Determining Compositionality of Word Expressions Using Word Space Models

Lubomír Krčmář, Karel Ježek
University of West Bohemia
Faculty of Applied Sciences
Department of Computer Science and Engineering
Pilsen, Czech Republic
{lkrcmar,jezek_ka}@kiv.zcu.cz

Pavel Pecina
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 research focuses on determining semantic compositionality of word expressions using word space models (WSMs). We discuss previous works employing WSMs and present differences in the proposed approaches which include types of WSMs, corpora, preprocessing techniques, methods for determining compositionality, and evaluation testbeds.

We also present results of our own approach for determining the semantic compositionality based on comparing distributional vectors of expressions and their components. The vectors were obtained by Latent Semantic Analysis (LSA) applied to the ukWaC corpus. Our results outperform those of all the participants in the Distributional Semantics and Compositionality (DISCO) 2011 shared task.

1 Introduction

A word expression is semantically compositional if its meaning can be understood from the literal meaning of its components. Therefore, semantically compositional expressions involve e.g. "small island" or "hot water"; on the other hand, semantically non-compositional expressions are e.g. "red tape" or "kick the bucket".

The notion of compositionality is closely related to idiomacy – the higher the compositionality the lower the idiomacy and vice versa (Sag et al., 2002; Baldwin and Kim, 2010).

Non-compositional expressions are often referred to as Multiword Expressions (MWEs). Baldwin and Kim (2010) differentiate the following sub-types of compositionality: lexical, syntactic, semantic, pragmatic, and statistical. This paper is concerned with semantic compositionality.

Compositionality as a feature of word expressions is not discrete. Instead, expressions populate a continuum between two extremes: idioms and free word combinations (McCarthy et al., 2003; Bannard et al., 2003; Katz, 2006; Fazly, 2007; Baldwin and Kim, 2010; Biemann and Giesbrecht, 2011). Typical examples of expressions between the two extremes are "zebra crossing" or "blind alley".

Our research in compositionality is motivated by the hypothesis that a special treatment of semantically non-compositional expressions can improve results in various Natural Language Processing (NPL) tasks, as shown for example by Acosta et al. (2011), who utilized MWEs in Information Retrieval (IR). Besides that, there are other NLP applications that can benefit from knowing the degree of compositionality of expressions such as machine translation (Carpuat and Diab, 2010), lexicography (Church and Hanks, 1990), word sense disambiguation (Finlayson and Kulkarni, 2011), part-of-speech (POS) tagging and parsing (Seretan, 2008) as listed in Ramisch (2012).

The main goal of this paper is to present an analysis of previous approaches using WSMs for determining the semantic compositionality of expressions. The analysis can be found in Section 2. A special attention is paid to the evaluation of the proposed models that is described in Section 3. Section 4 presents our first intuitive experimental setup and results of LSA applied to the DISCO 2011 task. Section 5 concludes the paper.

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 42–50,
Atlanta, Georgia, 13-14 June 2013.

2 Semantic Compositionality of Word Expressions Determined by WSMs

Several recent works, including Lin (1999), Schone and Jurafsky (2001), Baldwin et al. (2003), McCarthy et al. (2003), Katz (2006), Johannsen et al. (2011), Reddy et al. (2011a), and Krčmář et al. (2012), show the ability of methods based on WSMs to capture the degree of semantic compositionality of word expressions. We analyse the proposed methods and discuss their differences. As further described in detail and summarized in Table 1, the approaches differ in the type of WSMs, corpora, preprocessing techniques, methods for determining the compositionality, datasets for evaluation, and methods of evaluation itself.

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". For more information on WSMs, see e.g. Turney and Pantel (2010), Jurgens and Stevens (2010), or Sahlgren (2006).

WSMs and their parameters WSMs can be built by different algorithms including LSA (Landauer and Dumais, 1997), Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996), Random Indexing (RI) (Sahlgren, 2005), and Correlated Occurrence Analogue to Lexical Semantics (COALS) (Rohde et al., 2005). Every algorithm has its own specifics and can be configured in different ways. The configuration usually involves e.g. the choice of context size, weighting functions, or normalizing functions. While Schone and Jurafsky (2001), Baldwin et al. (2003), and Katz (2006) addopted LSA-based approaches, Johannsen et al. (2011) and Krčmář et al. (2012) employ COALS; the others use their own specific WSMs.

