Comparison of the Best Parameter Settings in the Creation and Comparison of Feature Vectors in Distributional Semantic Models Across Multiple Languages

András Dobó(✉) and János Csirik

University of Szeged, Szeged, Hungary
{dobo,csirik}@inf.u-szeged.hu

Abstract. Measuring the semantic similarity and relatedness of words is important for many natural language processing tasks. Although distributional semantic models designed for this task have many different parameters, such as vector similarity measures, weighting schemes and dimensionality reduction techniques, there is no truly comprehensive study simultaneously evaluating these parameters while also analysing the differences in the findings for multiple languages. We would like to address this gap with our systematic study by searching for the best combination of parameter settings in the creation and comparison of feature vectors in distributional semantic models for English, Spanish and Hungarian separately, and then comparing our findings across these languages.

During our extensive analysis we test a large number of possible settings for all parameters, with more than a thousand novel variants in case of some of them. As a result of this we were able to find such combinations of parameter settings that significantly outperform conventional settings combinations and achieve state-of-the-art results.

Keywords: Distributional semantic models · Semantic similarity and relatedness · Best combination of parameter settings · Comparison of findings across languages · English, Spanish and Hungarian

1 Introduction

In many NLP problems, including information retrieval [16], spelling correction [3] and noun compound interpretation [14] among many others, knowing the semantic similarity or semantic relatedness of words can be very useful. Although distributional semantic models (DSMs) [1] calculating these measures

Published by Springer Nature Switzerland AG 2019
J. MacIntyre et al. (Eds.): AIAI 2019, IFIP AICT 559, pp. 487–499, 2019.
https://doi.org/10.1007/978-3-030-19823-7_41

have many possible parameters, most research focuses only on one or two aspects of these models, while using some conventional settings for the rest of the parameters (e.g. cosine as vector similarity and (positive) pointwise mutual information as weighting). Therefore a truly comprehensive study evaluating the numerous parameters of DSMs for any language is still missing, and would be needed, as also suggested by [21]. Moreover, despite the fact that the best parameter settings for the parameters can differ for different languages, the vast majority of papers consider DSMs for only one language (mostly English), or consider multiple languages but without a real comparison of findings across languages. In this article we would like to address these gaps.

There are two distinct phases of DSMs in general: the extraction of statistical information from raw text, and the creation and comparison of feature vectors for words based on the extracted information. Within this study we focus on the parameters of the second phase, as the two phases are relatively distinct from each other, and the number of possible combinations of parameter settings (CPS) for the second phase is already well in the trillions. So we are searching for the best CPS of those 10 parameters during the creation and comparison of feature vectors in DSMs that we considered important, for English, Spanish and Hungarian separately, and then we compare our findings for the different languages.

A detailed description of our analysis can be found in [13], where tests were done only for English, without any comparison of findings across languages, and with far less settings tested for several parameters.

2 Background

Although there are a vast number of studies dealing with DSMs, most of them only consider one or two parameters of these models, and take the others granted with some standard setting. Most commonly, these models use cosine as vector similarity [1,2,6,8,17,21,27,29,30,32,34,35] and (positive) pointwise mutual information as weighting [2,17,18,21,29,30,35]. Further, they also usually do not care for the interaction of these parameters, and experiment with the considered parameters one by one, and not simultaneously. Of course there are some studies that experiment with several parameters with multiple possible settings [7,19,20], but even these are far from being truly comprehensive.

Moreover, most models were only tested for English and neglect any other languages despite the fact that DSMs might work differently across multiple languages. Of course, there are several studies in which results were presented for languages other than English, including Spanish [4,15,23] and Hungarian [11,24]. However, even those that include multiple languages usually only present some test results for the different languages separately, without any real analysis of the differences in the findings between the languages.

3 Data and Evaluation Methods

For input we used information extracted from the British National Corpus (BNC), the Spanish Wikicorpus (EsWiki) [28], and the 23.01.2012 dump of the Hungarian Wikipedia (HuWiki) for English, Spanish and Hungarian, respectively, with the help of the information extraction methods of [11], [12] and [29].

Tests were done on parts of the MEN [2] dataset for English, the Spanish WordSimilarity-353 [15], the Moldovan [23] and the Spanish Rubenstein-Goodenough [5] datasets for Spanish, and parts of the Hungarian version of the TOEFL [11] and Rubenstein-Goodenough datasets for Hungarian. The last was constructed the same way as the Hungarian TOEFL and Miller-Charles datasets in [11].

Out of these datasets only the MEN dataset is truly reliable, as the others are rather small and except for the Moldovan dataset just translated from English datasets, during which they can be distorted. The Hungarian datasets are especially small, and the type of the TOEFL dataset also makes the results on it even less reliable compared to the other datasets. However, due to the lack of truly suitable resources, we had to settle for these.

In case of the TOEFL dataset, the accuracy (A) of the models on the questions were calculated, while in case of all the other datasets, the Pearson's (P) and Spearman's (S) correlations with the gold standard scores and their modified harmonic mean (H) were computed, as follows:

$$H(P,S) = \frac{2 \times P \times S}{|P| + |S|} \tag{1}$$

For more information, please refer to [13].

4 Our Heuristic Analysis

As the number of possible CPSs are in the magnitude of trillions, we had to use a heuristic approach to find the best one in case of each language. First, each parameter was tested separately on a development dataset, where a candidate list of settings were selected for each parameter. Then all combinations of the selected settings of all parameters were tested on a different dataset, to find the best CPS. These were done for the three languages separately.

The 10 parameters tested, together with the number of settings tried for them, are listed in Table 1. For several parameters, a large number of novel settings were tested. These were either brand new settings, modified versions of conventionally used settings, or combinations of multiple settings. A detailed description of the tested parameters and their settings can be found in [13].

We have to note that when using singular value decomposition (SVD) for dimensionality reduction or the various smoothing options, we usually did a smaller number of runs than in other cases due to our limited resources. Further, in case of Spanish we had to set MWFFreq to 3 instead of NoLimit when using SVD due to the too many features otherwise, which would have made running SVD unmanageable.

First we have done this two-step heuristic analysis for English and evaluated the results extensively in [13]. Then we have repeated the same analysis, with a greatly increased number of settings for several parameters, for English, Spanish and Hungarian, and compared the findings across the different languages in this article.

5 Results

5.1 Results of the First Phase

During the first phase of our analysis multiple runs were done for each setting of every parameter, and the most promising ones in case of each parameter were selected to be included in the second phase. In case of English, we used half of the development part of the MEN dataset for evaluation, while for Spanish the Spanish WordSimilarity-353 dataset and for Hungarian half of the Hungarian TOEFL dataset was employed. The top 5 performing settings for each parameter are listed in Table 2 in case of each language.

Table 1. The tested parameters, with the number of settings tested for each

Parameter	Abbreviation	Count
Vector similarity	VecSim	1221
Weighting scheme	Weight	2907
Feature transformation	FeatTranf	22
Dimensionality reduction	DimRed	21
Smoothing	Smooth	5
Vector normalization	VNorm	3
Stop-word filtering	StopW	2
Minimum limits on word-feature frequencies	MWFFreq	6
Minimum limits on word-feature weights	MWFWeight	26
Minimum limits on feature frequencies	MFFreq	14

Although presenting the definition of all settings for every parameter would be impossible within this article due to their large number, below we briefly define a couple of them to help interpreting our most important results.

In case of vector similarity measures, we have defined many new variants based on one or more conventional measures. For example, the best one for

English is a combination of the Pearson, MarylandBridge [9] and AdjCos [31] measures, with some additional transformations:

$$
\begin{aligned}
&PearsMbAdjCosMod\text{-}3.Lb(u,v) = \begin{cases} 1, & d \geq 0.1 \\ \frac{d}{0.1}, & d < 0.1 \end{cases} \\
&d = 0.5 \times \left(\frac{\sum_{i=1}^{n} sgn(u_i - \bar{u}) \times lb(|u_i - \bar{u}| + 1) \times sgn(v_i - \bar{v}) \times lb(|v_i - \bar{v}| + 1)}{lbinv\left(\sum_{i=1}^{n} (lb(|u_i - \bar{u}| + 1))^2\right)} \right. \\
&\qquad \left. + \frac{\sum_{i=1}^{n} sgn(u_i - \bar{u}) \times lb(|u_i - \bar{u}| + 1) \times sgn(v_i - \bar{v}) \times lb(|v_i - \bar{v}| + 1)}{lbinv\left(\sum_{i=1}^{n} (lb(|v_i - \bar{v}| + 1))^2\right)} \right) \\
&lbinv(x) = min(max(sgn(x) \times (2^{|x|} - 1), -2^{100}), 2^{100})
\end{aligned}
\tag{2}
$$

