Determining Compositionality of Word Expressions Using Various Word Space Models and Measures

Lubomír Krčmář[1,2]
[1]University of West Bohemia, Faculty of Applied Sciences, NTIS – New Technologies for the Information Society, Pilsen, Czech Republic
lkrcmar@kiv.zcu.cz

Karel Ježek[2]
[2]University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, Pilsen, Czech Republic
jezek_ka@kiv.zcu.cz

Pavel Pecina[3]
[3]Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic
pecina@ufal.mff.cuni.cz

Abstract

This paper presents a comparative study of 5 different types of Word Space Models (WSMs) combined with 4 different compositionality measures applied to the task of automatically determining semantic compositionality of word expressions. Many combinations of WSMs and measures have never been applied to the task before.

The study follows Biemann and Giesbrecht (2011) who attempted to find a list of expressions for which the compositionality assumption – the meaning of an expression is determined by the meaning of its constituents and their combination – does not hold. Our results are very promising and can be appreciated by those interested in WSMs, compositionality, and/or relevant evaluation methods.

1 Introduction

Our understanding of WSM is in agreement with Sahlgren (2006): "The word space model is a computational model of word meaning that utilizes the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity". There are many types of WSMs built by different algorithms. WSMs are based on the Harris distributional hypothesis (Harris, 1954), which assumes that words are similar to the extent to which they share similar linguistic contexts. WSM can be viewed as a set of words associated with vectors representing contexts in which the words occur. Then, similar vectors imply (semantic) similarity of the words and vice versa. Consequently, WSMs provide a means to find words semantically similar to a given word. This capability of WSMs is exploited by many Natural Language Processing (NLP) applications as listed e.g. by Turney and Pantel (2010).

This study follows Biemann and Giesbrecht (2011), who attempted to find a list of non-compositional expressions whose meaning is not fully determined by the meaning of its constituents and their combination. The task turned out to be frustratingly hard (Johannsen et al., 2011). Biemann's idea and motivation is that non-compositional expressions could be treated as single units in many NLP applications such as Information Retrieval (Acosta et al., 2011) or Machine Translation (Carpuat and Diab, 2010). We extend this motivation by stating that WSMs could also benefit from a set of non-compositional expressions. Specifically, WSMs could treat semantically non-compositional expressions as single units. As an example, consider "kick the bucket", "hot dog", or "zebra crossing". Treating such expressions as single units might improve the quality of WSMs since the neighboring words of these expressions should not be related to their constituents ("kick", "bucket", "dog" or "zebra"), but instead to the whole expressions.

Recent works, including that of Lin (1999), Baldwin et al. (2003), Biemann and Giesbrecht (2011), Johannsen et al. (2011), Reddy et al. (2011a), Krčmář et al. (2012), and Krčmář et al. (2013), show the applicability of WSMs in determining the compositionality of word expressions. The proposed methods exploit various types of WSMs combined with various measures for determining the compositionality applied to various datasets. First, this leads to non-directly comparable results and second, many combinations of

Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, pages 64–73,
Sofia, Bulgaria, August 9 2013.

WSMs and measures have never before been applied to the task. The main contribution and novelty of our study lies in systematic research of several basic and also advanced WSMs combined with all the so far, to the best of our knowledge, proposed WSM-based measures for determining the semantic compositionality.

The explored WSMs, described in more detail in Section 2, include the Vector Space Model, Latent Semantic Analysis, Hyperspace Analogue to Language, Correlated Occurrence Analogue to Lexical Semantics, and Random Indexing. The measures, including substitutability, endocentricity, compositionality, and neighbors-in-common-based, are described in detail in Section 3. Section 4 describes our experiments performed on the manually annotated datasets – Distributional Semantics and Compositionality dataset (DISCO) and the dataset built by Reddy et al. (2011a). Section 5 summarizes the results and Section 6 concludes the paper.

2 Word Space Models

The simplest and oldest types of WSMs[1] are the Vector Space Model (VSM) and Hyperspace Analogue to Language (HAL). More recent and advanced models include Latent Semantic Analysis (LSA), which is based on VSM, and Correlated Occurrence Analogue to Lexical Semantics (COALS), which originates from HAL. Random Indexing (RI) is WSM joining the principles of LSA and HAL. Many other WSMs have been proposed too. Their description is outside the scope of this paper and can be found e.g. in Turney and Pantel (2010) or Jurgens and Stevens (2010).

