Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality

Bahar Salehi,♣♡ Paul Cook♡ and Timothy Baldwin♣♡
♣ NICTA Victoria Research Laboratory
♡ Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
bsalehi@student.unimelb.edu.au, paulcook@unimelb.edu.au, tb@ldwin.net

Abstract

We predict the compositionality of multiword expressions using distributional similarity between each component word and the overall expression, based on translations into multiple languages. We evaluate the method over English noun compounds, English verb particle constructions and German noun compounds. We show that the estimation of compositionality is improved when using translations into multiple languages, as compared to simply using distributional similarity in the source language. We further find that string similarity complements distributional similarity.

1 Compositionality of MWEs

Multiword expressions (hereafter MWEs) are combinations of words which are lexically, syntactically, semantically or statistically idiosyncratic (Sag et al., 2002; Baldwin and Kim, 2009). Much research has been carried out on the extraction and identification of MWEs[1] in English (Schone and Jurafsky, 2001; Pecina, 2008; Fazly et al., 2009) and other languages (Dias, 2003; Evert and Krenn, 2005; Salehi et al., 2012). However, considerably less work has addressed the task of predicting the meaning of MWEs, especially in non-English languages. As a step in this direction, the focus of this study is on predicting the compositionality of MWEs.

An MWE is fully compositional if its meaning is predictable from its component words, and it is non-compositional (or idiomatic) if not. For example, *stand up* "rise to one's feet" is compositional, because its meaning is clear from the meaning of the components *stand* and *up*. However, the meaning of *strike up* "to start playing" is largely unpredictable from the component words *strike* and *up*.

In this study, following McCarthy et al. (2003) and Reddy et al. (2011), we consider compositionality to be graded, and aim to predict the *degree* of compositionality. For example, in the dataset of Reddy et al. (2011), *climate change* is judged to be 99% compositional, while *silver screen* is 48% compositional and *ivory tower* is 9% compositional. Formally, we model compositionality prediction as a regression task.

An explicit handling of MWEs has been shown to be useful in NLP applications (Ramisch, 2012). As an example, Carpuat and Diab (2010) proposed two strategies for integrating MWEs into statistical machine translation. They show that even a large scale bilingual corpus cannot capture all the necessary information to translate MWEs, and that in adding the facility to model the compositionality of MWEs into their system, they could improve translation quality. Acosta et al. (2011) showed that treating non-compositional MWEs as a single unit in information retrieval improves retrieval effectiveness. For example, while searching for documents related to *ivory tower*, we are almost certainly not interested in documents relating to elephant tusks.

Our approach is to use a large-scale multi-way translation lexicon to source translations of MWEs and their component words, and then model the relative similarity between each of the component words and the MWE, using distributional similarity based on monolingual corpora for the source language and each of the target languages. Our hypothesis is that using distributional similarity in more than one language will improve the prediction of compositionality. Importantly, in order to make the method as language-independent and

[1] In this paper, we follow Baldwin and Kim (2009) in considering MWE "identification" to be a token-level disambiguation task, and MWE "extraction" to be a type-level lexicon induction task.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 472–481,
Gothenburg, Sweden, April 26-30 2014.

broadly-applicable as possible, we make no use of corpus preprocessing such as lemmatisation, and rely only on the availability of a translation dictionary and monolingual corpora.

Our results confirm our hypothesis that distributional similarity over the source language in addition to multiple target languages improves the quality of compositionality prediction. We also show that our method can be complemented with string similarity (Salehi and Cook, 2013) to further improve compositionality prediction. We achieve state-of-the-art results over two datasets.

2 Related Work

Most recent work on predicting the compositionality of MWEs can be divided into two categories: language/construction-specific and general-purpose. This can be at either the token-level (over token occurrences of an MWE in a corpus) or type-level (over the MWE string, independent of usage). The bulk of work on compositionality has been language/construction-specific and operated at the token-level, using dedicated methods to identify instances of a given MWE, and specific properties of the MWE in that language to predict compositionality (Lin, 1999; Kim and Baldwin, 2007; Fazly et al., 2009).

General-purpose token-level approaches such as distributional similarity have been commonly applied to infer the semantics of a word/MWE (Schone and Jurafsky, 2001; Baldwin et al., 2003; Reddy et al., 2011). These techniques are based on the assumption that the meaning of a word is predictable from its context of use, via the neighbouring words of token-level occurrences of the MWE. In order to predict the compositionality of a given MWE using distributional similarity, the different contexts of the MWE are compared with the contexts of its components, and the MWE is considered to be compositional if the MWE and component words occur in similar contexts.

