Judgment Language Matters: Multilingual Vector Space Models for Judgment Language Aware Lexical Semantics

Ira Leviant
IE&M faculty
Technion - IIT
irakr@campus.technion.ac.il

Roi Reichart
IE&M faculty
Technion - IIT
roiri@ie.technion.ac.il

Abstract

It is a common practice in the vector space model (VSM) literature to evaluate the models' ability to predict human similarity scores for a set of word pairs. However, existing evaluation sets, even those used to evaluate multilingual VSMs, consist of English words only. In this paper we show that this practice may have significant undesired effects on VSM evaluation. By translating the popular wordsim353 evaluation set to three languages and training state-of-the-art VSMs on corpora of the corresponding languages as well as on English, we show that: (a) The *judgment language* in which word pairs are presented to human evaluators, all fluent in that language, has a substantial impact on their produced similarity scores; (b) Given the judgment language of an evaluation set, this judgment language is a good choice for the VSM training corpus language; and (c) Monolingual VSMs can be combined into multilingual VSMs that can predict human similarity scores for a variety of judgment languages better than any monolingual model. Our results highlight the impact of the judgment language on the human generated similarity scores and point on the importance of constructing judgment language aware VSMs.[1]

1 Introduction

In the last two decades, research in computational lexical semantics has emphasized the development of *Vector Space Models (VSMs)* for word meaning representation. Most models in this paradigm, from the ones that apply simple word co-location techniques (see a survey in (Turney et al., 2010)) to state-of-the-art Neural Network (NN) models (e.g. (Mikolov et al., 2013a)), are based on the distributional semantics hypothesis (Harris, 1954), stating that words that occur in similar contexts are likely to have similar meanings.

VSMs produce a vector representation for each word in the lexicon. A common evaluation practice for such models is to compute a similarity score for each member of a word pair set by applying a similarity function to the vectors of the words participating in the pair. The correlation between the model's scores and the scores generated by human evaluators is then computed. Henceforth we refer to the language of the word pairs presented to the human evaluators as the *judgment language* and to the language of the corpus used for a VSM training as its *training language*.

Recent research on multilingual approaches to VSMs aim to exploit the impact of multilingual training. The resulted models were evaluated either in *intrinsic evaluation* against human scores produced for English word pairs only (Faruqui and Dyer, 2014) or in *extrinsic evaluation* on multilingual text mining tasks (Klementiev et al., 2012; Lauly et al., 2013; Khapra et al., 2013; Hermann and Blunsom, 2014b; Hermann and Blunsom, 2014a; Kočiský et al., 2014; Lauly et al., 2014; Al-Rfou et al., 2014). While works which facilitate extrinsic evaluation do recognize the connection between the VSM training language and the application's test set language, to the best of our knowledge no previous work explored the impact of the judgment language in intrinsic evaluation.

In this paper we explore two open issues regarding this impact: (a) the effect of the judgment language on the similarity scores human evaluators generate for word pairs; and (b) the effect of the VSM training language on the model's capability to predict human generated similarity scores for

[1] The data sets (translated word pairs with their human scores, and translated annotation guidelines) produced in this work will be made publicly available upon paper acceptance.

word pairs in different judgment languages.

To address these issues we translate the popular wordsim353 dataset (Finkelstein et al., 2001), consisting of 353 English word pairs annotated for similarity,[2] to three languages from different branches of the Indo-European language family: German (Germanic), Italian (Romance) and Russian (Slavic). Then, we employ the CrowdFlower crowdsourcing service [3] to collect similarity scores for the translated sets from human evaluators, fluent in the judgment languages (Section 3).

In Section 5 we explore the effect of the judgment language on the human generated similarity scores. Our hypothesis is that due to a variety of factors – linguistic, cultural and others – the judgment language should affect the produced scores. Indeed, our results suggest that inter evaluator agreement is significantly higher within a judgment language than it is when the judgment languages differ. This suggests that the concept of word similarity is judgment language dependent.

We then (Section 6) investigate the connection between the VSM training language and the human judgment language. We do this by training VSMs – co-location based as well as NN based – on monolingual comparable corpora from our four judgment languages (Section 4) and comparing their predicted similarity scores with the human scores produced for the different judgment languages. Our results suggest that, given an evaluation set judgment language, this language is a good choice for the VSM training corpus language, although not necessarily the best choice.