Corpora and text preprocessing Using different corpora and their preprocessing naturally leads to different WSMs. The preprocessing can differ e.g. in the choice of used word forms or in removal/retaining of low-frequency words. For example, while Lin (1999) employs a 125-million-word newspaper corpus, Schone and Jurafsky (2001) use a 6.7-million-word subset of the TREC databases, Baldwin et al. (2003) base their experiments on 90 million words from the British National Corpus (Burnard, 2000). Krčmář et al. (2012), Johannsen et al. (2011), and Reddy et al. (2011a) use the ukWaC corpus, consisting of 1.9 billion words from web texts (Baroni et al., 2009). As for preprocessing, Lin (1999) extracts triples with dependency relationships, Baldwin et al. (2003), Reddy et al. (2011a), and Krčmář et al. (2012) concatenate word lemmas with their POS categories. Johannsen et al. (2011) use word lemmas and remove low-frequency words while Reddy et al. (2011a), for example, keep only frequent content words.

Methods We have identified three basic methods for determining semantic compositionality:
1) The substitutability-based methods exploit the fact that replacing components of non-compositional expressions by words which are similar leads to anti-collocations (Pearce, 2002). Then, frequency or mutual information of such expressions (anti-collocations) is compared with the frequency or mutual information of the original expressions. For example, consider expected occurrence counts of "hot dog" and its anti-collocations such as "warm dog" or "hot terrier".
2) The component-based methods, utilized for example by Baldwin et al. (2003) or Johannsen et al. (2011), compare the distributional characteristics of expressions and their components. The context 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".
3) The compositionality-based methods compare two vectors of each analysed expression: the true co-occurrence vector of an expression and the vector obtained from vectors corresponding to the components of the expression using a compositionality function (Reddy et al., 2011a). The most common compositionality functions are vector addition or pointwise vector multiplication (Mitchell and Lapata, 2008). For example, the vectors for "hot dog" and "hot"$\oplus$"dog" are supposed to be different.

Evaluation datasets There is still no consensus on how to evaluate models determining semantic compositionality. However, by examining the discussed papers, we have observed an increasing ten-

Paper	Corpora	WSMs	Methods	Data (types)	Evaluation
Lin (1999)	125m, triples	own	SY	NVAA c.	dicts., P/R
Schone+Jurafsky(2001)	6.7m TREC	LSA	SY, CY	all types	WN, P/Rc
Baldwin et al. (2003)	BNC+POS	LSA	CT	NN, VP	WN, PC
McCarthy et al. (2003)	BNC+GR	own	CTn	PV	MA, WN, dicts., S
Katz (2006)	GNC	LSA	CY	PNV	MA, P/R, Fm
Krčmář et al. (2012)	ukWaC+POS	COALS	SY	AN, VO, SV	MA, CR, APD, CL
Johannsen et al. (2011)	ukWaC	COALS	SY, CT	AN, VO, SV	MA, CR, APD, CL
Reddy et al. (2011a)	ukWaC+POS	own	CT, CY	NN	MA, S, R2

Table 1: Overview of experiments applying WSMs to determine semantic compositionality of word expressions. BNC - British National Corpus, GR - grammatical relations, GNC - German newspaper corpus, TREC - TREC corpus; SY - substitutability-based methods, CT - component-based methods, CTn - component-based methods comparing WSM neighbors of expressions and their components, CY - compositionality-based methods; NVAP c. - noun, verb, adjective, adverb combinations, NN - noun-noun, VP - verb-particles, AN - adjective-noun, VO - verb-object, SV - subject-verb, PV - phrasal-verb, PNV - preposition-noun-verb; dicts. - dictionaries of idioms, WN - Wordnet, MA - use of manually annotated data, S - Spearman correlation, PC - Pearson correlation, CR - Spearman and Kendall correlations, APD - average point difference, CL - classification, P/R - Precision/Recall, P/Rc - Precision/Recall curves, Fm - F measure, R2 - goodness.

dency to exploit manually annotated data from a specific corpus, ranging from semantically compositional to non-compositional expressions (McCarthy et al., 2003; Katz, 2006; Johannsen et al., 2011; Reddy et al., 2011a; Krčmář et al., 2012).

This approach, as opposed to the methods based on dictionaries of MWEs (idioms) or Wordnet (Miller, 1995), has the following advantages: Firstly, the classification of a manually annotated data is not binary but finer-grained, enabling the evaluation to be more detailed. Secondly, the low-coverage problem of dictionaries, which originates for example due to the facts that new MWEs still arise or are domain specific, is avoided.[1] For example, Lin (1999), Schone and Jurafsky (2001), Baldwin et al. (2003) used Wordnet or other dictionary-type resources.