Identifying token instances of MWEs is not always easy, especially when the component words do not occur sequentially. For example consider *put on* in ***put*** *your jacket* ***on***, and ***put*** *your jacket* ***on*** *the chair*. In the first example *put on* is an MWE while in the second example, *put on* is a simple verb with prepositional phrase and not an instance of an MWE. Moreover, if we adopt a conservative identification method, the number of token occurrences will be limited and the distributional scores may not be reliable. Additionally, for morphologically-rich languages, it can be difficult to predict the different word forms a given MWE type will occur across, posing a challenge for our requirement of no language-specific preprocessing.

Pichotta and DeNero (2013) proposed a token-based method for identifying English phrasal verbs based on parallel corpora for 50 languages. They show that they can identify phrasal verbs better when they combine information from multiple languages, in addition to the information they get from a monolingual corpus. This finding lends weight to our hypothesis that using translation data and distributional similarity from each of a range of target languages, can improve compositionality prediction. Having said that, the general applicability of the method is questionable — there are many parallel corpora involving English, but for other languages, this tends not to be the case.

Salehi and Cook (2013) proposed a general-purpose type-based approach using translation data from multiple languages, and string similarity between the MWE and each of the component words. They use training data to identify the best-10 languages for a given family of MWEs, on which to base the string similarity, and once again find that translation data improves their results substantially. Among the four string similarity measures they experimented with, longest common substring was found to perform best. Their proposed method is general and applicable to different families of MWEs in different languages. In this paper, we reimplement the method of Salehi and Cook (2013) using longest common substring (LCS), and both benchmark against this method and combine it with our distributional similarity-based method.

3 Our Approach

To predict the compositionality of a given MWE, we first measure the semantic similarity between the MWE and each of its component words[2] using distributional similarity based on a monolingual corpus in the source language. We then repeat the process for translations of the MWE and its component words into each of a range of target languages, calculating distributional similarity using

[2]Note that we will always assume that there are two component words, but the method is easily generalisable to MWEs with more than two components.

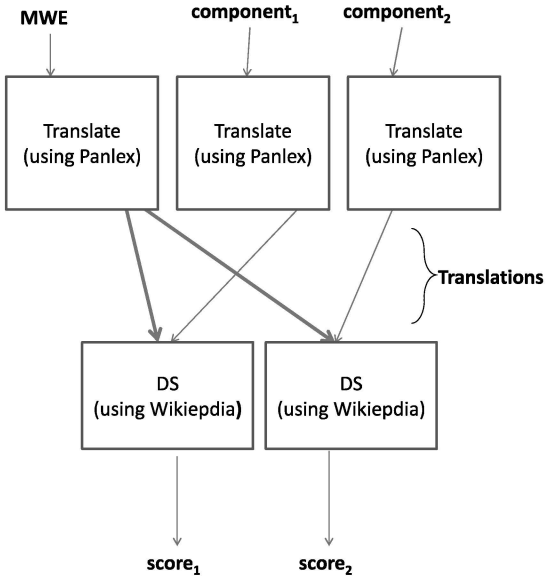


Figure 1: Outline of our approach to computing the distributional similarity (DS) of translations of an MWE with each of its component words, for a given target language. $score_1$ and $score_2$ are the similarity for the first and second components, respectively. We obtain translations from Panlex, and use Wikipedia as our corpus for each language.

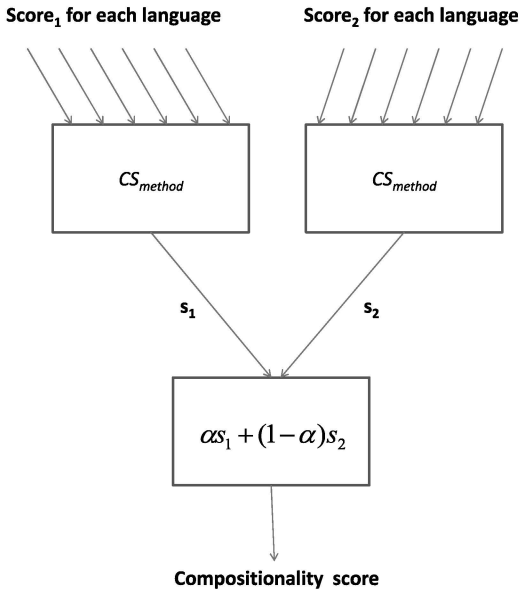


Figure 2: Outline of the method for combining distributional similarity scores from multiple languages, across the components of the MWE. CS_{method} refers to one of the methods described in Section 3.2 for calculating compositionality.

a monolingual corpus in the target language (Figure 1). We additionally use supervised learning to identify which target languages (or what weights for each language) optimise the prediction of compositionality (Figure 2). We hypothesise that by using multiple translations — rather than only information from the source language — we will be able to better predict compositionality.