Finally, to better understand the dependence between training and judgment languages, in Section 7 we explore a number of methods for combining monolingual VSMs. Our results show that, for most training language pairs, a method as simple as a linear combination of the similarity scores of the participating monolingual models, improves the predictive power with respect to the corresponding judgment languages. Interestingly, more sophisticated techniques, based on the Canonical Correlation Analysis method (CCA, (Hardoon et al., 2004)), fail to achieve such improvements.

2 Previous Work

Vector Space Modeling and Its Evaluation. Vector space modeling has become a key tool in computational lexical semantics. Earlier works (see (Turney et al., 2010)) directly exploited the distributional hypothesis (Harris, 1954), designing representations based on word co-location statistics that were then potentially post-processed using techniques such as Positive Pointwise Mutual Information (PPMI) and dimensionality reduction. Recently, much of the focus in this field has drifted to the development of Neural Networks (NNs) for word representation learning (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011; Huang et al., 2012; Mikolov et al., 2013a; Mikolov et al., 2013c; Levy and Goldberg, 2014; Pennington et al., 2014, inter alia).

Vector space models have been evaluated in two main forms. In intrinsic evaluation, the correlation between the similarity scores predicted by the model for word pairs in an evaluation set are compared against the human generated scores for the same pair set (Silberer and Lapata, 2012; Huang et al., 2012; Baroni et al., 2014; Levy and Goldberg, 2014; Pennington et al., 2014, inter alia). The similarity score assigned by the model to a given word pair is most often the cosine similarity between the vectors learned for the pair's members and the comparison between the model and the human scores is done through a correlation measure. In extrinsic evaluation, the potential of the learned vectors to enhance NLP applications is tested (Collobert and Weston, 2008; Collobert et al., 2011; Pennington et al., 2014, inter alia).

Several data sets consisting of word pairs scored by humans for semantic relationships (mostly similarity) are in use for intrinsic evaluation. Among these are RG-65 (Rubenstein and Goodenough, 1965), MC-30 (Miller and Charles, 1991), wordsim353 ((Finkelstein et al., 2001), WS3535 henceforth), YP-130 (Yang and Powers, 2006), MTurk-287 (Radinsky et al., 2011), MTurk-771 (Halawi et al., 2012), Stanford-rare-word (Luong et al., 2013), MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2014b) and Verb-143 (Baker et al., 2014). In all these datasets, however, the judgment language is English.

Multilingual Vector Space Modeling. Recently, there has been a growing interest in multilingual vector space modeling (Klementiev et al., 2012; Lauly et al., 2013; Khapra et al., 2013; Hermann and Blunsom, 2014b; Hermann and Blunsom, 2014a; Kočiský et al., 2014; Lauly et al.,

[2] In this paper we do not distinguish between *similarity* and *relatedness/association*. See a discussion in (Hill et al., 2014b).

[3] http://www.crowdflower.com/

2014; Al-Rfou et al., 2014; Faruqui and Dyer, 2014). These works train VSMs on multilingual data, either parallel or not, or combine VSMs trained on monolingual data. Evaluation is either intrinsic, using English word pairs, or extrinsic with tasks such as multilingual text mining that reflect the quality of the learned vectors for application models applied to multiple languages.

In this paper we address an aspect of vector space modeling that, to the best of our knowledge, has not been addressed before: the effect of the human judgment language. We explore the impact of this language on the human generated similarity scores and the ability of monolingual and multilingual VSMs to predict human similarity scores produced for a variety of judgment languages.

3 Multilingual Human Judgment Data

In this section we describe the data collection process. Our working data set is WS353 (Finkelstein et al., 2001), selected due to its popularity for intrinsic VSM evaluation. 153 of the 353 word pairs in this set were scored by 13 human subjects and the remaining 200 by 16, with all 29 evaluators possessing a near-native command of English. Evaluators were guided to estimate the relatedness of the words in each pair on a scale of 0 (totally unrelated) to 10 (highly related or identical).[4]

We start with a description of the translation process in which we produced German, Italian and Russian versions of the data set and the scoring guidelines. We then describe the collection of human scores for the translated data sets.

3.1 WS353 Translation

We started by translating the WS353 scoring guidelines to the target languages (German, Italian and Russian). For each language the translation was done by two native speakers, and disagreements were solved through a discussion mediated by an experiment manager. An external evaluator, fluent in both the target language and in English then verified the translation quality.