On the other hand, the best vector similarity measure for Spanish is rather different. It is a modified and transformed version of the Hindle$_r$ measure [22]:

$$
\begin{aligned}
&LinHindleRMod\text{-}7.1.2.Cu(u,v) = \frac{\sqrt[3]{\sum_{i=1}^{n} lhr\text{-}1.Cu(u_i, v_i)}}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}} \\
&lhr\text{-}1.Cu(x,y) = \begin{cases} min(x^3, y^3), & x \neq 0 \wedge y \neq 0 \\ 0, & otherwise \end{cases}
\end{aligned}
\tag{3}
$$

Weighting schemes were constructed similarly as vector similarity measures, and here too there were also numerous new variants. For example, the best one for English is a combination of PMIAplha [21], PMI with Laplace smoothing [33], Unisubtuples [26] and PMI with discounting factor [25]:

$$
\begin{aligned}
&PmiAl\text{-}Tc3Tw0S2P4(x,y) = \frac{f'_{xy}}{f'_{xy} + 1} \times \frac{min(f'_x, f'_y)}{min(f'_x, f'_y) + 1} \\
&\quad \times \left(lb\left(\frac{n'_\alpha \times f'_{xy}}{f'_x \times f'^{0.75}_y} \right) - 3.29 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right) \\
&a = f'_{xy}, \quad b = f'_x - f'_{xy}, \quad c = f'_y - f'_{xy}, \quad d = n' - f'_x - f'_y + f'_{xy} \\
&f'_x = f_x + 1, \;\; f'_y = f_y + 1, \;\; f'_{xy} = f_{xy} + 1, \;\; n' = n + 1, \;\; n'_\alpha = \left(\sum_{i=1}^{|V|} f_i^{0.75} \right) + 1
\end{aligned}
\tag{4}
$$

f_x, f_y: word frequencies, $f_{x,y}$: xy tuple frequency

n: total number of words in the corpus, $|V|$: size of the vocabulary

In case of feature transformation, we have experimented with transforming either raw frequencies or weights, both before and after normalization, and 7 different transformation functions were tried in all cases. For smoothing, we tried different versions of the Kneser-Ney smoothing [10]. In case of dimensionality reduction, we tried a couple of different techniques, including SVD and the method of [18]. And for minimum limits on word-feature weights we have tried the following two novel variants with multiple limit values:

$$
limit(w, minValue) = \begin{cases} w & if\ w \geq minValue \\ minValue & otherwise \end{cases}
\tag{5}
$$

$$zero(w, minValue) = \begin{cases} w & if\ w \geq minValue \\ 0 & otherwise \end{cases} \quad (6)$$

The other parameters are much less complex and more commonly used in NLP, thus one should be able to have enough understanding of them from Table 1.

A more detailed description of the different parameters and settings can be found in [13] (although with far less settings for several parameters).

5.2 Results of the Second Phase

In the second phase all possible combinations of the settings of each parameter were tested in case of all three languages, in order to find the best CPS for all

Table 2. The top 5 performing setting for each parameter in case of all 3 languages, in descending order of maximum H scores

Parameter	English		Spanish		Hungarian	
	Setting	H	Setting	H	Setting	H
VecSim	PearsMbAdjCosMod-3.Lb	0.71	LinHindleRMod-7.1.2.Cu	0.37	PearsMbMod-1.Lb	0.80
	PearsMbAdjCosMod-4.Lb	0.71	LinHindleRMod-6.1.2.Cu	0.36	PearsMbMod-4.Lb	0.80
	PearsMbAdjCosMod-2.Lb	0.71	LinHindleRMod-1.1.2.Cu	0.36	MbMod-6.Lb	0.80
	PearsMbAdjCosMod-6.Lb	0.71	LinHindleRMod-7.1.2.Sq	0.36	PearsMbMod-5.Lb	0.80
	PearsMbAdjCosMod-6.Sigm	0.71	LinHindleRMod-3.1.2.Sq	0.36	PearsMbMod-2.Lb	0.80
Weight	PmiAl-Tc3Tw0S2P4	0.70	PmiAlUnis-Tc4Tw3S2P2	0.38	PmiAl-Tc4Tw2S1P1	0.85
	PmiAl-Tc3Tw0S2P0	0.70	PmiAlUnis-Tc4Tw3S2P1	0.38	PmiAlUnisAm-Tc0Tw3S2P1	0.85
	PmiAlUnis-Tc3Tw0S0P4	0.70	PmiAlUnis-Tc4Tw2S1P1	0.38	PmiAlUnisAm-Tc0Tw2S2P2	0.85
	PmiAlUnis-Tc3Tw0S0P0	0.70	PmiAlUnisAm-Tc4Tw3S2P5	0.38	PmiAl-Tc4Tw3S0P2	0.85
	PmiAl-Tc4Tw0S2P5	0.70	PmiAlUnis-Tc4Tw2S1P2	0.38	PmiAlUnisAm-Tc0Tw2S2P1	0.85
FeatTransf	Weight AftNorm Lb	0.67	Weight AftNorm Lb	0.34	Weight AftNorm Sqrt	0.85
	Freq Sq	0.67	Weight AftNorm Sigm	0.34	Weight BefNorm Sqrt	0.85
	Weight BefNorm Sigm	0.67	NoTransf	0.34	Weight AftNorm Sigm	0.80
	Weight AftNorm Sigm	0.67	Weight BefNorm Lb	0.34	NoTransf	0.80
	NoTransf	0.67	Weight BefNorm Sigm	0.34	Weight AftNorm Lb	0.80
DimRed	SVD 200	0.70	SVD 100	0.37	IslamInkpen 0.025	0.80
	SVD 100	0.70	SVD 200	0.36	IslamInkpen 0.25	0.80
	SVD 300	0.69	SVD 500	0.35	SVD 200	0.80
	SVD 500	0.68	SVD 300	0.34	IslamInkpen 0.005	0.78
	IslamInkpen 0.05	0.67	IslamInkpen 0.01	0.34	IslamInkpen 0.01	0.78
Smooth	NoSmooth	0.67	Freq KNS	0.34	Freq KNS	0.83
	Weight KNS	0.65	Freq MDKNSPOMD	0.34	Freq MDKNSPOMD	0.80
	Freq KNS	0.62	NoSmooth	0.34	NoSmooth	0.78
	Freq MDKNSPOMD	0.62	Freq MKNS	0.33	Weight KNS	0.75
	Freq MKNS	0.57	Weight KNS	0.31	Freq MKNS	0.73
VNorm	L_2	0.67	L_2	0.34	L_2	0.80
	L_1	0.67	L_1	0.34	NN	0.78
	NN	0.67	NN	0.34	L_1	0.75
StopW	false	0.67	true	0.34	false	0.80
	true	0.67	false	0.34	true	0.75
MWFFreq	NoLimit	0.67	NoLimit	0.34	NoLimit	0.80
	2	0.60	2	0.30	3	0.68
	3	0.57	3	0.29	2	0.63
	5	0.54	7	0.28	5	0.58
	7	0.49	5	0.27	7	0.55
MWFWeight	Zero 0.05	0.68	Zero -0.2	0.34	Zero 0	0.80
	Zero 0.1	0.68	Limit -0.1	0.34	Limit -0.02	0.80
	Zero -0.05	0.68	Limit -0.2	0.34	Zero -0.05	0.80
	Limit -0.01	0.67	Limit -0.5	0.34	Limit -0.01	0.80
	Zero 0.02	0.67	NoLimit	0.34	Zero -0.01	0.80
MFFreq	NoLimit	0.67	100	0.36	2	0.80
	2	0.67	50	0.36	NoLimit	0.80
	3	0.67	30	0.35	3	0.80
	5	0.67	20	0.35	20	0.80
	7	0.67	15	0.35	15	0.80

languages. The second half of the development part of the MEN dataset was used for testing in case of English, while the Moldovan dataset and the second part of the Hungarian TOEFL dataset were used for Spanish and Hungarian, respectively. The top 5 performing CPSs for each language are presented in Table 3.

5.3 Results on the Test Datasets

The best CPS for English was tested on the test part of the MEN dataset (MT), and the best CPSs for Spanish and Hungarian were tested on the respective version of the Rubenstein-Goodenough dataset (RG) to give us the final results. The best CPS of each language was also evaluated on the datasets of the other languages, to provide us a way of comparison. The results of these test can be found in Table 4.

6 Evaluation and Discussion

In this section we evaluate our results presented in the previous sections. Please note that the scores are not fully comparable across languages, even when considering the same datasets on different languages, as except for the Moldovan dataset all of the used Spanish and Hungarian datasets were constructed by translating the English versions, and thus the results on them can be distorted and less reliable than on their English counterparts. Furthermore, the Spanish and Hungarian datasets, especially the latter ones, are rather small, which also makes them less reliable than the English ones.

As there are many differences in the syntax and morphology of the different languages, we anticipated from the beginning that there will be at least some small differences in our findings for the different languages. However, our intuition was that our findings for the different languages will be subtle, and we will be able to find good and rather language-independent CPSs. As English and Spanish belong to the family of Indo-European languages, while Hungarian does not, we expected that the results for English and Spanish will be similar due to this. Further, as both Spanish and Hungarian have very rich morphology, we expected that there will also be a higher similarity between our results for Spanish and Hungarian because of this. We anticipated that the least similarities will be between English and Hungarian, as these languages are the least similar to each other.

In the first phase of our analysis we could observe that some of the parameters worked exactly the same way or very similarly across languages. These parameters were the weighting scheme, feature transformation, vector normalization and minimum limits on word-feature frequencies. These finding are in line with our initial intuitions. Dimensionality reduction seemed to be similar for English and Spanish, while a bit different for Hungarian. Smoothing seemed to perform similarly for Spanish and Hungarian, while differently for English. Minimum limits on word-feature weights seem to behave a bit differently for all

Table 3. The top 5 performing CPSs for each language with their achieved scores, in descending order of maximum H scores

Lang	#	Parameter settings							P	S	H
En	1	**VecSim**		**Weight**			**FeatTransf**		0.72	0.71	0.71
		PearsMbAdjCosMod-3.Lb		PmiAl-Tc3Tw0S2P0			NoTransf				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		IslamInkpen 0.05	NoSmooth	L_1	false	NoLimit	Zero 0	NoLimit			
	2	**VecSim**		**Weight**			**FeatTransf**		0.72	0.71	0.71
		PearsMbAdjCosMod-3.Lb		PmiAl-Tc3Tw0S2P0			NoTransf				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		NoDimRed	NoSmooth	L_1	false	NoLimit	Zero 0	NoLimit			
	3	**VecSim**		**Weight**			**FeatTransf**		0.72	0.71	0.71
		PearsMbAdjCosMod-3.Lb		PmiAl-Tc3Tw0S2P0			NoTransf				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		NoDimRed	NoSmooth	L_1	false	NoLimit	Zero -0.05	NoLimit			
	4	**VecSim**		**Weight**			**FeatTransf**		0.72	0.71	0.71
		PearsMbAdjCosMod-3.Lb		PmiAl-Tc3Tw0S2P0			NoTransf				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		IslamInkpen 0.05	NoSmooth	L_1	false	NoLimit	Zero -0.05	NoLimit			
	5	**VecSim**		**Weight**			**FeatTransf**		0.72	0.71	0.71
		PearsMbAdjCosMod-4.Lb		PmiAl-Tc3Tw0S2P0			NoTransf				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		NoDimRed	NoSmooth	L_1	false	NoLimit	Zero -0.05	NoLimit			
Es	1	**VecSim**		**Weight**			**FeatTransf**		0.43	0.44	0.44
		Cos		Pmi-Tc1Tw3S2P0			Weight AftNorm Lb				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		SVD 100	NoSmooth	L_2	true	NoLimit	NoLimit	3			
	2	**VecSim**		**Weight**			**FeatTransf**		0.43	0.43	0.43
		Cos		PmiAl-Tc3Tw3S2P0			Weight AftNorm Lb				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		SVD 100	NoSmooth	L_2	true	NoLimit	NoLimit	3			
	3	**VecSim**		**Weight**			**FeatTransf**		0.43	0.43	0.43
		Cos		Pmi-Tc1Tw3S2P0			Weight AftNorm Lb				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		SVD 100	NoSmooth	L_2	true	NoLimit	NoLimit	100			
	4	**VecSim**		**Weight**			**FeatTransf**		0.43	0.43	0.43
		Cos		PmiAl-Tc3Tw3S2P0			Weight AftNorm Lb				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		SVD 100	NoSmooth	L_2	false	NoLimit	NoLimit	3			
	5	**VecSim**		**Weight**			**FeatTransf**		0.43	0.43	0.43
		Cos		Pmi-Tc1Tw3S2P0			Weight AftNorm Lb				
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**			
		SVD 100	NoSmooth	L_1	true	NoLimit	NoLimit	3			