VSM is based on the assumption that similar (related) words tend to occur in the same documents.[2] VSM stores occurrence counts of all word types in documents a given corpus in a co-occurrence matrix $\mathbf{C}$. The row vectors of the matrix correspond to the word types and the columns to the documents in the corpus. The numbers of occurrences c_{ij} in $\mathbf{C}$ are usually weighted by the product of the local and global weighting functions (Nakov et al., 2001). The local function weights c_{ij} by the same mathematical function; typically none (further denoted as no), $\log(c_{ij}+1)$ (denoted as log) or $\sqrt{c_{ij}}$ (denoted as $sqrt$). The purpose of local weighting is to lower the importance of highly occurring words in the document. The global function weights every value in row i of $\mathbf{C}$ by the same value calculated for row i. Typically: none (denoted as No), Inverse Document Frequency (denoted as Idf) or a function referred to as Entropy (Ent). Idf is calculated as $1+\log(ndocs/df(i))$ and Ent as $1+\{\sum_j p(i,j)\log p(i,j)\}/\log ndocs$, where $ndocs$ is the number of documents in the corpora, $df(i)$ is the number of documents containing word type i, and $p(i,j)$ is the probability of occurrence of word type i in document j.

LSA builds on VSM and was introduced by Landauer and Dumais (1997). The LSA algorithm works with the same co-occurrence matrix $\mathbf{C}$ which can be weighted in the same manner as in VSM. The matrix is than transformed by Singular Value Decomposition (SVD) (Deerwester et al., 1990) into $\mathbf{C}$. The purpose of SVD is to project the row vectors and column vectors of $\mathbf{C}$ into a lower-dimensional space and thus bring the vectors of word types and vectors of documents, respectively, with similar meanings near to each other.[3] The output number of dimensions is a parameter of SVD and typically ranges from 200 to 1000 (Landauer and Dumais, 1997; Rohde et al., 2005).

HAL was first explored by Lund and Burgess (1996). It differs from VSM and LSA in that it only exploits neighboring words as contexts for word types. HAL processes the corpus by moving a sliding double-sided window with a size ranging from 1 to 5 around the word type in focus and accumulating the weighted co-occurrences of the preceding and following words into a matrix. Typically, the linear weighting function is used to ensure that the occurrences of words which are closer to the word type in focus are more significant. The dimensions of the resulting co-occurrence matrix are of size $|V|$ and $2|V|$, where V denotes the vocabulary consisting of all the word types occurring in the processed corpora. Finally, the HAL co-occurrence matrix can be reduced by retaining the most informative columns only. The columns with the highest values of entropy ($-\sum_j p_j \log p_j$, where p_j denotes the prob-

[1] WSMs are also referred to as distributional models of semantics, vector space models, or semantic spaces.

[2] VSM was originally developed for the SMART information retrieval system (Salton, 1971).

[3] In this way, LSA is able to capture higher-order co-occurrences.

ability of a word in the investigated column j) can be considered as the most informative. The alternatives and their description can be found e.g. in Song et al. (2004).

COALS was introduced by Rohde et al. (2005). Compared to HAL, COALS also processes a corpus by using a sliding window and linear weighting, but differs in several aspects: the window size of COALS is 4 and this value is fixed; COALS does not distinguish between the preceding and following words and treats them equally; applying COALS supposes that all but the most frequent m columns reflecting the most common open-class words are discarded; COALS transforms weighted counts in the co-occurrence matrix in a special way (all the word pair correlations are calculated, negative values are set to 0, and non-negative ones are square rooted – $corr$); and optionally, Singular Value Decomposition (Deerwester et al., 1990) can be applied to the COALS co-occurrence matrix.

RI is described in Sahlgren (2005) and can be viewed as a mixture of HAL and LSA. First, RI assigns random vectors to each word type in the corpus. The random vectors, referred to as index vectors, are very sparse, typically with a length of thousands, and contain only several (e.g. 7) non-zero values from the {-1,1} set. Second, RI processes the corpus by exploiting a sliding window like HAL and COALS. However, RI does not accumulate the weighted co-occurrence counts of neighboring words to the vector of the word type in focus. Instead, RI accumulates the index vectors of the co-occurring words. For accounting the word order, the permutation variant of RI was also developed (Sahlgren et al., 2008). This variant permutes the index vectors of neighboring words of the word type in focus according to the word order.