The word pair translation process was more complicated. We followed the same protocol outlined above and further set a number of rules that guided our translators in challenging cases. Below we discuss the different types of translation ambiguities addressed in our guidelines.

Gender. In some cases English does not make gender distinctions that some of the other languages do. For example, the English word *cat* refers to both the female and the male cat while in Russian and Italian each gender has its own word (e.g. *gatto* and *gatta* in Italian). In such cases, if the other word in the English pair has a clear gender interpretation we followed this gender in the translation of both words, otherwise we chose one of the genders randomly and kept it fixed across the target languages.

Word Senses. It is common that some words in a given language have a sense set that is not conveyed by any of the words of another given language. For example, the English word *plane*, from the WS353 pair *(car,plane)*, has both the *airplane* and the *geometric plane* senses. However, to the best of our translators' knowledge, no German, Italian or Russian word has these two senses.

We assume that when the authors of WS3353 paired two words, they referred to their closest senses. Therefore, like for gender, we used the other word in the pair for sense disambiguation. In our example, *plane* is translated to the target language word which has the *airplane* meaning (e.g. *Flugzeug* in German, *aeroplano* in Italian), since this sense is closer to the meaning of *car*.

In cases where the other word in the pair does not clearly disambiguate the sense of its polysemous counterpart, we randomly chose one of the latter word's senses, and kept it fixed across the target languages. Consider, for example, the pair *(life,stock)*. *Stock* has two senses [5] - one corresponds to the *the supply of goods available for sale in a store* and the other to *a share of the value of a company which can be bought, sold, or traded as an investment*. Since it was not clear to our translators how the word *life* can facilitate sense selection, we randomly chose the latter sense and used it in the translation to all target languages.

Sense disambiguation is done on a POS basis as well. For example, in the pair *(attempt,peace)*, *attempt* can be a verb or a noun and none of these senses is necessarily closer to the meaning of *peace*. Reasoning that words with the same POS tend to have a closer meaning, in such cases we used the interpretation of the polysemous word which has the same POS as the other word in the pair. That is, in the current example *attempt* was

[4] The data set and annotation guidelines are available at http://www.cs.technion.ac.il/∼gabr/resources/data/wordsim353/wordsim353.html

[5] According to http://www.merriam-webster.com/dictionary/stock

assigned its noun sense, as *peace* is a noun.

Naturally, the same problem can happen on the opposite direction – the target language translation of a given English word may have multiple senses, some of which are not expressed by the English word. We guided our annotators to avoid such translations whenever possible, although in a few cases that was impossible.

Pair Exclusion. Three pairs were excluded from our sets due to translation difficulties. The pairs *(noon,midday)* and *(coast, shore)* were excluded because, according to our translators, none of the target languages includes two different words that convey the meaning of either set. The pair *(football,soccer)* was also excluded since it reflects a cultural distinction that is not made in the target languages. The resulted data sets in all four languages therefore consist of 350 word pairs.

Inter Annotator Agreement. Naturally, our ambiguity resolution guidelines were not able to promise full agreement between the translators. In practice, the disagreement rates between our two translators for the 700 words in the 350 translated pairs were: Russian (85 words, 12.1%), Italian (57 words, 8.1%) and German (113 words, 16.1%). To resolve these disagreements, for each language we asked one of the translators to decide on the final translation by choosing the translation which is more similar in meaning to the other word in the pair and, if this is not possible, the word which seems to her more common in the target language.

3.2 Word Pair Scoring

We now turn to the description of the word pair scoring process. We divided the 350 word pairs to 7 non-overlapping batches of 50 pairs each and employed the crowdflower crowdsourcing service to recruit fluent speakers of each target language to score each batch.[6] For all three languages evaluators were presented with the scoring guidelines translated to their judgment language and were asked to score the pairs on a 0-10 scale.

We verified the quality of our evaluators through a three step process. First, for each of the target languages we only recruited evaluators who were located at a country where the target language is the mother tongue of the majority of the population (i.e. Germany, Italy or Russia). Second, in order to make sure that our evaluators understand the task properly (which further implies they understand the annotation guidelines that were presented to them in the judgment language only), we generated 7 tests for each language, each consisting of two word pairs that do not appear in the data set. Each of these pairs consisted of words that were either very similar or very dissimilar. Before scoring the data, each evaluator was presented with a randomly sampled test in its language and was asked to score the respective two word pairs. Every evaluator that assigned a similar pair with a score lower than 7 or a non-similar pair with a score higher than 3 was excluded from the experiment.