3 Evaluation Methods

This section discusses evaluation methods including average point difference (APD), Spearman and Kendall correlations, and precision of classification (PoC) suggested by Biemann and Giesbrecht (2011); Precision/nBest, Recall/nBest and Precision/Recall curves proposed by Evert (2005); and Average Precision used by Pecina (2009). Our evaluation is based on the English part of the manually annotated datasets DISCO 2011 (Biemann and Giesbrecht, 2011), further referred to as DISCO-En-Gold.

Disco-En-Gold consists of 349 expressions divided into training (TrainD), validation (ValD), and test data (TestD) manually assigned scores from 0 to 100, indicating the level of compositionality (the lower the score the lower the compositionality and vice versa). The expressions are of the following types: adjective-noun (AN), verb-object (VO), and subject-verb (SV). Based on the numerical scores, the expressions are also classified into three disjoint classes (coarse scores): low, medium, and high compositional.[2] A sample of the Disco-En-Gold data is presented in Table 2.

Comparison of evaluation methods The purpose of the DISCO workshop was to find the best methods for determining semantic compositionality. The participants were asked to create systems capable of assigning the numerical values closest to the ones assigned by the annotators (Gold values). The proposed APD evaluation measure is calculated as the mean difference between the particular systems' val-

[1] The consequence of using a low-coverage dictionary can cause underestimation of the used method since the dictionary does not have to contain MWEs correctly found by that method.

[2] Several expressions with the numerical scores close to the specified thresholds were not classified into any class.

Type	Expression	Ns	Cs
EN_ADJ_NN	blue chip	11	low
EN_V_OBJ	buck trend	14	low
EN_ADJ_NN	open source	49	medium
EN_V_OBJ	take advantage	57	medium
EN_ADJ_NN	red squirrel	90	high
EN_V_SUBJ	student learn	98	high

Table 2: A sample of manually annotated expressions from Disco-En-Gold with their numerical scores (Ns) and coarse scores (Cs).

ues and the Gold values assigned to the same expressions. PoC is defined as the ratio of correct coarse predictions to the number of all the predictions.

Following Krčmář et al. (2012), we argue that for the purpose of comparison of the methods, the values assigned to a set of expressions by a certain model are not as important as is the ranking of the expressions (which is not sensitive to the original distribution of compositionality values). Similarly as Evert (2005), Pecina (2009), and Krčmář et al. (2012) we adopt evaluation based on ranking (although the measures such as PoC or APD might provide useful information too).

Evaluation based on ranking can be realized by measuring ranked correlations (Spearman and Kendall) or Precision/Recall scores and curves commonly used e.g. in IR (Manning et al., 2008). In IR, Precision is defined as the ratio of found relevant documents to all the retrieved documents with regards to a user's query. Recall is defined as the ratio of found relevant documents to all the relevant documents in a test set to the user's query. The Precision/Recall curve is a curve depicting the dependency of Precision upon Recall. Analogously, the scheme can be used for evaluation of the methods finding semantically non-compositional expressions. However, estimation of Recall is not possible without knowledge of the correct class[3] for every expression in a corpus. To bypass this, Evert (2005) calculates Recall with respect to the set of annotated data divided into non-compositional and compositional classes. The Precision/nBest, Recall/nBest, and Precision/Recall curves for the LSA experiment described in the following section are depicted in Figures 1 and 2.

Evert's (2005) curves allow us to visually compare the results of the methods in more detail. To facilitate comparison of several methods, we also suggest using average precision (AP) adopted from Pecina (2009), which reduces information provided by a single Precision/Recall curve to one value. AP is defined as a mean Precision at all the values of Recall different from zero.

4 LSA experiment

LSA is WSM based on the Singular Value Decomposition (SVD) factorization (Deerwester et al., 1990) applied to the co-occurrence matrix. In the matrix, the numbers of word occurrences in specified contexts[4] are stored. The row vectors of the matrix capture the word meanings.[5] The idea of using SVD is to project vectors corresponding to the words into a lower-dimensional space and thus bring the vectors of words with similar meaning near to each other.

We built LSA WSM and applied the component-based method to Disco-En-Gold. We used our own modification of the LSA algorithm originally implemented in the S-Space package (Jurgens and Stevens, 2010). The modification lies in treating expressions and handling stopwords. Specifically, we added vectors for the examined expressions to WSM in such a way that the original vectors for words were preserved. This differentiates our approach e.g. from Baldwin et al. (2003) or Johannsen et al. (2011) who label the expressions ahead of time and build WSMs treating them as single words. Treating the expressions as the single words affects the WSM vectors of their constituents. As an example, consider the replacement of occurrences of "short distance" by e.g. the EXP#123 label. This affects the WSM vectors of "short" and "distance" since the numbers of their occurrences and the numbers of contexts they occur in drops. Consequently, this also affects the methods for determining the compositionality which are based upon using the vectors of

[3] A semantically non-compositional expression or a semantically compositional expressions

[4] The commonly used contexts for words are documents or the preceding and following words in a specified window.

[5] WSMs exploit Harris' distributional hypothesis (Harris, 1954), which states that semantically similar words tend to appear in similar contexts.

expressions' constituents.

As for treating stopwords, we mapped the trigram expressions containing the determiners "the", "a", or "an" as the middle word to the corresponding bigram expressions without the determiners. The intuition is to extract more precise co-occurrence vectors for the VO expressions often containing some intervening determiner. As an example, compare the occurrences of "reinvent wheel" and "reinvent (determiner) wheel" in the ukWaC corpus which are 27 and 623, respectively, or the occurrences of "cross bridge" and "cross (determiner) bridge" being 50 and 1050, respectively.[6]

We built LSA WSM from the whole ukWaC POS-tagged corpus for all the word lemmas concatenated with their POS tags excluding stopwords. We treated the following strings as stopwords: the lemmas with frequency below 50 (omitting low-frequency words), the strings containing two adjacent non-letter characters (omitting strings such as web addresses and sequences of e.g. star symbols), and lemmas with a different POS tag from noun, proper noun, adjective, verb, and adverb (omitting closed-class words). As contexts, the entire documents were used.

The co-occurrence matrix for words was normalized by applying the log-entropy transformation and reduced to 300 dimensions. Using these settings, Landauer and Dumais (1997) obtained the best results. Finally, the co-occurrence vectors of expressions were expressed in the lower-dimensional space of words in a manner analogous to how a user's query is being expressed in lower-dimensional space of documents in IR (Berry et al., 1995). The Disco-En-Gold expressions were sorted in ascending order by the average cosine similarity between the vectors corresponding to the expressions and the vectors corresponding to their components.

Evaluation We have not tried to find the optimal parameter settings for the LSA-based model yet. Therefore, we present the results on the concatenation of TrainD with ValD giving us TrainValD and on TestD. The expressions "leading edge" and "broken link" were removed from TestD because they occur in the ukWaC corpus assigned with the required POS tags less than 50 times. APs with the Spearman and Kendall correlations between the compositionality values assigned by the LSA-based model and the Gold values are depicted in Table 3. The Spearman correlations of the LSA model applied to the whole TrainValD and TestD are highly significant with p-values < 0.001. For the AP evaluation, the expressions with numerical values less or equal to 50 were classified as non-compositional[7], giving us the ratio of non-compositional expressions in TrainValD and TestD equal to 0.26 and 0.20, respectively. The Precision/nBest and Recall/nBest graphs corresponding to the LSA-based model applied to TestD are depicted in Figure 1. The Precision/Recall graphs corresponding to the LSA-based model applied to TrainD and TestD are depicted in Figure 2.

For comparison, the graphs in Figures 1 and 2 also show the curves corresponding to the evaluation of Pointwise Mutual Information (PMI).[8] The co-occurrence statistics of the expressions in Disco-En-Gold was extracted from the window of size three, sliding through the whole lemmatized ukWaC corpus.

Discussion As suggested in Section 3, we compare the results of the methods using Spearman and Kendall correlations, AP, and Everts' curves. We present the results of the LSA and PMI models alongside the results of the best performing models participating in the DISCO task. Namely, Table 3 presents the correlation values of our models, the best performing WSM-based model (Reddy et al., 2011b), the best performing model based upon association measures (Chakraborty et al., 2011), and random baseline models.

The poor results achieved by employing PMI are similar to the results of random baselines and in accordance with those of participants of the DISCO workshop (Chakraborty et al., 2011). We hypothesize that the PMI-based model incorrectly assigns low values of semantic compositionality (high val-

[6]More precisely, the occurrences were calculated from the POS-tagged parallels of the expressions.

[7]Choice of this value can affect the results. The value of 50 was chosen since it is the middle value between the manually assigned scores ranging from 0 to 100.

[8]PMI is an association measure used to determine the strength of association between two or more words based on their occurrences and co-occurrences in a corpus (Pecina, 2009).

Model	Dataset	ρ-All	ρ-AN	ρ-VO	ρ-SV	τ-All	τ-AN	τ-VO	τ-SV	AP-All
LSA	TrainValD	0.47	0.54	0.36	0.57	0.32	0.38	0.24	0.44	0.61
PMI	TrainValD	0.02	-0.25	0.29	0.14	0.01	-0.18	0.20	0.10	0.28
baseline	TrainValD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26
LSA	TestD	0.50	0.50	0.56	0.41	0.35	0.36	0.39	0.30	0.53
Reddy-WSM	TestD	0.35	-	-	-	0.24	-	-	-	-
StatMix	TestD	0.33	-	-	-	0.23	-	-	-	-
PMI	TestD	-0.08	-0.07	0.13	-0.08	-0.06	-0.04	0.08	-0.07	0.21
baseline	TestD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20

Table 3: The values of AP, Spearman (ρ) and Kendall (τ) correlations between the LSA-based and PMI-based model respectively and the Gold data with regards to the expression type. Every zero value in the table corresponds to the theoretically achieved mean value of correlation calculated from the infinite number of correlation values between the ranking of scores assigned by the annotators and the rankings of scores being obtained by a random number genarator. Reddy-WSM stands for the best performing WSM in the DISCO task (Reddy et al., 2011b). StatMix stands for the best performing system based upon association measures (Chakraborty et al., 2011). Only ρ-All and τ-All are available for the models explored by Reddy et al. (2011b) and Chakraborty et al. (2011).

ues of PMI) to frequently occurring fixed expressions. For example, we observed that the calculated values of PMI for "international airport" and "religious belief" were high.

To the contrary, our results achieved by employing the LSA model are statistically significant and better than those of all the participants of the DISCO workshop. However, the data set is probably not large enough to provide statistically reliable comparison of the methods and it is not clear how reliable the dataset itself is (the interannotator agreement was not analyzed) and therefore we can not make any hard conclusions.

5 Conclusion

We analysed the previous works applying WSMs for determining the semantic compositionality of expressions. We discussed and summarized the majority of techniques presented in the papers. Our analysis reveals a large diversity of approaches which leads to incomparable results (Table 1). Since it has been shown that WSMs can serve as good predictors of semantic compositionality, we aim to create a comparative study of the approaches.

Our analysis implies to evaluate the proposed approaches using human annotated data and evaluation techniques based on ranking. Namely, we suggest using Spearman and Kendall correlations, Precision/nBest, Recall/nBest, Precision/Recall curves, and AP.

Using the suggested evaluation techniques, we present the results of our first experiments exploiting LSA (Figures 1, 2 and Table 3). The results of the LSA-based model, compared with random baselines, PMI-based model, and all the WSM-based and statistical-based models proposed by the participants of the DISCO task, are very promising.

Acknowledgments

We thank to Vít Suchomel for providing the ukWaC corpus and the anonymous reviewers for their helpful comments and suggestions. The research is supported 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 and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

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.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, volume 18 of *MWE '03*, pages 65–72, Stroudsburg, PA, USA.

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.

Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 37(4):573–595.

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.

Lou Burnard. 2000. User reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

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.

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

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

Afsaneh Fazly. 2007. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, University of Toronto.

Mark Alan Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 20–24, Stroudsburg, PA, USA.

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.

Graham Katz. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *In Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19.

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

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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

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.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38:39–41.

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

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. 2009. *Lexical Association Measures: Collocation Extraction*, volume 4 of *Studies in Computational and Theoretical Linguistics*. ÚFAL, Praha, Czechia.

Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, pages 61–66, Stroudsburg, PA, USA.

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.

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.

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

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, London, UK. Springer-Verlag.

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.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, University of Geneva.

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

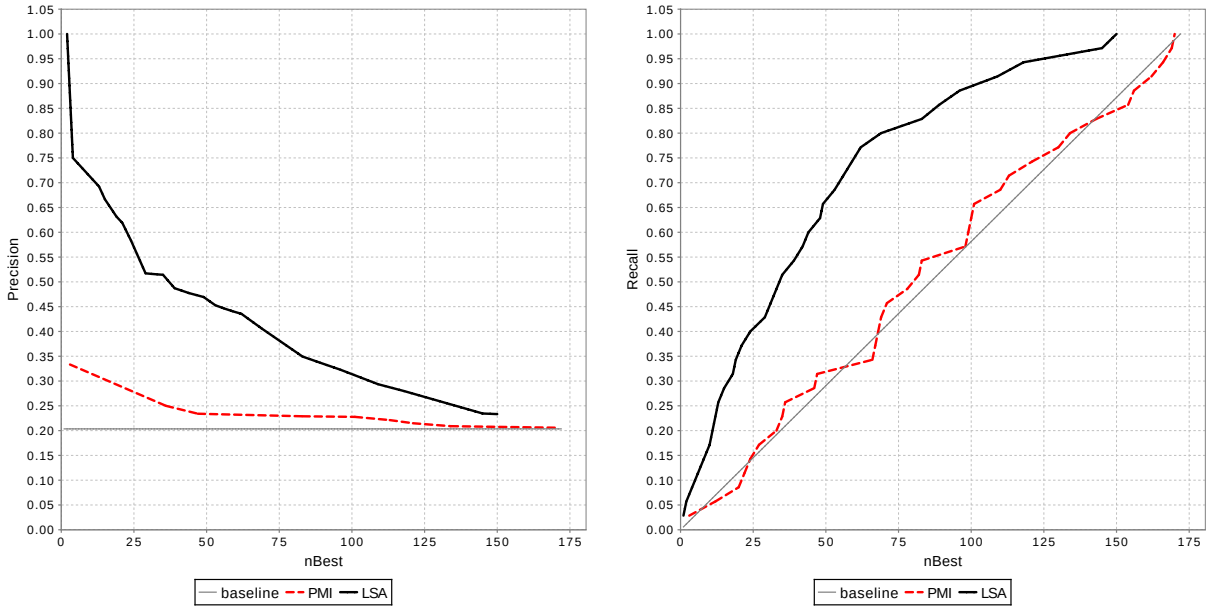

Figure 1: Smoothed graphs depicting the dependency of Precision (left) and Recall (right) upon the nBest selected non-compositional candidates from the ordered list of expressions in TestD created by the LSA and PMI-based models.

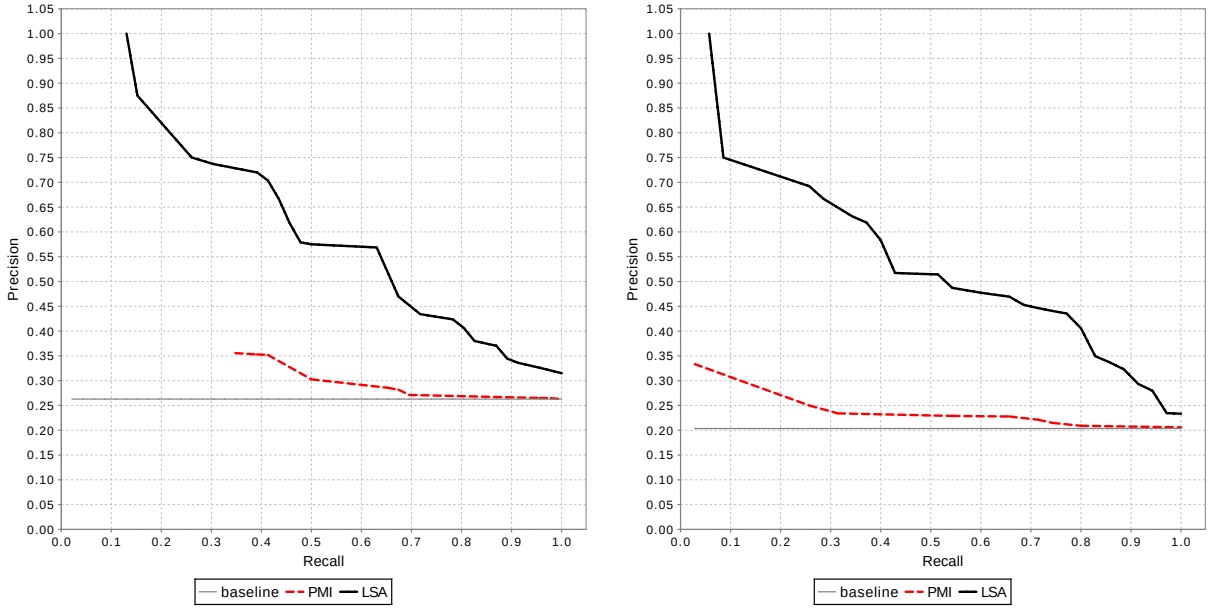

Figure 2: Smoothed graphs depicting the dependency of Precision upon Recall using the LSA and PMI-based models ordering the expressions in TrainValD (left) and TestD (right) according to their non-compositionality.