We optionally combine our proposed approach with string similarity, calculated based on the method of Salehi and Cook (2013), using LCS.

Below, we detail our method for calculating distributional similarity in a given language, the different methods for combining distributional similarity scores into a single estimate of compositionality, and finally the method for selecting the target languages to use in calculating compositionality.

3.1 Calculating Distributional Similarity

In order to be consistent across all languages and be as language-independent as possible, we calculate distributional similarity in the following manner for a given language.

Tokenisation is based on whitespace delimiters and punctuation; no lemmatisation or case-folding is carried out. Token instances of a given MWE or component word are identified by full-token n-gram matching over the token stream. We assume that all full stops and equivalent characters for other orthographies are sentence boundaries, and chunk the corpora into (pseudo-)sentences on the basis of them. For each language, we identify the 51st–1050th most frequent words, and consider them to be content-bearing words, in the manner of Schütze (1997). This is based on the assumption that the top-50 most frequent words are stop words, and not a good choice of word for calculating distributional similarity over. That is not to say that we can't calculate the distributional similarity for stop words, however (as we will for the verb particle construction dataset — see Section 4.3.2) they are simply not used as the dimensions in our calculation of distributional similarity.

We form a vector of content-bearing words across all token occurrences of the target word,

on the basis of these content-bearing words. Distributional similarity is calculated over these context vectors using cosine similarity. According to Weeds (2003), using dependency relations with the neighbouring words of the target word can better predict the meaning of the target word. However, in line with our assumption of no language-specific preprocessing, we just use word co-occurrence.

3.2 Calculating Compositionality

First, we need to calculate a combined compositionality score from the individual distributional similarities between each component word and the MWE. Following Reddy et al. (2011), we combine the component scores using the weighted mean (as shown in Figure 2):

$$\text{comp} = \alpha s_1 + (1 - \alpha) s_2 \quad (1)$$

where s_1 and s_2 are the scores for the first and the second component, respectively. We use different α settings for each dataset, as detailed in Section 4.3.

We experiment with a range of methods for calculating compositionality, as follows:

CS_{L1}: calculate distributional similarity using only distributional similarity in the source language corpus (This is the approach used by Reddy et al. (2011), as discussed in Section 2).

CS_{L2N}: exclude the source language, and compute the mean of the distributional similarity scores for the best-N target languages. The value of N is selected according to training data, as detailed in Section 3.3.

CS_{L1+L2N}: calculate distributional similarity over both the source language (CS_{L1}) and the mean of the best-N languages (CS_{L2N}), and combine via the arithmetic mean.[3] This is to examine the hypothesis that using multiple target languages is better than just using the source language.

$CS_{SVR(L1+L2)}$: train a support vector regressor (SVR: Smola and Schölkopf (2004)) over the distributional similarities for all 52 languages (source and target languages).

CS_{string}: calculate string similarity using the LCS-based method of Salehi and Cook (2013).[4]

$CS_{string+L1}$: calculate the mean of the string similarity (CS_{string}) and distributional similarity in the source language (Salehi and Cook, 2013).

CS_{all}: calculate the mean of the string similarity (CS_{string}) and distributional similarity scores (CS_{L1} and CS_{L2N}).

3.3 Selecting Target Languages

We experiment with two approaches for combining the compositionality scores from multiple target languages.

First, in CS_{L2N} (and CS_{L1+L2N} and CS_{all} that build off it), we use training data to rank the target languages according to Pearson's correlation between the predicted compositionality scores and the gold-standard compositionality judgements. Based on this ranking, we take the best-N languages, and combine the individual compositionality scores by taking the arithmetic mean. We select N by determining the value that optimises the correlation over the training data. In other words, the selection of N and accordingly the best-N languages are based on nested cross-validation over training data, independently of the test data for that iteration of cross-validation.