Finally, we ran an outlier detection procedure in order to exclude evaluators whose scores were substantially different from those of the other evaluators of their batch. [7] For each evaluator we computed the distance of its average word pair score from the average of the other evaluators and normalized by the standard deviation of the latter set. Evaluators whose statistic was above a predefined threshold [8] were excluded from the final data set. We performed this procedure periodically and once a batch had 13 annotators that passed the test we stopped collecting scores for that batch.[9]

4 Vector Space Models

We describe the monolingual VSMs we employ, their training data and evaluation protocol.

4.1 Models

Bag of Words (BOW). We constructed a VSM following the optimal performance guidelines given in (Kiela and Clark, 2014). After extracting the k most frequent words in the training corpus, we generated a matrix of co-occurrence counts with a row for each of the words in any of the pairs

[6] Crowdflower allows up to 100 pairs per job. Each of the crowdflower workers was paid 120 American Cents per a 50 pair batch. It took four days to complete the scoring process for each language.

[7] Some works that apply crowdsourcing compare some of the collected annotation to a pre-prepared gold standard. We consider our outlier detection process an alternative as it keeps only those annotators that enjoy a high agreement rate with each other.

[8] The threshold was set to 1.4, reasoning that if the word pair scores were sampled from a normal distribution with the empirical mean and variance, then $\sim$80% of the evaluators would be below the threshold.

[9] In order to treat all languages equally, we ran the same procedure for English as well. Consequently, when we split the data sets to batches we did not cross the boundaries of the original 153 and 200 word pair subsets of WS353. In one of the English batches we had to collect scores from one additional evaluator using crowdflower as only 12 of its evaluators passed the outlier test.

in our evaluation set, and a column for each of the k most frequent words.[10] Co-occurrence was counted within a window size C,[11] without crossing sentence boundaries. The entries of the matrix were then normalized to PPMI values (Church and Hanks, 1990). The resulting matrix's rows constitute the vector representations of the words.

word2vec. The Mikolov et al.'s Neural Network (NN) model for word embedding induction, whose vectors achieve state-of-the-art performance on several semantic tasks (Mikolov et al., 2013a; Mikolov et al., 2013b).[12] The model aims to predict the past and future context given a word by learning word representations that maximize the following objective:

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

Where T is the number of training tokens, and c a window size parameter. The objective respects sentence boundaries, conditioning only words from the same sentence on each other. [13]

Three parameters were tuned for this model. D: the vector dimensionality, F: a frequency cutoff for words to be included in the objective, and c: the window size. We followed Radim Rehurek's word2vec tutorial [14] and set $c = 5$, $D = 400$ and $F = 1$ for all training languages.

4.2 Training and Word Pair Scoring

We trained our VSMs on the Wikipedia corpora released by (Al-Rfou et al., 2014) .[15] This is a set of multilingual comparable corpora, as Wikipedia entries covering the same topic have similar content even if written in different languages. This allows us to focus on the effect of the training and judgment language change, while keeping the training topics fixed across languages.[16]

Before running the models we cleaned the corpora, removing any string that is not comprised of alphabetic characters only as well as stopwords, [17] and stemming the remaining words using an NLTK stemmer.[18] The size of the resulted corpora, which were used for VSM model training, is as follows (left number for word types, right number for word tokens): English (3.984 M, 1.4 G), German (5.099 M, 484.5 M), Italian (1.65 M, 281.6 M), Russian (2.809 M, 230 M).

The score assigned to each word pair by a model is the cosine similarity between the vectors the model induces for the participating words. For each training/judgment language pair we then compute the Spearman correlation coefficient score (ρ) between the ranking derived from a model's word pair scores and the ranking derived from the human scores.[19]

5 The Judgment Language Effect

Our first question is: what is the effect of the human judgment language on the produced word pair scores. To provide a quantitative answer, we run the following protocol, both within and across judgment languages. For each of the seven 50 word pair batches, we generated all possible K-size subsets of the 13 scores we collected for that batch, each K-size subset defining a unique partition of the batch evaluators (We set K to 6). Then, for the within language evaluation we calculated the correlation between the word pair scores (averaged across the subsets' evaluators) of the two subsets induced by each K-size subset selection. For the cross language evaluation, in turn, we calculated the correlation between the average word pair scores of each K-size subset of language 1 with its corresponding subset of language 2. The resulted ρ scores were averaged to get a final score for each language (in the within-language case) and language pair (in the cross-language case).[20]

Table 1 presents our results. The correlations within a judgment language are clearly better compared to their cross-language counterparts, with mean values at the $[0.8708 - 0.8961]$ range for the former compared to $[0.7196 - 0.7521]$ for the latter. Further, the standard deviation values are $[0.0326 - 0.0550]$ for within-language compared

[10] We experimented with $k \in \{1000, 2000, \ldots, 10000\}$ and set $k = 10000$ for all training languages.