3 Compositionality Measures

We experimented with four basically different compositionality measures (further referred to as Measures) (Krčmář et al., 2013). Each Measure employs a function to measure similarity of WSM vectors. We experimented with the following ones: cosine (*cos*), Euclidian (inverse to Euclidian distance) (*euc*), and Pearson correlation (*cor*). The mathematical formulas are presented below.

$$cos(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2 \sum_{i=1}^{n}(b_i)^2}}$$

$$euc(\mathbf{a},\mathbf{b}) = \frac{1}{1+\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}}$$

$$cor(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})^2 \sum_{i=1}^{n}(b_i-\bar{b})^2}}$$

$$\text{where } \bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}, \quad \bar{b} = \frac{\sum_{i=1}^{n} b_i}{n}$$

SU The substitutability-based Measure is based on the fact that the replacement of non-compositional expressions' constituents by the words similar to them leads to anti-collocations (Pearce, 2002). The compositionality of expressions is calculated as the ratio between the number of occurrences of the expression in a corpora and the sum of occurrences of its alternatives – possibly anti-collocations. In a similar way, we can compare pointwise mutual information scores (Lin, 1999). As an example, consider the possible occurrences of "hot dog" and "warm dog" in the corpora.

Formally, adopted from Krčmář et al. (2012), we calculate the compositionality score c_{su} for an examined expression as follows:

$$c_{su} = \frac{\sum_{i=1}^{H} W\langle a_i^h, m\rangle * \sum_{j=1}^{M} W\langle h, a_j^m\rangle}{W\langle h, m\rangle},$$

where $\langle h, m\rangle$ denotes the number of corpora occurrences of the examined expression consisting of a head and a modifying word, a_i^h and a_j^m denote i-th and j-th most similar word[4] in a certain WSM to the head and modifying word of the expression, respectively. W stands for a weighting function; following Krčmář et al. (2012), we experimented with no (no) and logarithm (log) weighting. The $*$ symbol stands for one of the two operators: addition (*plus*) and multiplication (*mult*).

EN The endocentricity-based Measure, also referred to as component or constituent-based, compares the WSM vectors of the examined expressions and their constituents. The vectors expected to be different from each other are e.g. the vector representing the expression "hot dog" and the vector representing the word "dog". Formally, the

[4]When exploiting POS tags, we constrained the similar words to be of the same POS category in our experiments.

compositionality score c_{en} can be calculated as follows:

$$c_{en} = f(x_h, x_m) \ ,$$

where x_h and x_m denote the similarity (*sim*) or inverse rank distance (*–dist*) between the examined expression and its head and modifying constituent, respectively, with regards to a certain WSM. Function f stands for a combination of its parameters: $0.5x_h + 0.5x_m$ (*avg*), $0x_h + 1x_m$ (*mOnly*), $1x_h + 0x_m$ (*hOnly*), $\min(x_h, x_m)$ (*min*), and $\max(x_h, x_m)$ (*max*).

CO The compositionality-based Measure compares the true co-occurrence vector of the examined expression and the vector obtained from the vectors corresponding to the constituents of the expression using some compositionality function (Reddy et al., 2011a). Commonly used compositionality functions are vector addition ($\oplus$) and pointwise vector multiplication ($\otimes$) (Mitchell and Lapata, 2008). The vectors expected to be different from each other are e.g. "hot dog" and "hot"$\oplus$"dog". Formally,

$$c_{co} = s(v_e, v_h * v_m) \ ,$$

where v_e, v_h, and v_m stand for vectors of an examined expression, its head and modifying constituents, respectively. $*$ stands for a vector operation.

NE The neighbors-in-common-based Measure is based on overlap of the most similar words to the examined expression and to its constituents (McCarthy et al., 2003). As an example, consider that "hot dog" is similar to "food" or "chips" and "dog" is similar to "cat" or "bark". On the other hand, the list of neighbors of a semantically compositional expression such as "black dog" is supposed to overlap with at least one of the lists of neighbors of both the expression constituents. Formally,

$$c_{ne} = o_N^h + o_N^m \ ,$$

where o_N^h and o_N^m stand for the number of same words occurring in the list of the most similar words to the examined expression and to its head and modifying constituent, respectively.

4 Experiments

We evaluated the ability of various combinations of WSMs and Measures to rank expressions as the human annotators had done ahead of time.

Datasets We experimented with the DISCO (Biemann and Giesbrecht, 2011) and Reddy (Reddy et al., 2011a) human annotated datasets, built for the task of automatic determining of semantic compositionality. The DISCO and Reddy datasets consist of manually scored expressions of adjective-noun (AN), verb-object (VO), and subject-verb (SV) types and the noun-noun (NN) type, respectively. The DISCO dataset consists of 349 expressions divided into training, validation, and test data (TestD); the Reddy dataset consists of one set containing 90 expressions. Since the DISCO validation data are of low size (35), we concatenated them with the training data (TrValD). To TrValD and TestD we added the Reddy dataset, which we had divided stratifically ahead of time. Numbers of expressions of all the different types are summarized in Table 1.

dataset	AN-VO-SV	AN	VO	SV	NN
TrValD	175	68	68	39	45
TestD	174	77	62	35	45

Table 1: Numbers of expressions of all the different types from the DISCO and Reddy datasets.