Second in $CS_{SVR(L1+L2)}$, we combine the compositionality scores from the source and all 51 target languages into a feature vector, and train an SVR over the data using LIBSVM.[5]

4 Resources

In this section, we describe the resources required by our method, and also the datasets used to evaluate our method.

4.1 Monolingual Corpora for Different Languages

We collected monolingual corpora for each of 52 languages (51 target languages + 1 source language) from XML dumps of Wikipedia. These languages are based on the 54 target languages

[3] We also experimented with taking the mean over all the languages — target and source — but found it best to combine the scores for the target languages first, to give more weight to the source language.

[4] Due to differences in our random partitioning, our reported results over the two English datasets differ slightly over the results of Salehi and Cook (2013) using the same method.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm

used by Salehi and Cook (2013), excluding Spanish because we happened not to have a dump of Spanish Wikipedia, and also Chinese and Japanese because of the need for a language-specific word tokeniser. The raw corpora were preprocessed using the WP2TXT toolbox[6] to eliminate XML tags, HTML tags and hyperlinks, and then tokenisation based on whitespace and punctuation was performed. The corpora vary in size from roughly 750M tokens for English, to roughly 640K tokens for Marathi.

4.2 Multilingual Dictionary

To translate the MWEs and their components, we follow Salehi and Cook (2013) in using Panlex (Baldwin et al., 2010). This online dictionary is massively multilingual, covering more than 1353 languages. For each MWE dataset (see Section 4.3), we translate the MWE and component words from the source language into each of the 51 languages.

In instances where there is no direct translation in a given language for a term, we use a pivot language to find translation(s) in the target language. For example, the English noun compound *silver screen* has direct translations in only 13 languages in Panlex, including Vietnamese (*màn bac*) but not French. There is, however, a translation of *màn bac* into French (*cinéma*), allowing us to infer an indirect translation between *silver screen* and *cinéma*. In this way, if there are no direct translations into a particular target language, we search for a single-pivot translation via each of our other target languages, and combine them all together as our set of translations for the target language of interest.

In the case that no translation (direct or indirect) can be found for a given source language term into a particular target language, the compositionality score for that target language is set to the average across all target languages for which scores can be calculated for the given term. If no translations are available for any target language (e.g. the term is not in Panlex) the compositionality score for each target language is set to the average score for that target language across all other source language terms.

[6] http://wp2txt.rubyforge.org/

4.3 Datasets

We evaluate our proposed method over three datasets (two English, one German), as described below.

4.3.1 English Noun Compounds (ENC)

Our first dataset is made up of 90 binary English noun compounds, from the work of Reddy et al. (2011). Each noun compound was annotated by multiple annotators using the integer scale 0 (fully non-compositional) to 5 (fully compositional). A final compositionality score was then calculated as the mean of the scores from the annotators. If we simplistically consider 2.5 as the threshold for compositionality, the dataset is relatively well balanced, containing 48% compositional and 52% non-compositional noun compounds. Following Reddy et al. (2011), in combining the component-wise distributional similarities for this dataset, we weight the first component in Equation 1 higher than the second ($\alpha = 0.7$).

4.3.2 English Verb Particle Constructions (EVPC)

The second dataset contains 160 English verb particle constructions (VPCs), from the work of Bannard (2006). In this dataset, a verb particle construction consists of a verb (the head) and a prepositional particle (e.g. *hand in*, *look up* or *battle on*).

For each component word (the verb and particle, respectively), multiple annotators were asked whether the VPC entails the component word. In order to translate the dataset into a regression task, we calculate the overall compositionality as the number of annotations of entailment for the verb, divided by the total number of verb annotations for that VPC. That is, following Bannard et al. (2003), we only consider the compositionality of the verb component in our experiments (and as such $\alpha = 1$ in Equation 1).

One area of particular interest with this dataset will be the robustness of the method to function words (the particles), both under translation and in terms of calculating distributional similarity, although the findings of Baldwin (2006) for English prepositions are at least encouraging in this respect. Additionally, English VPCs can occur in "split" form (e.g. ***put*** *your jacket* ***on***, from our earlier example), which will complicate identification, and the verb component will often be inflected and thus not match under our identification strategy (for both VPCs and the component verbs).