[11] We experimented with $C \in \{2, 3, \ldots, 8\}$ and set $C = 2$ for all training languages.

[12] We used the model implementation in: http://word2vec.googlecode.com/svn/trunk/word2vec.c

[13] We excluded this detail from the objective for brevity.

[14] http://radimrehurek.com/2014/02/word2vec-tutorial/

[15] https://sites.google.com/site/rmyeid/projects/polyglot

[16] Despite extensive efforts, we were not able to find additional sets of comparable corpora with sufficient size for the training of modern VSMs. We therefore perform our experiments with this set only.

[17] According to the NLTK list, http://www.nltk.org/

[18] http://www.nltk.org/howto/stem.html

[19] Result patterns are very similar when considering the Pearson and Kandall Tao scores. We hence keep our presentation concise and report only the Spearman scores.

[20] We have 1716 K-size subsets for each batch and a total of 1716*7 correlations for each scenario.

	English		German		Italian		Russian	
	mean	std	mean	std	mean	std	mean	std
English	0.8961	0.0326	0.7521	0.1050	0.7388	0.0917	0.7394	0.1104
German	—	—	0.8640	0.0550	0.7004	0.1050	0.7199	0.0763
Italian	—	—	—	—	0.8708	0.0549	0.7196	0.1203
Russian	—	—	—	—	—	—	0.8797	0.0329

Table 1: Average Spearman ρ correlation coefficient between human similarity judgments in the within and the cross language setups. The (L_1, L_2) table entry (which is further divided into *mean* and *std* columns) corresponds to the comparison of evaluators with judgment language L_1 to evaluators with judgment language L_2.

$T \mid J$	English	German	Italian	Russian
English	0.5998	0.5229	0.4884	0.4959
German	0.3869	0.4140	0.3602	0.4075
Italian	0.4854	0.4097	0.4512	0.4270
Russian	0.4025	0.3772	0.3599	0.4264

(a) BOW

$T \mid J$	English	German	Italian	Russian
English	0.7238	0.6182	0.6143	0.5854
German	0.5541	0.5945	0.5037	0.5529
Italian	0.6009	0.4821	0.5685	0.5034
Russian	0.5922	0.5336	0.4964	0.6067

(b) word2vec

Table 2: Spearman ρ correlation coefficient between human similarity scores and the similarity scores of the VSMs. The (T, J) entry in each matrix is the ρ value for a VSM trained on a language T with the human similarity scores produced for judgment language J.

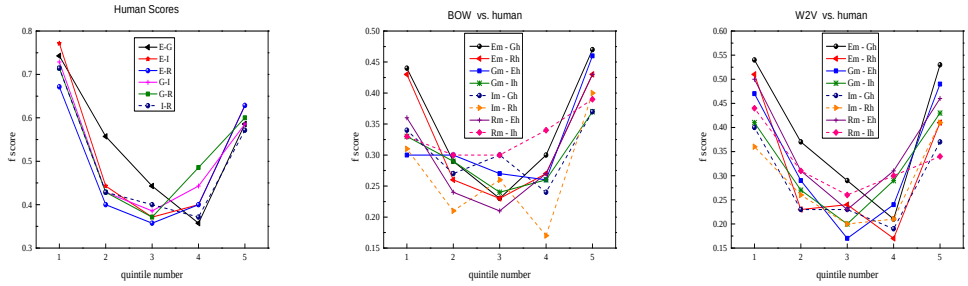

Figure 1: Relative F-score of the word pair lists in corresponding quintiles of (from left to right): (a) human rankings with different judgment languages; (b) BOW ranking with training language l_1 vs. human ranking with judgment language l_2; and (c) word2vec ranking with training language l_1 vs. human ranking with judgment language l_2. Languages are denoted with their first letter, m stands for model and h for human.

to $[0.0763 - 0.1203]$ for cross-language, reflecting the weaker dependence of the human judgment in the within-language setup on the involved word-pairs and human evaluators. [21] These results suggest that the concept of word similarity may be judgment language dependent.

