Detecting Non-compositional MWE Components using Wiktionary

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 propose a simple unsupervised approach to detecting non-compositional components in multiword expressions based on Wiktionary. The approach makes use of the definitions, synonyms and translations in Wiktionary, and is applicable to any type of MWE in any language, assuming the MWE is contained in Wiktionary. Our experiments show that the proposed approach achieves higher F-score than state-of-the-art methods.

1 Introduction

A multiword expression (MWE) is a combination of words with lexical, syntactic or semantic idiosyncrasy (Sag et al., 2002; Baldwin and Kim, 2009). An MWE is considered (semantically) "non-compositional" when its meaning is not predictable from the meaning of its components. Conversely, compositional MWEs are those whose meaning is predictable from the meaning of the components. Based on this definition, a component is compositional within an MWE, if its meaning is reflected in the meaning of the MWE, and it is non-compositional otherwise.

Understanding which components are non-compositional within an MWE is important in NLP applications in which semantic information is required. For example, when searching for *spelling bee*, we may also be interested in documents about *spelling*, but not those which contain only *bee*. For *research project*, on the other hand, we are likely to be interested in documents which contain either *research* or *project* in isolation, and for *swan song*, we are only going to be interested in documents which contain the phrase *swan song*, and not just *swan* or *song*.

In this paper, we propose an unsupervised approach based on Wikitionary for predicting which components of a given MWE have a compositional usage. Experiments over two widely-used datasets show that our approach outperforms state-of-the-art methods.

2 Related Work

Previous studies which have considered MWE compositionality have focused on either the identification of non-compositional MWE token instances (Kim and Baldwin, 2007; Fazly et al., 2009; Forthergill and Baldwin, 2011; Muzny and Zettlemoyer, 2013), or the prediction of the compositionality of MWE types (Reddy et al., 2011; Salehi and Cook, 2013; Salehi et al., 2014). The identification of non-compositional MWE tokens is an important task when a word combination such as *kick the bucket* or *saw logs* is ambiguous between a compositional (generally non-MWE) and non-compositional MWE usage. Approaches have ranged from the unsupervised learning of type-level preferences (Fazly et al., 2009) to supervised methods specific to particular MWE constructions (Kim and Baldwin, 2007) or applicable across multiple constructions using features similar to those used in all-words word sense disambiguation (Forthergill and Baldwin, 2011; Muzny and Zettlemoyer, 2013). The prediction of the compositionality of MWE types has traditionally been couched as a binary classification task (compositional or non-compositional: Baldwin et al. (2003), Bannard (2006)), but more recent work has moved towards a regression setup, where the degree of the compositionality is predicted on a continuous scale (Reddy et al., 2011; Salehi and Cook, 2013; Salehi et al., 2014). In either case, the modelling has been done either over the whole MWE (Reddy et al., 2011; Salehi and Cook, 2013), or relative to each component within the MWE (Baldwin et al., 2003; Bannard, 2006). In this paper, we focus on the binary classification of MWE types relative to each component of the

MWE.

The work that is perhaps most closely related to this paper is that of Salehi and Cook (2013) and Salehi et al. (2014), who use translation data to predict the compositionality of a given MWE relative to each of its components, and then combine those scores to derive an overall compositionality score. In both cases, translations of the MWE and its components are sourced from PanLex (Baldwin et al., 2010; Kamholz et al., 2014), and if there is greater similarity between the translated components and MWE in a range of languages, the MWE is predicted to be more compositional. The basis of the similarity calculation is unsupervised, using either string similarity (Salehi and Cook, 2013) or distributional similarity (Salehi et al., 2014). However, the overall method is supervised, as training data is used to select the languages to aggregate scores across for a given MWE construction. To benchmark our method, we use two of the same datasets as these two papers, and repurpose the best-performing methods of Salehi and Cook (2013) and Salehi et al. (2014) for classification of the compositionality of each MWE component.

3 Methodology

Our basic method relies on analysis of lexical overlap between the component words and the definitions of the MWE in Wiktionary, in the manner of Lesk (1986). That is, if a given component can be found in the definition, then it is inferred that the MWE carries the meaning of that component. For example, the Wiktionary definition of *swimming pool* is "An artificially constructed pool of water used for swimming", suggesting that the MWE is compositional relative to both *swimming* and *pool*. If the MWE is not found in Wiktionary, we use Wikipedia as a backoff, and use the first paragraph of the (top-ranked) Wikipedia article as a proxy for the definition.