Dataset	Language	Frequency	Family
ENC	Italian	100	Romance
	French	99	Romance
	German	86	Germanic
	Vietnamese	83	Viet-Muong
	Portuguese	62	Romance
EVPC	Bulgarian	100	Slavic
	Breton	100	Celtic
	Occitan	100	Romance
	Indonesian	100	Indonesian
	Slovenian	100	Slavic
GNC	Polish	100	Slavic
	Lithuanian	99	Baltic
	Finnish	74	Uralic
	Bulgarian	72	Slavic
	Czech	40	Slavic

Table 1: The 5 best languages for the ENC, EVPC and GNC datasets. The language family is based on Voegelin and Voegelin (1977).

4.3.3 German Noun Compounds (GNC)

Our final dataset is made up of 246 German noun compounds (von der Heide and Borgwaldt, 2009; Schulte im Walde et al., 2013). Multiple annotators were asked to rate the compositionality of each German noun compound on an integer scale of 1 (non-compositional) to 7 (compositional). The overall compositionality score is then calculated as the mean across the annotators. Note that the component words are provided as part of the dataset, and that there is no need to perform decompounding. Following Schulte im Walde et al. (2013), we weight the first component higher in Equation 1 ($\alpha = 0.8$) when calculating the overall compositionality score.

This dataset is significant in being non-English, and also in that German has relatively rich morphology, which we expect to impact on the identification of both the MWE and the component words.

5 Results

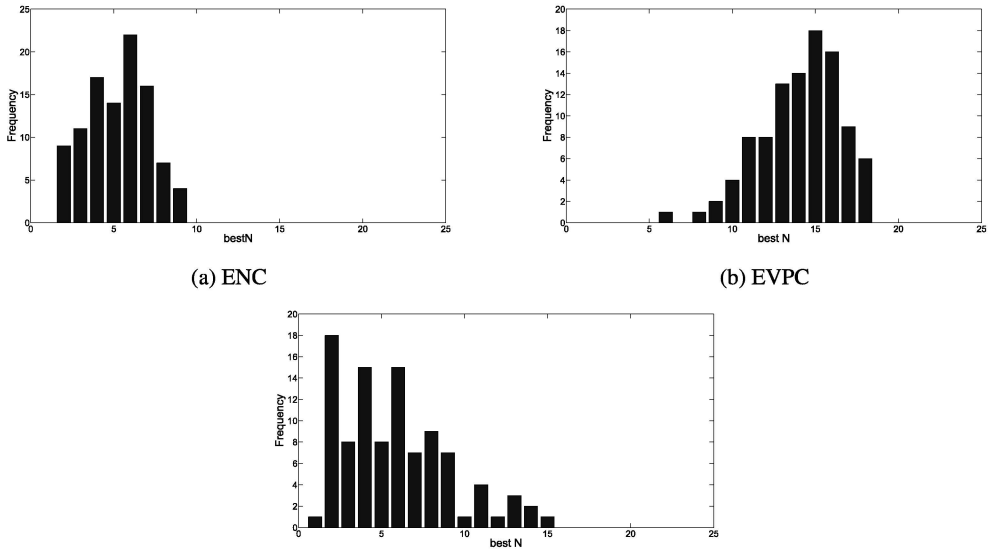
All experiments are carried out using 10 iterations of 10-fold cross validation, randomly partitioning the data independently on each of the 10 iterations, and averaging across all 100 test partitions in our presented results. In the case of CS_{L2N} and other methods that make use of it (i.e. CS_{L1+L2N} and CS_{all}), the languages selected for a given training fold are then used to compute the compositionality scores for the instances in the test set. Figures 3a, 3b and 3c are histograms of the number of times each N is selected over 100 folds on ENC, EVPC and GNC datasets, respectively. From the histograms, $N = 6$, $N = 15$ and $N = 2$ are the most commonly selected settings for ENC, EVPC and GNC, respectively. That is, multiple languages are generally used, but more languages are used for English VPCs than either of the compound noun datasets. The 5 most-selected languages for ENC, EVPC and GNC are shown in Table 1. As we can see, there are some languages which are always selected for a given dataset, but equally the commonly-selected languages vary considerably between datasets.

Further analysis reveals that 32 (63%) target languages for ENC, 25 (49%) target languages for EVPC, and only 5 (10%) target languages for GNC have a correlation of $r \geq 0.1$ with gold-standard compositionality judgements. On the other hand, 8 (16%) target languages for ENC, 2 (4%) target languages for EVPC, and no target languages for GNC have a correlation of $r \leq -0.1$.

5.1 ENC Results

English noun compounds are relatively easy to identify in a corpus,[7] because the components occur sequentially, and the only morphological variation is in noun number (singular vs. plural). In other words, the precision for our token matching method is very high, and the recall is also acceptably high. Partly as a result of the ease of identification, we get a high correlation of $r = 0.700$ for CS_{L1} (using only source language data). Using only target languages (CS_{L2N}), the results drop to $r = 0.434$, but when we combine the two (CS_{L1+L2N}), the correlation is higher than using only source or target language data, at $r = 0.725$. When we combine all languages using SVR, the results rise slightly higher again to $r = 0.744$, which is slightly above the correlation of the state-of-the-art method of Salehi and Cook (2013), which combines their method with the method of Reddy et al. (2011) ($CS_{string+L1}$). These last two results support our hypothesis that using translation data can improve the prediction of compositionality. The results for string similarity on its own (CS_{string}, $r = 0.644$) are slightly lower than those using only source language distributional similarity, but when combined with

[7] Although see Lapata and Lascarides (2003) for discussion of the difficulty of reliably identifying low-frequency English noun compounds.



(a) ENC

(b) EVPC

(c) GNC

Figure 3: Histograms displaying how many times a given N is selected as the best number of languages over each dataset. For example, according to the GNC chart, there is a peak for $N = 2$, which shows that over 100 folds, the best-2 languages achieved the highest correlation on 18 folds.

Method	Summary of the Method	ENC	EVPC	GNC
CS_{L1}	Source language	0.700	0.177	0.141
CS_{L2N}	Best-N target languages	0.434	0.398	0.113
CS_{L1+L2N}	Source + best-N target languages	0.725	0.312	0.178
$CS_{SVR(L1+L2)}$	SVR (Source + all 51 target languages)	**0.744**	0.389	0.085
CS_{string}	String Similarity (Salehi and Cook, 2013)	0.644	0.385	**0.372**
$CS_{string+L1}$	CS_{string} +CS_{L1} (Salehi and Cook, 2013)	0.739	0.360	0.353
CS_{all}	CS_{L1} + CS_{L2N} + CS_{string}	0.732	**0.417**	0.364

Table 2: Pearson's correlation on the ENC, EVPC and GNC datasets

CS_{L1+L2N} (i.e. CS_{all}) there is a slight rise in correlation (from $r = 0.725$ to $r = 0.732$).

5.2 EVPC Results

English VPCs are hard to identify. As discussed in Section 2, VPC components may not occur sequentially, and even when they do occur sequentially, they may not be a VPC. As such, our simplistic identification method has low precision and recall (hand analysis of 927 identified VPC instances would suggest a precision of around 74%). There is no question that this is a contributor to the low correlation for the source language method (CS_{L1}; $r = 0.177$). When we use target languages instead of the source language (CS_{L2N}), the correlation jumps substantially to $r = 0.398$.

When we combine English and the target languages (CS_{L1+L2N}), the results are actually lower than just using the target languages, because of the high weight on the target language, which is not desirable for VPCs, based on the source language results. Even for $CS_{SVR(L1+L2)}$, the results ($r = 0.389$) are slightly below the target language-only results. This suggests that when predicting the compositionality of MWEs which are hard to identify in the source language, it may actually be better to use target languages only. The results for string similarity (CS_{string}: $r = 0.385$) are similar to those for CS_{L2N}. However, as with the ENC dataset, when we combine string similarity and distributional similarity (CS_{all}), the results improve, and we achieve the state-of-the-art for the dataset.

In Table 3, we present classification-based eval-

Method	Precision	Recall	F-score ($\beta = 1$)	Accuracy
Bannard et al. (2003)	60.8	66.6	63.6	60.0
Salehi and Cook (2013)	**86.2**	71.8	77.4	69.3
CS_{all}	79.5	**89.3**	**82.0**	**74.5**

Table 3: Results (%) for the binary compositionality prediction task on the EVPC dataset

uation over a subset of EVPC, binarising the compositionality judgements in the manner of Bannard et al. (2003). Our method achieves state-of-the-art results in terms of overall F-score and accuracy.

5.3 GNC Results

German is a morphologically-rich language, with marking of number and case on nouns. Given that we do not perform any lemmatization or other language-specific preprocessing, we inevitably achieve low recall for the identification of noun compound tokens, although the precision should be nearly 100%. Partly because of the resultant sparseness in the distributional similarity method, the results for CS_{L1} are low ($r = 0.141$), although they are lower again when using target languages ($r = 0.113$). However, when we combine the source and target languages (CS_{L1+L2N}) the results improve to $r = 0.178$. The results for $CS_{SVR(L1+L2)}$, on the other hand, are very low ($r = 0.085$). Ultimately, simple string similarity achieves the best results for the dataset ($r = 0.372$), and this result actually drops slightly when combined with the distributional similarities.

To better understand the reason for the lacklustre results using SVR, we carried out error analysis and found that, unlike the other two datasets, about half of the target languages return scores which correlate negatively with the human judgements. When we filter these languages from the data, the score for SVR improves appreciably. For example, over the best-3 languages overall, we get a correlation score of $r = 0.179$, which is slightly higher than CS_{L1+L2N}.

We further investigated the reason for getting very low and sometimes negative correlations with many of our target languages. We noted that about 24% of the German noun compounds in the dataset do not have entries in Panlex. This contrasts with ENC where only one instance does not have an entry in Panlex, and EVPC where all VPCs have translations in at least one language in Panlex. We experimented with using string similarity scores in the case of such missing translations, as opposed to the strategy described in Section 4.2. The results for $CS_{SVR(L1+L2)}$ rose to $r = 0.269$, although this is still below the correlation for just using string similarity.

Our results on the GNC dataset using string similarity are competitive with the state-of-the-art results ($r = 0.45$) using a window-based distributional similarity approach over monolingual German data (Schulte im Walde et al., 2013). Note, however, that their method used part-of-speech information and lemmatisation, where ours does not, in keeping with the language-independent philosophy of this research.

6 Conclusion and Future Work

In this study, we proposed a method to predict the compositionality of MWEs based on monolingual distributional similarity between the MWE and each of its component words, under translation into multiple target languages. We showed that using translation and multiple target languages enhances compositionality modelling, and also that there is strong complementarity between our approach and an approach based on string similarity.

In future work, we hope to address the question of translation sparseness, as observed for the GNC dataset. We also plan to experiment with unsupervised morphological analysis methods to improve identification recall, and explore the impact of tokenization. Furthermore, we would like to investigate the optimal number of stop words and content-bearing words for each language, and to look into the development of general unsupervised methods for compositionality prediction.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References

Otavio Acosta, Aline Villavicencio, and Viviane 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*, pages 101–109, Portland, USA.

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Timothy Baldwin, Jonathan Pool, and Susan M Colowick. 2010. Panlex and lextract: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 37–40, Beijing, China.

Timothy Baldwin. 2006. Distributional similarity and preposition semantics. In Patrick Saint-Dizier, editor, *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*, pages 197–210. Springer, Dordrecht, Netherlands.

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*, pages 65–72, Sapporo, Japan.

Colin James Bannard. 2006. *Acquiring Phrasal Lexicons from Corpora*. Ph.D. thesis, University of Edinburgh.

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*, pages 242–245, Los Angeles, USA.

Gaël Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 41–48, Sapporo, Japan.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):450–466.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Su Nam Kim and Timothy Baldwin. 2007. Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of the 7th Meeting of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 40–48, Melbourne, Australia.

Mirella Lapata and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics (EACL-2003)*, pages 235–242, Budapest, Hungary.

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*, pages 317–324, College Park, 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*, pages 73–80, Sapporo, Japan.

Pavel Pecina. 2008. *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

Karl Pichotta and John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA.

Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, pages 61–66, Jeju Island, Korea.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*, pages 210–218, Chiang Mai, Thailand.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing Computational Linguistics (CICLing-2002)*, pages 189–206, Mexico City, Mexico.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 266–275, Atlanta, USA.

Bahar Salehi, Narjes Askarian, and Afsaneh Fazly. 2012. Automatic identification of Persian light verb constructions. In *Proceedings of the 13th International Conference on Intelligent Text Processing Computational Linguistics (CICLing-2012)*, pages 201–210, New Delhi, India.

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108, Hong Kong, China.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, Atlanta, USA.

Hinrich Schütze. 1997. *Ambiguity Resolution in Language Learning*. CSLI Publications, Stanford, USA.

Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Charles Frederick Voegelin and Florence Marie Voegelin. 1977. *Classification and index of the world's languages*, volume 4. New York: Elsevier.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter, Basis und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Julie Elizabeth Weeds. 2003. *Measures and applications of lexical distributional similarity*. Ph.D. thesis, University of Sussex.