To further investigate the judgment language effect, for each judgment language we assign each word pair with the average of the scores assigned to it by the human annotators and rank the word pairs accordingly. Then, for each pair of judgment languages we compute the relative F-score between corresponding quintiles in the rankings.

The leftmost graph of Figure 1 reveals that the overlap between quintiles is substantially stronger for the top and bottom quintiles (top and bottom 20% of the word pairs according to each of the rankings) compared to quintiles 2-4. This analysis suggests that the judgment language is most influential for word pairs that are of medium similarity across judgment languages.

Our next natural question is how the relations between the VSM training language and the human judgment language affect the correlation between the model and the human scores.

6 The VSM Training Language Effect

Table 2 presents the Spearman ρ correlation coefficient between human similarity scores and the similarity scores of (a) co-occurrence based BOW VSMs; and (b) NN-based word2vec VSMs.

The table reveals that for both model types, the training language that leads to similarity scores that best correlate with the human scores in each of the judgment languages (except from Russian for the word2vec model) is English. In five of six cases, using English as the training language leads to better performance even compared to when the

[21] We further ran the Welch's t-test for each set of correlations computed for an individual language with each set of correlations computed for a pair of languages. In all 24 cases the null hypothesis stating that the two sets have an equal mean was rejected with $Pvalue < 0.001$.

$L_1 \mid L_2$	English	German	Italian	Russian
English	—	-0.0090	0.0223	-0.0041
German	0.1136	—	0.0517	0.0354
Italian	0.0653	0.0281	—	0.0243
Russian	0.0883	0.0464	0.0547	—

(a) BOW

$L_1 \mid L_2$	English	German	Italian	Russian
English	—	0.0090	0.0335	0.0219
German	0.0722	—	0.0330	0.0352
Italian	0.0940	0.0552	—	0.0372
Russian	0.0381	0.0375	0.0355	—

(b) word2vec

Table 3: **Improvements in Spearman ρ correlation coefficient** achieved by the similarity scores resulted from a linear interpolation of the similarity scores of pairs of monolingual models. The $(L_1, L2)$ entry of each table is **the improvement** in correlation for judgment language L_1 achieved by the interpolation of the scores produced by monolingual models trained on languages L_1 and L_2 over a VSM trained on language L_1. For the BOW models, where one of the involved training languages is English, the English score was weighted $\lambda = 0.75$ (and the other language's score was weighted 0.25). When Italian (but not English) is one of the training languages, its weight is $\lambda = 0.67$. When only German and Russian are used for training their scores are assigned the same weight. For the word2vec models, where one of the involved training languages is English, the English score was weighted $\lambda = 0.67$ (and the other language's score was weighted 0.33). Otherwise, both involved languages are weighted $\lambda = 0.5$. See discussion in text.

$L_1 \mid L_2$	English	German	Italian	Russian
English	—	-0.0515	-0.0277	-0.0528
German	-0.0544	—	-0.0547	-0.0559
Italian	-0.0202	-0.0699	—	-0.0070
Russian	-0.1032	-0.1169	-0.0557	—

(a) BOW

$L_1 \mid L_2$	English	German	Italian	Russian
English	—	0.3372	0.3372	0.3372
German	0.2841	—	0.2947	0.2947
Italian	0.2592	0.2349	—	0.2246
Russian	0.2314	0.2071	0.2027	—

(b) word2vec

Table 4: **Spearman ρ correlation coefficient** of the similarity scores resulted from a CCA combination of monolingual models, with corresponding human judgment scores. The $(L_1, L2)$ entry of each table is the correlation of (1) the similarity scores resulted from CCA combination of the word embeddings produced by monolingual models trained on languages L_1 and L_2 with (2) the human similarity scores produced for judgment language L_1.

judgment language itself is used to train the VSM. Except from English, for both models the best choice of training language for a given judgment language is the judgment language itself.