WSM construction Since the DISCO and Reddy data were extracted from the ukWaC corpus (Baroni et al., 2009), we also build our WSMs from the same corpus. We use our own modification of the S-Space package (Jurgens and Stevens, 2010). The modification lies in treating multiword expressions and handling stopwords. Specifically, we extended the package with the capability of building WSM vectors for the examined expressions in such a way that the WSM vectors previously built for words are preserved. This differentiates our approach e.g. from Baldwin et al. (2003), who label the expressions in the corpus ahead of time and treat them as single words.[5] As for treating stopwords, we map trigrams containing determiners as the middle word into bigrams without the determiners. The intuition is to extract better co-occurrence statistics for VO expressions often containing an intervening determiner. As an example, compare the occurrences of "reinvent (de-

[5]Since many single word occurrences disappear, the WSM vectors for words change. The more expressions are treated as single words, the more WSM changes. Consequently, we believe that this approach cannot be used for building a list of all expressions occurring in an examined corpus ordered by their compositionality score.

terminer) wheel" and "reinvent wheel" in ukWaC being 623 and 27, respectively.

We experimented with lemmas (*noT*) or with lemmas concatenated with their part of speech (POS) tags (*yesT*). We labeled the following strings in ukWaC as stopwords: low-frequency words (lemmas with frequency < 50), strings containing two adjacent non-letter characters (thus omitting sequences of various symbols), and closed-class words.

For our experiments, we built WSMs using various parameters examined in previous works (see Section 2) and parameters which are implied from our own experience with WSMs. Figure 1 summarizes all the parameters we used for building WSMs.

Measure settings We examined various Measure settings (see Section 3), summarized in Table 2. For all the vector comparisons, we used the *cos* similarity. Only for HAL we also examined *euc* and for COALS *cor*, since these are the recommended similarity functions for these particular WSMs (Lund and Burgess, 1996; Rohde et al., 2005).

Met.	par.	possible values
all	sim.	*cos*, *euc* if HAL, *cor* if COALS
SU	H	0,1,...,20,30,...,100
SU	M	0,1,...,20,30,...,100
SU	W	*no*, *log*
SU	$*$	*plus*, *mult*
EN	x	*sim*, *–dist*
EN	f	*avg*, *mOnly*, *hOnly*, *min*, *max*
CO	$*$	$\oplus$, $\otimes$
NE	N	10,20,...,50,100,200,...,500,1000

Table 2: All the parameters of Measures for determining semantic compositionality described in Section 3 used in our experiments.

Experimental setup Following Biemann and Giesbrecht (2011), Reddy et al. (2011a), Krčmář et al. (2012), and Krčmář et al. (2013), we use the Spearman correlation (ρ) for the evaluation of all the combinations of WSMs and Measures (Setups). Since the distribution of scores assigned to Reddy's NN dataset might not have corresponded to the distribution of DISCO scores, we decided not to map them to the same scale. Thus, we do not create a single list consisting of all the examined expressions. Instead, we order our Setups according to the weighted average of Spearman correlations calculated across all the expression types. The weights are directly proportional to the frequencies of the particular expression types. Thus, the Setup score (*wAvg*) is calculated as follows:

$$wAvg = \frac{|AN|\rho_{AN} + |VO|\rho_{VO} + |SV|\rho_{SV} + |NN|\rho_{NN}}{|AN| + |VO| + |SV| + |NN|}.$$

Having the evaluation testbed, we tried to find the optimal parameter settings for all WSMs combined with all Measures with the help of TrValD. Then, we applied the found Setups to TestD.

Notes Because several expressions or their constituents concatenated with their POS tags did not occur sufficiently often (for expressions: ≥ 0, for constituents: ≥ 50) in ukWaC, we removed them from the experiments; we removed "number crunching", "pecking order", and "sacred cow" from TrValD and "leading edge", "broken link", "spinning jenny", and "sitting duck" from TestD.

5 Results

The Setups achieving the highest *wAvg* when applied to TrValD are depicted in Table 3. The same Setups and their results when applied to TestD are depicted in Table 4. The values of Spearman correlations in TestD confirm many of the observations from TrValD[6]:

Almost all the combinations of WSMs and Measures achieve correlation values which are statistically significant. This is best illustrated by the $\rho(AN-VO-SV)$ column in Table 4, where a lot of correlation values are statistically ($p<0.05$) or highly statistically ($p<0.001$) significant, with regards to the number of expressions (172).

The results suggest that for every expression type, the task of determining compositionality is of varying difficulty. While determining the compositionality of the NN expression type seems to be the simplest (the highest correlations observed), determining the compositionality of the SV expression type seems to be hard since the majority of values in the ρSV column are not statistically significant; taking into account the number of SV expressions in TestD – 35, the statistically significant value of ρ at the $p<0.05$ level is 0.34.

The correlation values differ with regards to the expression type. Certain WSMs combined with

[6]A test of statistical difference between two values of the Spearman correlation is adopted from Papoulis (1990).

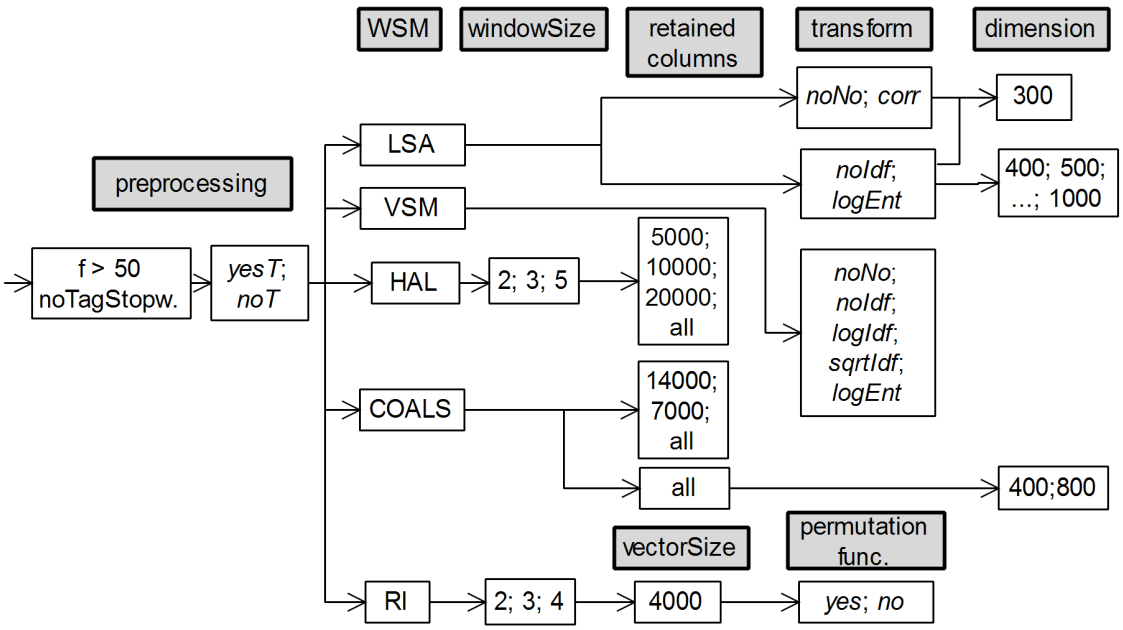

Figure 1: All the parameters of WSMs described in Section 2 used in all our experiments. Semicolon denotes OR. All the examined combinations of parameters are implied from reading the diagram from left to right.

certain Measures, although achieving high correlations upon certain expression types, fail to correlate with the rest of the expression types. Compare e.g. the correlation values of VSM and LSA combined with the SU Measure upon the AN and SV types with the correlation values upon the VO and NN types.

The results, as expected, illustrate that employing more advanced alternatives of basic WSMs is more appropriate. Specifically, LSA outperforms VSM and COALS outperforms HAL in 21 and 23 correlation values out of 24, respectively. Concerning RI, the values of correlations seem to be close to the values of VSM and HAL.

An interesting observation showing the appropriateness of using $wAvg(of\rho)$ as a good evaluation score is supported by a comparison of the $wAvg(of\rho)$ and $\rho(AN{-}VO{-}SV)$ columns. The columns suggest that some Setups might only be able to order the expressions of the same type and might not be able to order the expressions of different types among each other. As an example, compare the value of $\rho = 0.42$ in $wAvg(of\rho)$ with $\rho = 0.28$ in $\rho(AN{-}VO{-}SV)$ in the row corresponding to COALS combined with SU. Consider also that all the values of correlations are higher or equal to the value in $\rho(AN{-}VO{-}SV)$.

As for the parameters learned from applying all the combinations of differently set WSM algorithms and Measures to TrValD, their diversity is well illustrated in Tables 5 and 6. Due to this diversity, we cannot recommend any particular settings except for one. All our SU Measures benefit from weighting numbers of expression occurrences by logarithm.