Lang	#	Parameter settings							A
Hu	1	**VecSim**		**Weight**			**FeatTransf**		0.65
		MbCosAm		NPmiAlpha-Tc4Tw4S0P4			Weight AftNorm Sqrt		
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**	
		SVD 200	NoSmooth	L_2	false	NoLimit	NoLimit	2	
	2	**VecSim**		**Weight**			**FeatTransf**		0.65
		MbCosAm		NPmiAlpha-Tc4Tw4S0P4			Weight AftNorm Sqrt		
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**	
		SVD 200	NoSmooth	L_1	false	NoLimit	NoLimit	2	
	3	**VecSim**		**Weight**			**FeatTransf**		0.65
		Cos		NPmiAlpha-Tc4Tw4S0P4			Weight AftNorm Sqrt		
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**	
		SVD 200	NoSmooth	L_2	false	NoLimit	NoLimit	2	
	4	**VecSim**		**Weight**			**FeatTransf**		0.65
		Cos		NPmiAlpha-Tc4Tw4S0P4			Weight AftNorm Sqrt		
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**	
		SVD 200	NoSmooth	L_1	false	NoLimit	NoLimit	2	
	5	**VecSim**		**Weight**			**FeatTransf**		0.63
		PearsMbMod-1.Lb		NPmiAlpha-Tc4Tw4S0P4			Weight AftNorm Sqrt		
		DimRed	**Smooth**	**VNorm**	**StopW**	**MWFFreq**	**MWFWeight**	**MFFreq**	
		SVD 200	NoSmooth	L_2	false	NoLimit	NoLimit	2	

Table 4. Results on the test datasets, in descending order of maximum H scores

Lang	Test set	CPS	P	S	H
En	MT	BestEn	0.71	0.71	0.71
		BestHu	0.67	0.68	0.67
		BestEs	0.63	0.63	0.63
Es	RG	BestHu	0.83	0.83	0.83
		BestEn	0.82	0.80	0.81
		BestEs	0.80	0.79	0.80
Hu	RG	BestEn	0.73	0.72	0.72
		BestEs	0.65	0.61	0.63
		BestHu	0.58	0.68	0.62

three languages. However, it was interesting to see that the results for vector similarity measures, stop-word filtering and minimum limits on feature frequencies were rather similar for English and Hungarian, but different for Spanish, which is contrary to what we anticipated.

In the second phase, although there were similarities in the found best CPSs across the different languages, one could also observe many differences. Here too, the weighting schemes, feature transformation and minimum limits on word-feature frequencies were mostly similar. Compared to the first phase vector similarity, smoothing and minimum limits on feature frequencies were also alike for all languages. The other parameters showed a different behaviour for at least one language compared to the others.

We have to note here that there were actually two distinct CPSs with the same best score for English, and they were only different in their DimRed parameter setting. We have chosen the one with the "IslamInkpen 0.05" setting as BestEn, as that setting achieved better performance in the first phase than the "NoDimRed" setting in the other CPS. Furthermore, for Hungarian there were even more CPSs with the same best score. We have used a similar approach in selecting the BestHu version, as we have done in case of the BestEn version. However, as these different CPSs with the same best results have different settings in case of some parameters, one has to be careful drawing conclusions from the best CPSs of the different languages, and thus any conclusions drawn from them should be taken with some reservations.

The final conclusions for the parameters are the following:

- VecSim: for all languages measures based on cosine similarity achieve the best results
- Weight: measures based on PMI dominate the top of the table by far in case of all languages
- FeatTransf: no transformation and transforming the word-feature weights after normalization preforms best for all languages

- DimRed: dimensionality reduction seems to help in most situations: while in case of English the IslamInkpen version performed the best alongside no dimensionality reduction, for Spanish and Hungarian SVD is superior to these options
- Smooth: the no smoothing option clearly outperforms all others for all languages
- VNorm: for English the L_1 option clearly seems to be the best, while for Spanish and Hungarian the best CPSs use either L_1 or L_2 normalization, and most CPSs achieve the same or very similar results with either
- StopW: stop-word filtering seems to improve the results to some extent in case of Spanish, while it does not in case of English and Hungarian
- MWFFreq: no limit is by far superior to the other options for all languages
- MWFWeight: no limit seems to be the best option in case of Spanish and Hungarian, while the Zero option with different parameters seems to excel in case of English
- MFFreq: a low limit or no limit seems to be best in case of all languages (as noted before, in case of SVD for Spanish we had to use a limit of 3 instead of no limit for computational reasons)

As we anticipated, there were parameters where the results for Spanish and Hungarian were similar, but different for English. However, it was interesting that we did not find any parameters that were alike for English and Spanish, but different for Hungarian. Further, to our surprise we found such a parameter, where the results were similar for English and Hungarian, but different for Spanish. These latter findings were in contrast to our initial intuition.

Although all Spanish scores in the second phase are much lower than the English and Hungarian ones, these are almost completely due to the dataset used, and do not mean that the found Spanish CPSs are worse than their English and Hungarian counterparts, as it was noted in the beginning of this section and can be seen from our results on the test datasets (see Table 4) too. It simply suggests that the dataset used for Spanish in this phase is considerably tougher than the ones used for English and Hungarian.

Table 5. Comparison of our best English CPS with a conventional and a state-of-the-art CPS, using the information extracted from the BNC with the method of [29] and the MT dataset for all tests

CPS	P	S	H
BestEn	0.67	0.67	0.67
OSC of [29]	0.59	0.58	0.58
CosPPmi	0.56	0.58	0.57

It was interesting to see that in the cross-language experiments on the test datasets the order of the Best CPSs of the different languages with respect to their performance is different in case of the datasets of the three languages. The best English CPS was always superior to its Spanish counterpart, but it has no

absolute superiority over the best Hungarian CPS. Further, there is also no clear ranking between the best Spanish and Hungarian CPSs. It was also interesting to see that in case of the Spanish dataset, although the best Spanish CPS achieved rather good results, actually it achieved the lowest score out of the three best CPSs tested. It was the same for the best Hungarian CPS on the Hungarian dataset too.

All in all, there seems to be no clear ranking between the best CPSs of the different languages, and all of them achieved good results on the datasets of all languages. So, although we got different best CPSs for the different languages, all of them seem to be rather language-independent. These findings give us a strong intuition that our heuristic approach was good, and that our found best CPSs for all languages and their results are robust and reliable.

To further prove that our heuristic approach was successful, that our results are robust and reliable, and that our found best CPSs perform much better than conventional CPSs, we compared our best English CPS with the conventional cosine with positive pointwise mutual information setting (CosPPmi) and the state-of-the-art original settings combination (OSC) of [29], using the information extracted from the BNC with the method of [29] and the MT dataset for all tests. The results of these tests, presented in Table 5, clearly show that our found best CPS is robust, and is not just superior to conventional settings, but to a current state-of-the-art CPS too.

The fact that the best CPSs found in the second phase are not simply made up of the best parameter settings in the first phase proves that our intuition was correct, and the parameters of DSMs need to be tested simultaneously, rather than separately.

7 Conclusions

Within this article we have presented a systematic analysis of the parameters in the creation and comparison of feature vectors in distributional semantic models for English, Spanish and Hungarian, including novel parameters and novel parameter settings. To our best knowledge, we are the first to do such a detailed analysis for these parameters, and also to do such an extensive comparison of them across multiple languages.

With our heuristic approach we were searching for the best combination of parameter settings for all three languages. In accordance with our intuition, there were several parameters that worked very similarly in case of all three languages. We also found such parameters that were alike for Spanish and Hungarian, and different for English, which we also anticipated. However, it was interesting to see that there was such a parameter that worked similarly for English and Hungarian, but not for Spanish, and we did not find any parameters that worked similarly for the two Indo-European languages, but differently for Hungarian.

Although we have found that the very best results are produced by different settings combinations for the different languages, our cross-language tests showed that all of them work rather well for all languages. Based on this we think that

our heuristic approach was successful, and we could find such combinations of parameter settings that are rather language-independent, and give robust and reliable results. Further, our best English CPS, incorporating multiple novel parameter settings, significantly outperformed both conventional and state-of-the-art parameter combinations.

Although our results seem rather robust and reliable for Spanish and Hungarian too, it would be interesting to redo our analysis on larger and more reliable Spanish and Hungarian datasets to check whether we could find even better CPSs for these languages, when such datasets will become available in the future.

References

1. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: 52nd ACL, pp. 238–247 (2014)
2. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res. **48**, 1–47 (2013)
3. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: NAACL Workshop on WordNet and Other Lexical Resources, p. 2 (2001)
4. Camacho-Collados, J., Pilehvar, M.T., Collier, N., Navigli, R.: SemEval-2017 task 2: multilingual and cross-lingual semantic word similarity. In: SemEval-2017, pp. 15–26 (2017)
5. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: A framework for the construction of monolingual and cross-lingual word similarity datasets. In: 53rd ACL, pp. 1–7 (2015)
6. Christopoulou, F., Briakou, E., Iosif, E., Potamianos, A.: Mixture of topic-based distributional semantic and affective models. In: 12th IEEE ICSC, pp. 203–210 (2018)
7. Curran, J.R.: From Distributional to Semantic Similarity. University of Edinburgh (2004)
8. De Deyne, S., Perfors, A., Navarro, D.J.: Predicting human similarity judgments with distributional models: the value of word associations. In: 26th IJCAI, pp. 4806–4810 (2017)
9. Deza, M.M., Deza, E.: Encyclopedia of distances. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-642-00234-2
10. Dobó, A.: Multi-D Kneser-Ney smoothing preserving the original marginal distributions. Res. Comput. Sci. **147**(6) (2018)
11. Dobó, A., Csirik, J.: Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása. In: IX. Magyar Számítógépes Nyelvészeti Konferencia, pp. 213–224 (2012)
12. Dobó, A., Csirik, J.: Computing semantic similarity using large static corpora. In: van Emde Boas, P., Groen, F.C.A., Italiano, G.F., Nawrocki, J., Sack, H. (eds.) SOFSEM 2013. LNCS, vol. 7741, pp. 491–502. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35843-2_42
13. Dobó, A., Csirik, J.: A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models. J. Quant. Linguist. (2019). https://doi.org/10.1080/09296174.2019.1570897

14. Dobó, A., Pulman, S.G.: Interpreting noun compounds using paraphrases. Procesamiento del Lenguaje Natural **46**, 59–66 (2011)
15. Hassan, S., Mihalcea, R.: Cross-lingual semantic relatedness using encyclopedic knowledge. In: 2009 CoNLL, pp. 1192–1201 (2009)
16. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., Milios, E.: Information retrieval by semantic similarity. Int. J. Seman. Web Inf. Syst. **2**(3), 55–73 (2006)
17. Iosif, E., Georgiladakis, S., Potamianos, A.: Cognitively motivated distributional representations of meaning. In: 10th LREC (2016)
18. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Disc. From Data **2**(2), 1–25 (2008)
19. Kiela, D., Clark, S.: A systematic study of semantic vector space model parameters. In: 2nd CVSC at EACL, pp. 21–30 (2014)
20. Lapesa, G., Evert, S.: A large scale evaluation of distributional semantic models: parameters, interactions and model selection. Trans. Assoc. Comput. Linguist. **2**, 531–545 (2014)
21. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Trans. Assoc. Comput. Linguist. **3**, 211–225 (2015)
22. Lin, D.: Automatic retrieval and clustering of similar words. In: 36th ACL, pp. 768–774 (1998)
23. Moldovan, C.D., Ferré, P., Demestre, J., Sánchez-Casas, R.: Semantic similarity: normative ratings for 185 Spanish noun triplets. Behav. Res. Methods **47**(3), 788–799 (2015)
24. Novák, A., Novák, B.: Magyar szóbeágyazási modellek kézi kiértékelése. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia, pp. 66–77 (2018)
25. Pantel, P., Lin, D.: Discovering word senses from text. In: 8th ACM SIGKDD, vol. 41, p. 613 (2002)
26. Pecina, P.: Lexical association measures and collocation extraction. Lang. Res. Eval. **44**(1–2), 137–158 (2010)
27. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. EMNLP **2014**, 1532–1543 (2014)
28. Reese, S., Boleda, G., Cuadros, M., Padró, L., Rigau, G.: Wikicorpus: a word-sense disambiguated multilingual wikipedia corpus. In: 7th LREC, pp. 1418–1421 (2010)
29. Salle, A., Idiart, M., Villavicencio, A.: Matrix factorization using window sampling and negative sampling for improved word representations. In: 54th ACL, p. 419 (2016)
30. Salle, A., Idiart, M., Villavicencio, A.: LexVec (2018). https://github.com/alexandres/lexvec/blob/master/README.md. Accessed 04 July 2018
31. Shalaby, W., Zadrozny, W.: Measuring semantic relatedness using mined semantic analysis. arXiv preprint arXiv:1512.03465 (2016)
32. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An open multilingual graph of general knowledge. In: 31st AAAI, pp. 4444–4451 (2017)
33. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**, 141–188 (2010)
34. Vakulenko, M.: Calculation of semantic distances between words: from synonymy to antonymy. J. Quant. Linguist. 1–13 (2018)
35. Yih, W.T., Arbor, A.: Measuring word relatedness using heterogeneous vector space models University of Michigan. In: NAACL-HLT 2012, pp. 616–620 (2012)