As the English and German corpora are substantially larger than their Russian and Italian counterparts (Section 4) we validated our results by evaluating both VSMs when trained on the English or the German training corpus after these corpora were cut to the size of the Russian or the Italian corpus. The results we got were very similar.[22]

Our results are generally in line with the initial hypothesis: the training/judgment language combination has a noticeable effect on the performance of the VSM. Moreover, for all the languages in our experiments, the judgment language is a good choice for the VSM training language, being second only to English (in the cases English is not already the judgment language). As the training corpus size has been ruled out as a possible explanation to the dominance of English as a training language, we will have to search for further explanations in future work.[23]

To further investigate the mutual impact of the training and judgment languages, we replicated the quintile analysis of Section 5, this time comparing the rankings of a model trained with language l_1 to the human scores obtained with judgment language l_2. Results are presented in the two rightmost graphs of Figure 1.[24]

Interestingly, like in the respective analysis for judgment language pairs, the impact of language transfer is most prominent for word pairs that are considered of medium similarity - both by humans across judgment languages, and by models across training languages. Note, however, that in this analysis, the agreement between model and human scores on the top and bottom quintiles is weaker compared to the cross judgment language human agreement on the corresponding quintiles we explored in Section 5.

We next investigate whether the expressive power of two training languages is additive.

[22] up to 0.02 difference in the Spearman, Pearson and Kandall Tau scores.

[23] The dominance of English may stem from the fact that all our evaluation sets are translations of a set originally authored in English. We will explore this hypothesis in the future by repeating our experiments using evaluation sets that are originally authored in the other judgment languages and are then translated. As carefully evaluating the effect of the original evaluation set language requires additional extensive experimentation, we consider it beyond the scope of this paper.

[24] For brevity, we present only the curves for 8 out of 16 possible train/judgment language combinations, the patterns for the other combinations are very similar.

7 VSM Training Language Combination

We explored two simple methods for the combination of VSMs trained on corpora of different languages, $l1$ and $l2$. In the first method, *linear interpolation*, we combine the similarity scores produced by two VSMs for a word pair (w_i, w_j) using the following linear equation (for $\lambda \in [0, 1]$) [25]:

$$Score(w_i, w_j) = \lambda \cdot sc_{l1}(w_i, w_j) + (1 - \lambda) \cdot sc_{l2}(w_i, w_j)$$

Where $sc_{li}(w_i, w_j)$ is the score produced by the model trained on the li language. In the second method we followed the CCA protocol of (Silberer and Lapata, 2012), originally proposed for the combination of word embeddings derived from different modalities. For each pair of languages, (l_1, l_2), we calculated a pair of projection matrices to the shared subspace through the CCA method (Hardoon et al., 2004), using the vectors induced by monolingual models trained on an l_1 and an l_2 corpora. We then constructed a multilingual vector representation for each word by concatenating the l_1 and l_2 projected representations. [26]

Tables 3 and 4 present our results. The numbers clearly show that linear interpolation is an effective method of combining two monolingual models, leading to improvements with both the BOW and the word2vec models. This result is further demonstrated in table 5 that presents the improvement of combined word2vec models over monolingual word2vec models that were trained on the same language that was used for the human judgment. The table shows (top four lines) both the number of judgment languages for which improvement is achieved (out of 4) and the average improvement in Spearman ρ. Comparison to the effect of using each of the monolingual models (bottom line, out of 3 languages reveals the impact of this combination method [27] [28].

[25] We experimented with $\lambda \in \{0.25, 0.33, 0.5, 0.67, 0.75\}$ and got improvements for most combinations of training language pairs, judgment languages and λs. We selected λs that prefer the English (for both BOW and word2vec) and the Italian (for BOW only) models, reflecting their better monolingual performance. λ values are at Table 3's caption.

[26] Following (Faruqui and Dyer, 2014) we also tried to take only one of the monolingual projected vectors as the multilingual representation. The results were very similar to what we report below: no improvement over monolingual models.

[27] We present these numbers only for the word2vec model due to space limitations. The effect for the BOW model is identical in terms of the number of judgment languages for which each model combination improves over a model trained on the same language that was used for the human judgment, but the average ρ improvement is smaller.

[28] A simple concatenation of the monolingual vectors is also an effective combination method for monolingual models, leading to improvements that are similar to what we report for linear interpolation. However, simple concatenation is effective for the BOW model only when PPMI normalization is applied to the row counts, as opposed to linear interpolation which is effective regardless of this step. We therefore focus on linear interpolation, the more robust method, as a combination method in this paper.

	Eng.	Ger.	Ital.	Russ.
Eng.	—	4/4 (0.0409)	4/4 (0.0452)	4/4 (0.0419)
Ger.	—	—	3/4 (0.0063)	2/4 (-0.0124)
Ital.	—	—	—	2/4 (0.0074)
Russ.	—	—	—	—
Mono.	2/3 (0.0160)	0/3 (-0.0961)	0/3 (-0.1128)	0/3 (-0.0882)

Table 5: Results summary for the linear interpolation model combination method with word2vec VSMs. In the four top lines the (L_1, L_2) entry contains the number of judgment languages for which a combination of L_1 and L_2 trained VSMs outperforms a VSM of the same type when its judgment and training languages are identical (the average improvement is given in parenthesis). For comparison, the L_2 entry of the bottom line presents the same statistics for a monolingual VSM trained on L_2 only.

Like in Section 6, we controlled against the corpus size effect. Even when the English and German corpora were cut to the size of the Russian or the Italian corpus the result difference was up to 0.01 Spearman, Pearson or Kandall Tau points.

Interestingly, as opposed to the results reported in (Faruqui and Dyer, 2014) for the English judgment language, our CCA combination experiments resulted in no improvement over the monolingual models (Table 4). This is also in contrast to the positive results reported for this method for the combination of visual and textual representations (Silberer and Lapata, 2012; Hill et al., 2014a).

8 Conclusions

In this paper we demonstrated the importance of the *judgment language*, the language in which word pairs are presented to the human annotators, to the resulted similarity ratings. By translating WS353 to three languages and collecting similarity scores for the translated sets, we were further able to show the impact of the VSM training language on the correlation of its similarity ratings with the ratings produced by humans that used different judgment languages, and to explore the power of monolingual model combination.

In future work we intend to extend our inquiry to relations beyond word similarity and to a larger number of training and judgment languages. In addition, we intend to explore more sophisticated methods for multilingual VSM construction. Finally, we would like to extend our analysis beyond

quantitative exploration. A good way to start tackling this is through a qualitative analysis of the performance of the various VSMs on specific word pairs in the different judgment languages - we already have initial promising findings in this direction.

References

[Al-Rfou et al.2014] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Polyglot: Distributed word representations for multilingual nlp. In *Proc. of CoNLL*.

[Baker et al.2014] Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proc. of EMNLP*.

[Baroni et al.2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of ACL*.

[Bengio et al.2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *JMLR*.

[Bruni et al.2014] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47.

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

[Collobert and Weston2008] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of ICML*.

[Collobert et al.2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, 12:2493–2537.

[Faruqui and Dyer2014] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of EACL*.

[Finkelstein et al.2001] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proc. of WWW*.

[Halawi et al.2012] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proc. ACM SIGKDD*.

[Hardoon et al.2004] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

[Harris1954] Zellig Harris. 1954. Distributional structure. *Word*.

[Hermann and Blunsom2014a] Karl Moritz Hermann and Phil Blunsom. 2014a. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*.

[Hermann and Blunsom2014b] Karl Moritz Hermann and Phil Blunsom. 2014b. Multilingual models for compositional distributed semantics. In *Proc. of ACL*.

[Hill et al.2014a] Felix Hill, Roi Reichart, and Anna Korhonen. 2014a. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2(10):285–296.

[Hill et al.2014b] Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv:1408.3456 [cs.CL]*.

[Huang et al.2012] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*.

[Khapra et al.2013] Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2013. Multilingual deep learning. In *Proceedings of NIPS Workshop on Deep Learning*.

[Kiela and Clark2014] Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proc. of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), EACL*.

[Klementiev et al.2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proc. of COLING*.

[Kočiský et al.2014] Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *Proc. of ACL*.

[Lauly et al.2013] Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2013. Learning multilingual word representations using a bag-of-words autoencoder. In *Proc. of NIPS Workshop on Deep Learning*.

[Lauly et al.2014] Stanislas Lauly, Hugo Larochelle, Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, Amrita Saha, et al. 2014. An autoencoder approach to learning bilingual word representations. In *Proc. of NIPS*.

[Levy and Goldberg2014] Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proc. of ACL (Volume 2: Short Papers)*.

[Luong et al.2013] Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proc. of CoNLL*.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

[Mikolov et al.2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*.

[Mikolov et al.2013c] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proc. of NAACL-HLT*.

[Miller and Charles1991] George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

[Radinsky et al.2011] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proc. of WWW'*.

[Rubenstein and Goodenough1965] Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

[Silberer and Lapata2012] Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proc. of EMNLP-CoNLL*.

[Turney et al.2010] Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

[Yang and Powers2006] Dongqiang Yang and David MW Powers. 2006. Verb similarity on the taxonomy of wordnet. In *Proc. of GWC-06*.