As detailed below, we further extend the basic method to incorporate three types of information found in Wiktionary: (1) definitions of each word in the definitions, (2) synonyms of the words in the definitions, and (3) translations of the MWEs and components.

3.1 Definition-based Similarity

The basic method uses Boolean lexical overlap between the target component of the MWE and a definition. A given MWE will often have multiple definitions, however, begging the question of how to combine across them, for which we propose the following three methods.

First Definition (FIRSTDEF): Use only the first-listed Wiktionary definition for the MWE, based on the assumption that this is the predominant sense.

All Definitions (ALLDEFS): In the case that there are multiple definitions for the MWE, calculate the lexical overlap for each independently and take a majority vote; in the case of a tie, label the component as non-compositional.

Idiom Tag (ITAG): In Wiktionary, there is facility for users to tag definitions as idiomatic.[1] If, for a given MWE, there are definitions tagged as idiomatic, use only those definitions; if there are no such definitions, use the full set of definitions.

3.2 Synonym-based Definition Expansion

In some cases, a component is not explicitly mentioned in a definition, but a synonym does occur, indicating that the definition is compositional in that component. In order to capture synonym-based matches, we optionally look for synonyms of the component word in the definition,[2] and expand our notion of lexical overlap to include these synonyms.

For example, for the MWE *china clay*, the definition is *kaolin*, which includes neither of the components. However, we find the component word *clay* in the definition for *kaolin*, as shown below.

> A fine clay, rich in kaolinite, used in ceramics, paper-making, etc.

This method is compatible with the three definition-based similarity methods described above, and indicated by the +SYN suffix (e.g. FIRSTDEF+SYN is FIRSTDEF with synonym-based expansion).

3.3 Translations

A third information source in Wiktionary that can be used to predict compositionality is sense-level translation data. Due to the user-generated nature of Wiktionary, the set of languages for which

[1] Although the recall of these tags is low (Muzny and Zettlemoyer, 2013).

[2] After removing function words, based on a stopword list.

	ENC	EVPC
WordNet	91.1%	87.5%
Wiktionary	96.7%	96.2%
Wiktionary+Wikipedia	100.0%	96.2%

Table 1: Lexical coverage of WordNet, Wiktionary and Wiktionary+Wikipedia over our two datasets.

translations are provided varies greatly across lexical entries. Our approach is to take whatever translations happen to exist in Wiktionary for a given MWE, and where there are translations in that language for the component of interest, use the LCS-based method of Salehi and Cook (2013) to measure the string similarity between the translation of the MWE and the translation of the components. Unlike Salehi and Cook (2013), however, we do not use development data to select the optimal set of languages in a supervised manner, and instead simply take the average of the string similarity scores across the available languages. In the case of more than one translation in a given language, we use the maximum string similarity for each pairing of MWE and component translation.

Unlike the definition and synonym-based approach, the translation-based approach will produce real rather than binary values. To combine the two approaches, we discretise the scores given by the translation approach. In the case of disagreement between the two approaches, we label the given MWE as non-compositional. This results in higher recall and lower precision for the task of detecting compositionality.

3.4 An Analysis of Wiktionary Coverage

A dictionary-based method is only as good as the dictionary it is applied to. In the case of MWE compositionality analysis, our primary concern is lexical coverage in Wiktionary, i.e., what proportion of a representative set of MWEs is contained in Wiktionary. We measure lexical coverage relative to the two datasets used in this research (described in detail in Section 4), namely 90 English noun compounds (ENCs) and 160 English verb particle constructions (EVPCs). In each case, we calculated the proportion of the dataset that is found in Wiktionary, Wiktionary+Wikipedia (where we back off to a Wikipedia document in the case that a MWE is not found in Wiktionary) and WordNet (Fellbaum, 1998). The results are found in Table 1, and indicate perfect coverage in Wiktionary+Wikipedia for the ENCs, and very high coverage for the EVPCs. In both cases, the coverage of WordNet is substantially lower, although still respectable, at around 90%.

4 Datasets

As mentioned above, we evaluate our method over the same two datasets as Salehi and Cook (2013) (which were later used, in addition to a third dataset of German noun compounds, in Salehi et al. (2014)): (1) 90 binary English noun compounds (ENCs, e.g. *spelling bee* or *swimming pool*); and (2) 160 English verb particle constructions (EVPCs, e.g. *stand up* and *give away*). Our results are not directly comparable with those of Salehi and Cook (2013) and Salehi et al. (2014), however, who evaluated in terms of a regression task, modelling the overall compositionality of the MWE. In our case, the task setup is a binary classification task relative to each of the two components of the MWE.

