mwetoolkit+sem: Integrating Word Embeddings in the mwetoolkit for Semantic MWE Processing

Silvio Cordeiro[1,2], **Carlos Ramisch**[2], **Aline Villavicencio**[1]
[1] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[2] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr avillavicencio@inf.ufrgs.br

Abstract

This paper presents mwetoolkit+sem: an extension of the mwetoolkit that estimates semantic compositionality scores for multiword expressions (MWEs) based on word embeddings. First, we describe our implementation of vector-space operations working on distributional vectors. The compositionality score is based on the cosine distance between the MWE vector and the composition of the vectors of its member words. Our generic system can handle several types of word embeddings and MWE lists, and may combine individual word representations using several composition techniques. We evaluate our implementation on a dataset of 1042 English noun compounds (Farahmand et al., 2015), comparing different configurations of the underlying word embeddings and word-composition models. We show that our vector-based scores model non-compositionality better than standard association measures such as log-likelihood.

Keywords: Lexical semantics, multiword expressions, compositionality, word embeddings.

1. Introduction

Multiword expressions (MWEs) are often defined as word combinations "whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components" (Choueka, 1988). A broader definition of MWEs considers that, beyond non-compositionality, MWEs are word combinations formed by at least 2 lexemes and that present some lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Baldwin and Kim, 2010). Prototypical examples of MWEs cited in related work are often idioms such as *kick the bucket* or *let the cat out of the bag*.

With the growing interest in processing MWEs, there is an increasing need for tools that represent their lexical, syntactic and semantic characteristics. The mwetoolkit (Ramisch, 2015) provides one such language-independent framework for MWE discovery and identification in corpora. It has been successfully used for modeling lexical and syntactic characteristics of MWEs in many languages. It supports association scores that estimate the degree of conventionality of a candidate MWE based on the individual frequencies of component words. However, the ability to model one of the most salient properties of these expressions, that of semantic idiosyncrasy, was so far lacking.

We now introduce mwetoolkit+sem, containing a new set of capabilities that allow the use of distributional semantic models (DSMs) to estimate the semantic compositionality of MWEs. To this end, we have implemented and tested state-of-the-art techniques for estimating the compositionality of multiword combinations given the vector representations (or embeddings) of their component words. Our general principle is to combine the individual co-occurrence vectors of words and compare the result to the vector representing the co-occurrence pattern of the whole MWE. If both are close, the MWE is quite compositional, otherwise it is idiomatic.

In this paper, we compare standard association scores with the newly implemented models for predicting compositionality scores. We show that association scores capture conventional/compositional MWEs while compositionality scores capture idiomaticity. We evaluate mwetoolkit+sem on an existing dataset of English noun compounds (Farahmand et al., 2015) and show that the implemented compositionality scores correlate well with human judgments.

The developed framework allows treating idiomatic expressions as semantic units and representing compositional expressions as the combination of individual meanings. This information can in turn be exploited by NLP systems in tasks such as machine translation. The framework is freely available as part of the mwetoolkit[1].

This paper is structured as follows: in §2. we briefly discuss some techniques for meaning identification. In §3. we introduce mwetoolkit+sem and the semantic compositionality tools available. Their application for noun-compound compositionality prediction and the results obtained are presented in §4. and 5. We finish with conclusions and future work.

2. Related Work

The term "multiword expression" is often used as a synonym for "idiom", that is, an expression whose meaning of the component words is not directly found in the meaning of the whole combination (Choueka, 1988). However, state-of-the-art methods for automatic MWE discovery tend to use lexical, syntactic and statistical cues, rarely recurring to meaning (Evert, 2004; Seretan, 2011; Ramisch, 2015).

Meaning composition for MWEs requires accurate meaning representation of single words. To date, many models have been proposed for representing the lexical semantics of single words. We focus on distributional models, based on Harris' distributional hypothesis. DSMs have

[1] `http://mwetoolkit.sf.net`

been around for a while (Landauer and Dumais, 1997). However, the recent enthusiasm about neural networks and word embeddings has made DSMs more accurate and faster to build using very large corpora. Many tools are nowadays available for building word embeddings, like Dissect (Dinu et al., 2013), minimantics[2], word2vec[3] (Mikolov et al., 2013) and Glove[4] (Pennington et al., 2014).

Modeling semantic compositionality in DSMs is a hot topic in NLP. As word meanings can be represented as vectors, composition can be effectively modeled through simple operations like vector addition and multiplication (Mitchell and Lapata, 2010). Some authors have proposed models for estimating the degree of semantic idiomaticity of MWEs, focusing on noun compounds. Reddy et al. (2011) suggest a compositionality measure which is the cosine similarity between the MWE vector and the sum of the vectors of the component words. This model was also used by Salehi et al. (2015), in combination with word translation information coming from parallel corpora. Yazdani et al. (2015) propose and evaluate more sophisticated composition functions, based on linear, non-linear and neural network projections. The mwetoolkit+sem framework is based on vector addition and cosine similarity.

3. The mwetoolkit+sem Framework

Given the many state-of-the-art tools available for building DSMs, we assume that word embeddings are built offline by one of these dedicated tools. We implemented internal file readers that enable the automatic detection and reading of a variety of embedding formats in the toolkit, including:

- *Minimantics*: maps each target word to a set of context words. Each target-context pair appears in one line, along with association scores (frequency, PMI, etc). Users must explicitly select the score to be used.
- *word2vec*: each line describes a mapping from a target word to a real-valued n-dimensional vector, representing the all the contexts. The first line contains the number of embeddings and the value of n.
- *GloVe*: identical to *word2vec* with no first line header.

File readers see the word embedding files as a list of named embeddings, each of which associates a target's word form (e.g. its lemma) to a mapping between context identifiers and real values. For fixed-length embeddings, where there are no clear semantics attached to each dimension, we read the file as if each of the n values corresponded to artificial context identifiers $[c_0, \ldots, c_{n-1}]$. On the other hand, in models such as Minimantics, context identifiers are preserved as context-word forms.

For imputation of missing values we adopt two strategies. If a single-word is not found in the embeddings file, the zero vector is used instead. If an MWE candidate is not found in the embeddings file, we arbitrarily assign it the average compositionality score of all other MWEs in the list of candidates following Salehi et al. (2015).

For the semantic processing, we developed a new tool, called *feat_compositionality*, described in Figure 1. It outputs a compositionality score for each MWE in a list of input candidates, based on an input word-embeddings file.

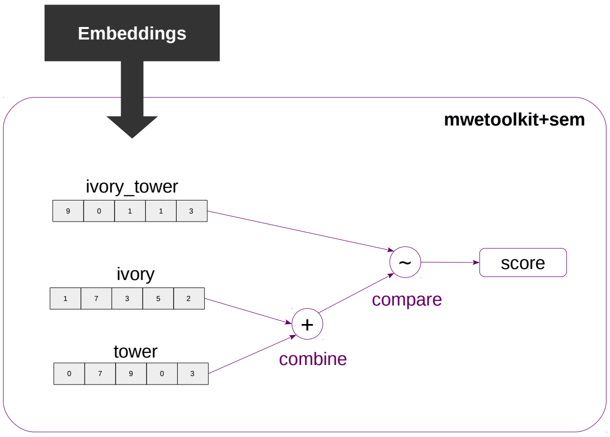

Figure 1: Overview of *feat_compositionality*

The first step combines the vectors $\overrightarrow{w_i}$ representing each word w_i in an MWE. The appropriate choice of a combinator depends on the semantics of the underlying vector space:

- *PointwiseAddition*: where pointwise vector addition is used to combine two embeddings (Mikolov et al., 2013). Weights can be applied to the combinator by explicitly specifying a list of multiplicative constants α_i, one for each component word w_i (option *combination-weights*).
- *PointwiseMultiplication*: using pointwise vector multiplication, where the elements of one vector are scaled by the other (Mitchell and Lapata, 2010).

Due to the way some DSMs are built, vectors are naturally scaled by the frequency of each word. Under the assumption that frequency does not determine semantics, it is convenient to normalize the input vector before and after combination. This is performed by default (using Euclidean norm), but can be explicitly disabled (options *no-normalize-input* and *no-normalize-combined*).

Once the embeddings of the words inside the MWE candidate have been combined (e.g. *bounty*$\oplus$*hunter*), we compare the result with the embedding of the MWE itself (the embedding for the token *bounty_hunter*) using cosine similarity. In short, the compositionality score using weighted pointwise addition for an MWE candidate composed of words w_1 through w_m is:

$$comp(w_1 \ldots w_m) = \cos\left(\frac{\overrightarrow{w_1 \ldots w_m}}{||\overrightarrow{w_1 \ldots w_m}||}, \sum_{j=1}^{m} \alpha_j \frac{\overrightarrow{w_j}}{||\overrightarrow{w_j}||}\right)$$

This model can be applied to MWEs with more than 2 words. For instance, it is possible to compute compositionality scores for 3-word compounds such as *liver_cell_line*.

4. Experimental Setup

To evaluate our framework, we predict non-compositionality for a set of 1042 English noun com-

[2] https://github.com/ceramisch/minimantics

[3] http://word2vec.googlecode.com/svn/trunk

[4] https://github.com/stanfordnlp/GloVe

pounds (Farahmand et al., 2015). Four human judges annotated each compound in this dataset as to whether it is *conventional* and *non-compositional*.[5] Our hypothesis is that, while standard association scores can predict conventionality, mwetoolkit+sem is much better at predicting non-compositionality. We compare the performance of our framework using different underlying distributional models, as well as different word combination weights.

We train an instance of each of these distributional semantic models: *minimantics*, *word2vec (cbow)* and *GloVe*. For training, we feed an MWE-annotated corpus where MWEs are joined as a single token as in *bounty_hunter*), as performed by Ferret (2014). We fix the following parameters:

- Corpus: UKWaC, containing 2G words of English texts crawled from the web (Baroni et al., 2009);
- Context window: lemma of each content word 8 words to the left/right of the target;
- Context weight decay: linear, that is, $\left[\frac{8}{8}, \frac{7}{8}, \frac{6}{8}, \ldots, \frac{1}{8}\right]$ (Levy et al., 2015);
- Dimensions per embedding: 250.

We train an additional model, $minimantics_B$, with a window of size 1 and dimension of 500, to verify the impact of the parameters. We compare our model for compositionality prediction with a simple baseline that uses the *log-likelihood* (LL) association score. LL compares an MWE frequency with the frequency of each component word.

We implemented several evaluation measures used in the literature to compare the model predictions with human judgments.

- Spearman's Rho (ρ): measures correlation between the ranks provided by the model predictions and by human judgments.
- Normalized Discounted Cumulative Gain (NDCG): a precision measure that penalizes more intensely wrong predictions at the top of the ranking.
- Best F-score (F_1): the highest F-score considering the first k predictions, for all values of k.
- Precision at k (P@k): precision for the top k predictions.
- Average precision (AP): average of precision calculated at each relevant prediction (Gurrutxaga and Alegria, 2013).

Our dataset contains four binary judgments per compound. For ρ, we use the sum of the binary judgments to rank the compounds. For NDCG, F_1, P@k and AP, a compound is considered relevant (i.e. conventional or non-compositional) if at least two judges consider it relevant (Yazdani et al., 2015).

5. Results

Table 1 presents the results when evaluating the predictive ability of different models concerning conventionality. Except for the baseline, all of the other models use a 50% : 50% combination weight (i.e. an average between the vector of the head and the vector of the modifier). All measures range from 0 to 1 (except for ρ, which ranges from -1 to 1); values close to 1 indicate better results.

Association scores are specifically designed for measuring the likelihood of non-random co-occurrence, and thus the baseline (LL) fares much better at this task. The other models, when obtaining good results, do so due to the fact that non-compositionality implies some level of conventionality.

	ρ	NDCG	F_1	P@100	AP
Baseline (LL)	**0.47**	**0.95**	**0.70**	**0.86**	**0.72**
minimantics	0.03	0.87	0.63	0.51	0.49
minimantics_B	-0.03	0.87	0.63	0.58	0.47
word2vec	-0.19	0.83	0.62	0.41	0.41
GloVe	0.18	0.89	0.64	0.58	0.55

Table 1: Evaluating for conventionality

While association scores can reasonably predict human judgments of conventionality, they do not perform as well when compared to judgments of non-compositionality, as shown in Table 2. The prediction based on distributional models, on the other hand, correlates much better with the human non-compositionality scores. These results are comparable with what has been found in other works (Yazdani et al., 2015), even though we have not tuned the parameters of our models.

	ρ	NDCG	F_1	P@100	AP
Baseline (LL)	-0.19	0.63	0.32	0.09	0.15
minimantics	0.17	0.72	0.36	0.32	0.27
minimantics_B	0.21	0.71	0.40	0.27	0.27
word2vec	**0.31**	**0.84**	**0.46**	**0.46**	**0.40**
GloVe	0.07	0.68	0.35	0.14	0.21
Yazdani2015	0.41	0.86	0.49	0.54	N/A

Table 2: Evaluating for non-compositionality (50% : 50%)

These results indicate that predicting conventionality seems to be an easier task than non-compositionality, as reflected in the higher scores obtained for almost all measures, except for ρ. As conventionality represents the preference of speakers for a given word or expression, this is often reflected in its frequency, and may be more directly measurable than compositionality.

In addition to the evaluation of a uniform combination of component words, the implemented tools also allow one to emphasize the higher relevance of one of the component words for compositionality prediction, by applying different weights. While more compositional cases will be better captured by a uniform combination of both words, partly compositional cases will be more related to one of the component words (e.g. *crocodile tears*). Tables 3 and 4 present a non-compositionality evaluation when considering only the head of each compound (weight 0% : 100%) or only the modifier (weight 100% : 0%). The results show that fully focusing on one of the components, the head, reduces the power of compositionality prediction in the

[5] Since our model predicts *compositionality*, and the human judges annotated *non-compositionality*, we inverted the model predictions. Thus, measures based on precision focus on selecting non-compositional compounds.

model. On the other hand, focusing on the modifier generated slightly better results than in Table 2. This suggests that there is a higher prevalence of partly compositional compounds in the dataset whose meaning is related to the modifier. By simply applying different combination weights, mwetoolkit+sem is able to detect irregular patterns such as the head-modifier imbalance that is present in the target dataset.

	ρ	NDCG	F_1	P@100	AP
minimantics	0.08	0.69	0.33	0.25	0.22
minimantics$_B$	**0.15**	0.70	**0.37**	**0.30**	0.25
word2vec	**0.15**	**0.79**	0.36	0.29	**0.29**
GloVe	0.01	0.67	0.33	0.13	0.19

Table 3: Evaluating for non-compositionality (0% : 100%)

	ρ	NDCG	F_1	P@100	AP
minimantics	0.25	0.76	0.44	0.45	0.34
minimantics$_B$	0.23	0.72	0.43	0.33	0.29
word2vec	**0.33**	**0.82**	**0.49**	**0.53**	**0.44**
GloVe	0.13	0.69	0.38	0.18	0.23

Table 4: Evaluating for non-compositionality (100% : 0%)

Our goal is not to tune the system so that it overcomes state-of-the-art results. Instead, we provide a simple implementation that allows easy estimation of non-compositionality from the output of several distributional semantic models. This can be helpful for researchers studying MWE semantics and building more sophisticated models.

Table 5 presents the MWEs that were ranked at the extremities in the system output. Errors are marked in bold. The bottom-10 MWEs have been ranked by both humans and the system as fully compositional, with a very low non-compositionality score. The only exception seems to be *web site*, which was judged by humans as fairly non-compositional (2/4). This happens to be the most frequent compound in our corpus, which is built precisely by web site crawling. In this domain, it is quite reasonable to assume that a *web site* is a *site* present on the *web*, thus the individual vectors of *web* and *site* would correspond to compositional meanings, even if non-literal. Overall, compositional compounds seem to be correctly modeled in the system.

Detecting non-compositionality, on the other hand, can be somewhat more complicated: while non-compositional expressions such as *think tank* are being correctly classified, other MWEs such as *carnival crowd* and *background target* are simply not frequent enough in the corpus. Therefore, as the distributional vectors were built based on limited occurrence contexts, the compositionality predictions are not reliable.

Finally, the implemented tools are computationally efficient, and using *feat_compositionality* for evaluating the models took less than 5 min to execute for each embedding configuration, in spite of processing the full dataset of 1042 compounds.[6]

Compound	System	Human	Freq.
think tank	.133	4	7083
blood bath	.010	4	157
action figure	.003	4	843
front line	-.012	4	13534
grass root	-.023	4	3454
carnival crowd	-.034	0	10
brain drain	-.046	4	637
background target	-.059	0	5
supporting act	-.060	3	142
computer innovation	-.071	0	9
visual art	-.817	0	11208
user interface	-.818	0	14046
computer network	-.819	0	4556
government funding	-.820	0	5322
football league	-.823	0	6375
data analysis	-.832	0	9138
web site	-.849	2	266750
web page	-.855	0	89306
debut album	-.863	0	7005
contact detail	-.863	0	50793

Table 5: System scores, human scores and UKWaC frequency of top-10 and bottom-10 MWEs ranked for non-compositionality using word2vec CBOW embeddings and 50% : 50% composition.

6. Conclusions and Future Work

This paper introduced mwetoolkit+sem, an integrated framework for processing MWEs. It offers semantic and word embedding capabilities in addition to the standard techniques for lexical and syntactic MWE representation. It includes a tool that processes a list of potential MWEs and outputs a compositionality prediction for each of them. This complements the various association scores available in the toolkit, which fare better at predicting conventionality, with a new measure that performs better at predicting semantic compositionality.

The results we obtained are compatible with the state of the art even with simple methods and without any optimization. As future work, we plan on providing support for other similarity scores and sense composition functions (e.g. matrices, tensors and neural networks). We also intend on performing an extensive evaluation of techniques examining conventionality and compositionality on other datasets and languages.

7. Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche through projects PARSEME-FR (ANR-14-CERA-0001) and ORFEO (ANR-12-CORP-0005), and by French-Brazilian cooperation projects CAMELEON (CAPES-COFECUB #707/11) and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7). Part of the results presented in this paper were obtained through research on a project titled "Simplificação Textual de Expressões Complexas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

[6] In comparison, building the models took longer and required more computational resources. For instance, word2vec model occupied around 0.5GB of disk.

8. Bibliographical References

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In *Handbook of Natural Language Processing, Second Edition.*, pages 267–292.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. pages 209–226.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In Christian Fluhr et al., editors, *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988)*, pages 609–624, Cambridge, MA, USA, March. CID.

Dinu, G., Pham, N. T., and Baroni, M. (2013). Dissect - distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Sofia, Bulgaria, August. Association for Computational Linguistics.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany. 353 p.

Farahmand, M., Smith, A., and Nivre, J. (2015). A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Denver, Colorado, June. Association for Computational Linguistics.

Ferret, O. (2014). Compounds and distributional thesauri. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2979–2984, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1590.

Gurrutxaga, A. and Alegria, I. n. (2013). Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 116–125, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Ramisch, C. (2015). *Multiword Expressions Acquisition - A Generic and Open Framework*. Theory and Applications of Natural Language Processing. Springer.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, Chiang Mai, Thailand, November.

Salehi, B., Cook, P., and Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado, May–June. Association for Computational Linguistics.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Dordrecht, Netherlands, 1st edition. 212 p.

Yazdani, M., Farahmand, M., and Henderson, J. (2015). Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal, September. Association for Computational Linguistics.