The correlation values in TestD are slightly lower – probably due to overfitting – than the ones observed in TrValD. HAL combined with the Measures using euc similarity was not as successful as when combined with cos.[7]

For comparison, the results of Reddy et al. (2011b) and Chakraborty et al. (2011) as the results of the best performing Setups based on WSMs and association measures, respectively, applied to the DISCO data, are presented (Biemann and Giesbrecht, 2011). The correlation values of our Setups based on LSA and COALS, respectively, are mostly higher. However, the improvements are not statistically significant. Also, the recent results achieved by Krčmář et al. (2012) employing COALS and Krčmář et al. (2013) employ-

[7]However, using HAL combined with euc, we observed significant negative correlations which deserve further exploration.

ing LSA are depicted.

Discussion As described above, we observed different values of correlations for different expression types. This motivates us to think about other classes of expressions different from types; Measures could be e.g. varyingly successful with regards to different occurrence frequency classes of expressions (Evert, 2005). However, with such small datasets, as shown e.g. by the fact that the majority of our results are statistically indistinguishable, we cannot carry out any deeper investigations. A large dataset would provide a more reliable comparison. Ideally, this would consist of all the candidate expressions occurring in some smaller corpus. Also, we would prefer the annotated dataset not to be biased towards non-compositional expressions and to be provided with an inner-annotator agreement (Pecina, 2008); which is unfortunately not the case of the DISCO dataset.

6 Conclusion

Our study suggests that different WSMs combined with different Measures perform reasonably well in the task of determining the semantic compositionality of word expressions of different types. Especially, LSA and COALS perform well in our experiments since their results are better than those of their basic variants (VSM and HAL, respectively) and, although not statistically significantly, they outperform the best results of the previously proposed approaches (Table 4).

Importantly, our results demonstrate (Section 5) that the datasets used for the experiments are small for: first, a statistical learning of optimal parameters of both WSM algorithms and Measures; second, a thorough (different types) and reliable (statistically significant) comparison of our and the previously proposed approaches.

Therefore, we plan to build a larger manually-annotated dataset. Finally, we plan to extract a list of semantically non-compositional expressions from a given corpus and experiment with using it in NLP applications.

Acknowledgments

We thank to Vít Suchomel for providing the ukWaC corpus and the anonymous reviewers for their helpful comments and suggestions. This work was supported by the European Regional Development Fund (ERDF), project NTIS – New Technologies for the Information Society, European Centre of Excellence, CZ.1.05/1.1.00/02.0090; by Advanced Computing and Information Systems (grant no. SGS-2013-029); and by the Czech Science Foundation (grant no. P103/12/G084). Also, the access to the CERIT-SC computing facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ.1.05/3.2.00/08.0144 is highly appreciated.

References

Otavio Costa Acosta, Aline Villavicencio, and Viviane P. Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 101–109, Stroudsburg, PA, USA.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. *Proceedings of the ACL 2003 workshop on Multiword expressions analysis acquisition and treatment*, pages 89–96.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources And Evaluation*, 43(3):209–226.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 21–28.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 242–245, Stroudsburg, PA, USA.

Tanmoy Chakraborty, Santanu Pal, Tapabrata Mondal, Tanik Saikh, and Sivaju Bandyopadhyay. 2011. Shared task system description: Measuring the compositionality of bigrams using statistical methodologies. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 38–42, Portland, Oregon, USA.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman.

WSM	Measure	wAvg(of ρ)	ρAN-VO-SV	ρAN	ρVO	ρSV	ρNN
VSM_1	SU_1	0.31	0.11	-0.03	0.36	**0.31**	**0.75**
VSM_2	EN_1	0.36	0.32	**0.41**	0.30	0.10	**0.61**
VSM_3	CO_1	0.40	0.34	**0.40**	0.26	**0.39**	**0.64**
VSM_1	NE_1	0.34	0.26	0.20	**0.48**	0.07	**0.60**
LSA_1	SU_2	0.34	0.19	-0.05	**0.46**	**0.42**	**0.71**
LSA_2	EN_2	0.56	**0.53**	**0.54**	**0.51**	**0.59**	**0.65**
LSA_3	CO_1	0.55	**0.53**	**0.49**	**0.56**	**0.63**	**0.58**
LSA_2	NE_2	0.50	**0.45**	**0.46**	0.37	**0.64**	**0.62**
HAL_1	SU_3	0.45	0.36	0.28	**0.50**	**0.40**	**0.67**
HAL_2	EN_3	0.36	0.35	**0.47**	0.28	0.27	0.38
HAL_3	CO_1	0.23	0.15	0.28	0.12	-0.01	**0.54**
HAL_4	NE_3	0.27	0.25	0.31	0.21	0.17	0.39
$COALS_1$	SU_4	0.48	**0.41**	0.28	**0.56**	**0.49**	**0.68**
$COALS_2$	EN_2	0.58	**0.54**	**0.6**	**0.63**	**0.37**	**0.68**
$COALS_2$	CO_1	0.59	**0.54**	**0.6**	**0.64**	**0.37**	**0.70**
$COALS_2$	NE_4	0.58	**0.56**	**0.61**	**0.58**	**0.46**	**0.67**
RI_1	SU_5	0.52	**0.44**	**0.45**	**0.51**	**0.52**	**0.68**
RI_2	EN_3	0.45	**0.44**	**0.41**	**0.57**	**0.33**	0.45
RI_3	CO_1	0.21	0.13	0.13	0.16	0.11	**0.54**
RI_2	NE_5	0.43	**0.43**	**0.43**	**0.53**	0.21	**0.49**

Table 3: The Spearman correlations ρ of the best performing (wAvg) combinations of particular WSMs and Measures from all the tested Setups applied to TrValD. The highest correlation values in the particular columns and the correlation values which are not statistically different from them ($p < 0.05$) are in bold (yet we do not know how to calculate the stat. significance for the wAvg(of ρ) column). The parameters of WSMs and Measures corresponding to the indexes are depicted in Tables 5 and 6, respectively.

1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Stefan Evert. 2005. *The statistics of word cooccurrences : word pairs and collocations*. Ph.D. thesis, Universität Stuttgart, Holzgartenstr. 16, 70174 Stuttgart.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Anders Johannsen, Hector Martinez Alonso, Christian Rishøj, and Anders Søgaard. 2011. Shared task system description: frustratingly hard compositionality prediction. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 29–32, Stroudsburg, PA, USA.

David Jurgens and Keith Stevens. 2010. The s-space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 30–35, Stroudsburg, PA, USA.

Lubomír Krčmář, Karel Ježek, and Massimo Poesio. 2012. Detection of semantic compositionality using semantic spaces. *Lecture Notes in Computer Science*, 7499 LNAI:353–361.

Lubomír Krčmář, Karel Ježek, and Pavel Pecina. 2013. Determining compositionality of word expressions using word space models. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 42–50, Atlanta, Georgia, USA.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 317–324, Stroudsburg, PA, USA.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions analysis acquisition and treatment*, volume 18 of *MWE '03*, pages 73–80.

WSM	Measure	wAvg(of ρ)	ρAN-VO-SV	ρAN	ρVO	ρSV	ρNN
VSM_1	SU_1	0.28	0.03	0.01	**0.51**	**0.04**	**0.62**
VSM_2	EN_1	0.26	0.19	0.08	0.29	**0.04**	**0.69**
VSM_3	CO_1	0.32	0.26	0.24	0.23	**0.25**	**0.65**
VSM_1	NE_1	0.32	0.19	**0.36**	0.25	-0.13	**0.73**
LSA_1	SU_2	0.31	0.06	0.05	**0.50**	**0.20**	**0.59**
LSA_2	EN_2	0.50	**0.40**	**0.39**	**0.55**	**0.32**	**0.78**
LSA_3	CO_1	0.48	**0.36**	**0.29**	**0.60**	**0.42**	**0.69**
LSA_2	NE_2	0.44	**0.33**	**0.34**	0.40	**0.44**	**0.67**
HAL_1	SU_3	0.29	0.16	0.09	0.32	**0.34**	**0.56**
HAL_2	EN_3	0.36	0.28	**0.33**	0.35	**0.26**	**0.53**
HAL_3	CO_1	0.24	0.22	0.25	0.16	**0.15**	0.42
HAL_4	NE_3	0.21	0.14	0.02	0.33	**0.06**	0.47
$COALS_1$	SU_4	0.42	0.28	0.28	**0.54**	**0.30**	**0.59**
$COALS_2$	EN_2	0.49	**0.44**	**0.52**	**0.51**	**0.07**	**0.72**
$COALS_2$	CO_1	0.47	**0.40**	**0.47**	**0.51**	**0.07**	**0.74**
$COALS_2$	NE_4	0.52	**0.48**	**0.55**	**0.50**	**0.21**	**0.74**
RI_1	SU_5	0.30	0.14	0.14	0.29	**0.12**	**0.72**
RI_2	EN_3	0.44	**0.34**	**0.37**	**0.54**	**0.20**	**0.63**
RI_3	CO_1	0.23	0.23	**0.29**	0.17	**0.17**	0.26
RI_2	NE_5	0.31	0.26	0.26	0.42	**0.04**	0.44
Reddy-WSM		-	**0.35**	-	-	-	-
StatMix		-	**0.33**	-	-	-	-
Krcmar-COALS		-	**0.42**	**0.42**	**0.69**	**0.24**	-
Krcmar-LSA		-	**0.50**	**0.50**	**0.56**	**0.41**	-