The ENC dataset was originally constructed by Reddy et al. (2011), and annotated on a continuous $[0, 5]$ scale for both overall compositionality and the component-wise compositionality of each of the modifier and head noun. The sampling was random in an attempt to make the dataset balanced, with 48% of compositional English noun compounds, of which 51% are compositional in the first component and 60% are compositional in the second component. We generate discrete labels by discretising the component-wise compositionality scores based on the partitions $[0, 2.5]$ and $(2.5, 5]$. On average, each NC in this dataset has 1.4 senses (definitions) in Wiktionary.

The EVPC dataset was constructed by Bannard (2006), and manually annotated for compositionality on a binary scale for each of the head verb and particle. For the 160 EVPCs, 76% are verb-compositional and 48% are particle-compositional. On average, each EVPC in this dataset has 3.0 senses (definitions) in Wiktionary.

5 Experiments

The baseline for each dataset takes the form of looking for a user-annotated idiom tag in the Wiktionary lexical entry for the MWE: if there is an idiomatic tag, both components are considered to be non-compositional; otherwise, both components are considered to be compositional. We expect this method to suffer from low precision for two

Method	First Component			Second Component		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	66.7	68.2	67.4	66.7	83.3	74.1
LCS	60.0	77.7	67.7	81.6	68.1	64.6
DS	62.1	88.6	73.0	80.5	86.4	71.2
DS+DSL2	62.5	92.3	74.5	78.4	89.4	70.6
LCS+DS+DSL2	66.3	87.5	75.4	82.1	80.6	70.1
FIRSTDEF	59.4	93.2	72.6	54.2	88.9	67.4
ALLDEFS	59.5	**100.0**	74.6	52.9	**100.0**	69.2
ITAG	60.3	**100.0**	75.2	54.5	**100.0**	70.6
FIRSTDEF+SYN	64.9	84.1	73.3	63.8	83.3	72.3
ALLDEFS+SYN	64.5	90.9	75.5	60.4	88.9	71.9
ITAG+SYN	64.5	90.9	75.5	61.8	94.4	**74.7**
FIRSTDEF+SYN $_{\text{COMB(LCS+DS+DSL2)}}$	**82.9**	85.3	84.1	81.9	80.0	69.8
ALLDEFS+SYN $_{\text{COMB(LCS+DS+DSL2)}}$	81.2	88.1	**84.5**	87.3	80.6	73.3
ITAG+SYN $_{\text{COMB(LCS+DS+DSL2)}}$	81.0	88.1	84.1	**88.0**	81.1	73.9

Table 2: Compositionality prediction results over the ENC dataset, relative to the first component (the modifier noun) and the second component (the head noun)

reasons: first, the guidelines given to the annotators of our datasets might be different from what Wiktionary contributors assume to be an idiom. Second, the baseline method assumes that for any non-compositional MWE, all components must be equally non-compositional, despite the wealth of MWEs where one or more components are compositional (e.g. from the Wiktionary guidelines for idiom inclusion,[3] *computer chess*, *basketball player*, *telephone box*).

We also compare our method with: (1) "LCS", the string similarity-based method of Salehi and Cook (2013), in which 54 languages are used; (2) "DS", the monolingual distributional similarity method of Salehi et al. (2014); (3) "DS+DSL2", the multilingual distributional similarity method of Salehi et al. (2014), including supervised language selection for a given dataset, based on cross-validation; and (4) "LCS+DS+DSL2", whereby the first three methods are combined using a supervised support vector regression model. In each case, the continuous output of the model is equal-width discretised to generate a binary classification. We additionally present results for the combination of each of the six methods proposed in this paper with LCS, DS and DSL2, using a linear-kernel support vector machine (represented with the suffix "$_{\text{COMB(LCS+DS+DSL2)}}$" for a given method). The results are based on cross-validation, and for direct comparability, the partitions are exactly the same as Salehi et al. (2014).

Tables 2 and 3 provide the results when our proposed method for detecting non-compositionality is applied to the ENC and EVPC datasets, respectively. The inclusion of translation data was found to improve all of precision, recall and F-score across the board for all of the proposed methods. For reasons of space, results without translation data are therefore omitted from the paper.

