average percent improvement (28.4%) using the MT model, while ESA accounts for the lowest value (-2.9%). As previously noticed, this can be explained by the sensitivity of these models to the corpus size due to the dimensional reduction strategy (LSA) or the broader context window (ESA). The remaining models accounted for substantial improvements (W2V = 21.7%, GloVe = 19.5%).

Arabic and French achieved the highest percent gains (47% and 38%, respectively), while German accounts for worst results (-4%). These numbers are consistent with the corpus size. For German, the result shows that the corpus volume of the German Wikipedia crossed a threshold size (34% of the English corpus) above which improvements for computing semantic similarity for the target word-pairs dataset might be marginally relevant, while the translation error accounts negatively in the final result.

The average improvement for the MT over the language specific model for each word-pairs dataset is consistently significant: MC = 20%, RG = 30% and WS353 = 14%.

#### 4.5 Summary

Below, the interpretation of the results are summarised as the core research questions which we aim to answer with this paper:

**Question 1:** Does machine translation to English perform better than the word vectors in the original language (for which languages and for which distributional semantic models)?

Machine translation to English consistently performs better for all languages, with the exception of German, which presents equivalent results for the language-specific models. The MT approach provides an average improvement of 16.7% over language-specific distributional semantic models.

**Question 2:** Which DSMs or MT-DSMs work best for the set of analysed languages? W2V-MT consistently performs as the best model for all word-pairs datasets and languages, except German, in which the difference between MT-W2V and language-specific W2V is not significant.

**Question 3:** What is the quality of state-of-the-art machine translation approaches for word-pairs?

The average translation accuracy for all languages and all word-pairs datasets is 59%. Translation quality varies according to the nature of the word-pair (better translations are provided for word pairs which are semantically related compared to semantically similar word pairs), reaching a maximum of 85% and a minimum of 36% across different languages.

For the distributional semantics user/practitioner, as a general practice, we recommend using W2V built over an English corpus, supported by machine translation. Additionally, the accuracy of state-of-the-art machine translation approaches work better for translating semantically related word pairs (in contrast to semantically similar word pairs).

## 5 Conclusion

This work provides a comparative analysis of the performance of four state-of-the-art distributional semantic models over 11 languages, contrasting the native language-specific models with the use of machine translation over English-based DSMs. The experimental results show that there is a significant improvement (average of 16.7% for the Spearman correlation) by using off-the-shelf machine translation approaches



and that the benefit of using a more informative (English) corpus outweighs the possible errors introduced by the machine translation approach. The average accuracy of the machine translation approach is 59%. Moreover, for all languages, W2V showed consistently better results, while ESA showed to be more robust concerning lower corpora sizes. For all languages, the combination of machine translation over the W2V English distributional model provided the best results consistently (average Spearman correlation of 0.68).

Future work will focus on the analysis and translation of two other word-pairs datasets: SimLex-999[10] and MEN-3000[3].

## Acknowledgments

This publication has emanated from research supported by the National Council for Scientific and Technological Development, Brazil (CNPq) and by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

## References

1. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual nlp. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. pp. 183–192. Association for Computational Linguistics, Sofia, Bulgaria (August 2013), <http://www.aclweb.org/anthology/W13-3520>
2. Barzegar, S., Sales, J.E., Freitas, A., Handschuh, S., Davis, B.: Dinfra: A one stop shop for computing multilingual semantic relatedness. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1027–1028. SIGIR '15, ACM, New York, NY, USA (2015), <http://doi.acm.org/10.1145/2766462.2767870>
3. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. *J. Artif. Int. Res.* 49(1), 1–47 (Jan 2014), <http://dl.acm.org/citation.cfm?id=2655713.2655714>
4. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: A framework for the construction of monolingual and cross-lingual word similarity datasets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP). pp. 1–7. Citeseer (2015)
5. Faruqi, M., Dyer, C.: Community evaluation and exchange of word vectors at wordvectors.org (2014)
6. Faruqi, M., Dyer, C.: 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. Association for Computational Linguistics, Gothenburg, Sweden (April 2014), <http://www.aclweb.org/anthology/E14-1049>
7. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proceedings of the 10th international conference on World Wide Web. pp. 406–414. ACM (2001)
8. Freitas, A.: Schema-agnostic queries over large-schema databases: a distributional semantics approach. Ph.D. thesis, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway (2015)

9. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611. IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007), <http://dl.acm.org/citation.cfm?id=1625275.1625535>
10. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* (2015)
11. Jurgens, D., Stevens, K.: The s-space package: An open source package for word space models. In: Proceedings of the ACL 2010 System Demonstrations. pp. 30–35. ACLDemos '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), <http://dl.acm.org/citation.cfm?id=1858933.1858939>
12. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. *Discourse processes* 25(2-3), 259–284 (1998)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop Papers (2013)
14. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. *Language and cognitive processes* 6(1), 1–28 (1991)
15. Navigli, R., Ponzetto, S.P.: Babelrelate! a joint multilingual approach to computing semantic relatedness. In: AAAI Conference on Artificial Intelligence (2012)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)* 12, 1532–1543 (2014)
17. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. *Communications of the ACM* 8(10), 627–633 (1965)
18. Sales, J.E., Freitas, A., Davis, B., Handschuh, S.: A compositional-distributional semantic model for searching complex entity categories. In: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (\*SEM). pp. 199–208 (2016)
19. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.* 37(1), 141–188 (Jan 2010), <http://dl.acm.org/citation.cfm?id=1861751.1861756>
20. Utt, J., Pad, S.: Crosslingual and multilingual construction of syntax-based vector space models. *Transactions of the Association of Computational Linguistics* 2, 245–258 (2014)
21. Zou, W.Y., Socher, R., Cer, D.M., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: EMNLP. pp. 1393–1398 (2013)