Table 4: The Spearman correlations ρ of the best performing (wAvg) combinations of particular WSMs and Measures trained in TranValD applied to TestD. The highest correlation values in the particular columns and the correlation values which are not statistically different from them ($p < 0.05$) are in bold (yet we do not know how to calculate the stat. significance for the wAvg(of ρ) column). Reddy-WSM and StatMix stand for the best performing system based on WSMs and association measures, respectively, applied to the DISCO task (Biemann and Giesbrecht, 2011). Krcmar-COALS and Krcmar-LSA stand for the best published results achieved upon the dataset presented in Krčmář et al. (2012) and Krčmář et al. (2013), respectively. The parameters of WSMs and Measures corresponding to the indexes are depicted in Tables 5 and 6, respectively.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio.

Preslav Nakov, Antonia Popova, and Plamen Mateev. 2001. Weight functions impact on lsa performance. In *Proceedings of the EuroConference Recent Advances in Natural Language Processing (RANLP'01)*, pages 187–193.

Athanasios Papoulis. 1990. *Probability & statistics*. Prentice Hall.

Darren Pearce. 2002. A Comparative Evaluation of Collocation Extraction Techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC*.

Pavel Pecina. 2008. Reference data for Czech collocation extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, pages 11–14, Marrakech, Morocco. European Language Resources Association.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011a. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Siva Reddy, Diana McCarthy, Suresh Manandhar, and Spandana Gella. 2011b. Exemplar-based word-space model for compositionality detection: Shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 54–60, Portland, Oregon, USA.

WSM	parameters			
VSM	**tags**	**trans.**		
VSM_1	*noT*	*noNo*		
VSM_2	*yesT*	*noNo*		
VSM_3	*yesT*	*noIdf*		
LSA	**tags**	**trans.**	**dim.**	
LSA_1	*noT*	*logEnt*	900	
LSA_2	*yesT*	*noNo*	300	
LSA_3	*noT*	*noIdf*	300	
HAL	**tags**	**win s.**	**ret. c.**	
HAL_1	*noT*	5	20000	
HAL_2	*yesT*	5	20000	
HAL_3	*noT*	2	10000	
HAL_4	*yesT*	5	all	
COALS	**tags**	**ret. c.**		
$COALS_1$	*noT*	7000		
$COALS_2$	*yesT*	7000		
RI	**tags**	**win. s.**	**vec. s.**	**perm.**
RI_1	*noT*	2	4000	*no*
RI_2	*noT*	4	4000	*no*
RI_3	*noT*	2	4000	*yes*

Table 5: Parameters of WSMs (Section 2) which, combined with particular Measures, achieved the highest average correlation in TrValD.

Measure	parameters				
SU	**sim.**	$*$	**W**	**H**	**M**
SU_1	*cos*	*plus*	*log*	30	3
SU_2	*cos*	*plus*	*log*	100	5
SU_3	*cos*	*mult*	*log*	12	2
SU_4	*cos*	*mult*	*log*	80	4
SU_5	*cos*	*mult*	*log*	4	3
EN	**sim.**	**func.**	**x**		
EN_1	*cos*	*min*	*sim*		
EN_2	*cos*	*avg*	*sim*		
EN_3	*cos*	*min*	*–dist*		
CO	**sim.**	$*$			
CO_1	*cos*	$\oplus$			
NE	**sim.**	**O**			
NE_1	*cos*	1000			
NE_2	*cos*	500			
NE_3	*cos*	50			
NE_4	*cor*	500			
NE_5	*cos*	20			

Table 6: Parameters of Measures (Section 3) which, combined with particular WSMs, achieved the highest average correlation in TrValD.

Douglas L. Rohde, Laura M. Gonnerman, and David C. Plaut. 2005. An improved model of semantic similarity based on lexical co-occurrence. *Unpublished manuscript*.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In V. Sloutsky, B. Love, and K. Mcrae, editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1300–1305. Cognitive Science Society, Austin, TX.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, Leipzig, Germany.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Gerard Salton. 1971. *The SMART Retrieval System; Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Dawei Song, Peter Bruza, and Richard Cole. 2004. Concept learning and information inferencing on a highdimensional semantic space. In *ACM SIGIR 2004 Workshop on Mathematical/Formal Methods in Information Retrieval*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188.