Overall, the simple unsupervised methods proposed in this paper are comparable with the unsupervised and supervised state-of-the-art methods of Salehi and Cook (2013) and Salehi et al. (2014), with ITAG achieving the highest F-score for the ENC dataset and for the verb components of the EVPC dataset. The inclusion of synonyms boosts results in most cases.

When we combine each of our proposed methods with the string and distributional similarity methods of Salehi and Cook (2013) and Salehi et al. (2014), we see substantial improvements over the comparable combined method of "LCS+DS+DSL2" in most cases, demonstrating both the robustness of the proposed methods and their complementarity with the earlier methods. It is important to reinforce that the proposed methods make no language-specific assumptions and are therefore applicable to any type of MWE and any language, with the only requirement being that the MWE of interest be listed in the Wiktionary for

[3] http://en.wiktionary.org/wiki/Wiktionary:Idioms_that_survived_RFD

Method	First Component			Second Component		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	24.6	36.8	29.5	59.6	40.5	48.2
LCS	36.5	49.2	39.3	61.5	63.7	60.3
DS	32.8	34.1	33.5	80.9	19.6	29.7
DS+DSL2	31.8	72.4	44.2	74.8	27.5	36.6
LCS+DS+DSL2	36.1	62.6	45.8	77.9	42.8	49.2
FIRSTDEF	24.8	84.2	38.3	54.5	94.0	69.0
ALLDEFS	25.0	**97.4**	39.8	53.6	**97.6**	**69.2**
ITAG	26.2	89.5	40.5	54.6	91.7	68.4
FIRSTDEF+SYN	32.9	65.8	43.9	60.4	65.5	62.9
ALLDEFS+SYN	28.4	81.6	42.1	62.5	77.4	69.1
ITAG+SYN	30.5	65.8	41.7	57.8	61.9	59.8
FIRSTDEF+SYN $_{\text{COMB(LCS+DS+DSL2)}}$	34.0	65.3	44.7	**83.6**	67.3	65.4
ALLDEFS+SYN $_{\text{COMB(LCS+DS+DSL2)}}$	**37.4**	70.9	**48.9**	80.4	65.9	63.0
ITAG+SYN $_{\text{COMB(LCS+DS+DSL2)}}$	35.6	70.9	47.4	83.5	64.9	64.2

Table 3: Compositionality prediction results over the EVPC dataset, relative to the first component (the head verb) and the second component (the particle)

that language.

6 Error Analysis

We analysed all items in each dataset where the system score differed from that of the human annotators. For both datasets, the majority of incorrectly-labelled items were compositional but predicted to be non-compositional by our system, as can be seen in the relatively low precision scores in Tables 2 and 3. In many of these cases, the prediction based on definitions and synonyms was compositional but the prediction based on translations was non-compositional. In such cases, we arbitrarily break the tie by labelling the instance as non-compositional, and in doing so favour recall over precision.

Some of the incorrectly-labelled ENCs have a gold-standard annotation of around 2.5, or in other words are semi-compositional. For example, the compositionality score for *game* in *game plan* is 2.82/5, but our system labels it as non-compositional; a similar thing happens with *figure* and the EVPC *figure out*. Such cases demonstrate the limitation of approaches to MWE compositionality that treat the problem as a binary classification task.

On average, the EVPCs have three senses, which is roughly twice the number for ENCs. This makes the prediction of compositionality harder, as there is more information to combine across (an effect that is compounded with the addition of synonyms and translations). In future work, we hope to address this problem by first finding the sense which matches best with the sentences given to the annotators.

7 Conclusion

We have proposed an unsupervised approach for predicting the compositionality of an MWE relative to each of its components, based on lexical overlap using Wiktionary, optionally incorporating synonym and translation data. Our experiments showed that the various instantiations of our approach are superior to previous state-of-the-art supervised methods. All code to replicate the results in this paper has been made publicly available at `https://github.com/bsalehi/wiktionary_MWE_compositionality`.

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

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Dam-

erau, 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.

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

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Richard Forthergill and Timothy Baldwin. 2011. Fleshing it out: A supervised approach to MWE-token and MWE-type classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 911–919, Chiang Mai, Thailand.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland.

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.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Ontario, Canada.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, USA.

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. *SEM, Proc. of the Main Conference and the Shared Task: Semantic Textual Similarity*, 1:266–275.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality.