Detecting Compositionality in Multi-Word Expressions

Ioannis Korkontzelos
Department of Computer Science
The University of York
Heslington, York, YO10 5NG, UK
johnkork@cs.york.ac.uk

Suresh Manandhar
Department of Computer Science
The University of York
Heslington, York, YO10 5NG, UK
suresh@cs.york.ac.uk

Abstract
Identifying whether a multi-word expression (MWE) is compositional or not is important for numerous NLP applications. Sense induction can partition the context of MWEs into semantic uses and therefore aid in deciding compositionality. We propose an unsupervised system to explore this hypothesis on compound nominals, proper names and adjective-noun constructions, and evaluate the contribution of sense induction. The evaluation set is derived from WordNet in a semi-supervised way. Graph connectivity measures are employed for unsupervised parameter tuning.

1 Introduction and related work
Multi-word expressions (MWEs) are sequences of words that tend to co-occur more frequently than chance and are either idiosyncratic or decomposable into multiple simple words (Baldwin, 2006). Deciding idiomaticity of MWEs is highly important for machine translation, information retrieval, question answering, lexical acquisition, parsing and language generation.

Compositionality refers to the degree to which the meaning of a MWE can be predicted by combining the meanings of its components. Unlike syntactic compositionality (e.g. by and large), semantic compositionality is continuous (Baldwin, 2006).

In this paper, we propose a novel unsupervised approach that compares the major senses of a MWE and its semantic head using distributional similarity measures to test the compositionality of the MWE. These senses are induced by a graph based sense induction system, whose parameters are estimated in an unsupervised manner exploiting a number of graph connectivity measures (Korkontzelos et al., 2009). Our method partitions the context space and only uses the major senses, filtering out minor senses. In our approach the only language dependent components are a PoS tagger and a parser.

There are several studies relevant to detecting compositionality of noun-noun MWEs (Baldwin et al., 2003) verb-particle constructions (Bannard et al., 2003; McCarthy et al., 2003) and verb-noun pairs (Katz and Griesbrecht, 2006). Datasets with human compositionality judgements are available for these MWE categories (Cook et al., 2008). Here, we focus on compound nominals, proper names and adjective-noun constructions.

Our contributions are threefold: firstly, we experimentally show that sense induction can assist in identifying compositional MWEs. Secondly, we show that unsupervised parameter tuning (Korkontzelos et al., 2009) results in accuracy that is comparable to the best manually selected combination of parameters. Thirdly, we propose a semi-supervised approach for extracting non-compositional MWEs from WordNet, to decrease annotation cost.

2 Proposed approach
Let us consider the non-compositional MWE “red carpet”. It mainly refers to a strip of red carpeting laid down for dignitaries to walk on. However, it is possible to encounter instances of “red carpet” referring to any carpet of red colour. Our method first applies sense induction to identify the major semantic uses (senses) of a MWE (“red carpet”) and its semantic head (“carpet”). Then, it compares these uses to decide MWE compositionality. The more diverse these uses are, the more possibly the MWE is non-compositional. Our algorithm consists of 4 steps:

A. Corpora collection and preprocessing. Our approach receives as input a MWE (e.g. “red carpet”). The dependency output of Stanford Parser (Klein and Manning, 2003) is used to locate the
MWE semantic head. Two different corpora are collected (for the MWE and its semantic head). Each consists of webtext snippets of length 15 to 200 tokens in which the MWE/semantic head appears. Given a MWE, a set of queries is created: All synonyms of the MWE extracted from WordNet are collected\(^1\). The MWE is paired with each synonym to create a set of queries. For each query, snippets are collected by parsing the web-pages returned by Yahoo!. The union of all snippets produces the MWE corpus. The corpus for a semantic head is created equivalently.

To keep the computational time reasonable, only the longest 3, 000 snippets are kept from each corpus. Both corpora are PoS tagged (GENIA tagger). In common with Agirre et al. (2006), only nouns are kept and lemmatized, since they are more discriminative than other PoS.

**B. Sense Induction** methods can be broadly divided into vector-space models and graph based models. Sense induction methods are evaluated under the SemEval-2007 framework (Agirre and Soroa, 2007). We employ the collocational graph-based sense induction of Klapaftis and Manandhar (2008) in this work (henceforth referred to as KM). The method consists of 3 stages:

**Corpus preprocessing** aims to capture nouns that are contextually related to the target MWE/sem head. Log-likelihood ratio (\(G^2\)) (Dunning, 1993) with respect to a large reference corpus, \(WebIT\ 5\)-gram Corpus (Brants and Franz, 2006), is used to capture the contextually relevant nouns. \(P_1\) is the \(G^2\) threshold below which nouns are removed from corpora.

**Graph creation.** A collocation is defined as a pair of nouns cooccurring within a snippet. Each noun within a snippet is combined with every other, generating \(\binom{n}{2}\) collocations. Each collocation is represented as a weighted vertex. \(P_2\) thresholds collocation frequencies and \(P_3\) collocation weights. Weighted edges are drawn based on cooccurrence of the corresponding vertices in one or more snippets (e.g. \(w_8\) and \(w_{7,9}\), fig. 1). In contrast to KM, frequencies for weighting vertices and edges are obtained from Yahoo! web-page counts to deal with data sparsity.

**Graph clustering** uses Chinese Whispers\(^2\) (Biemann, 2006) to cluster the graph. Each cluster now represents a sense of the target word.

KM produces larger number of clusters (uses) than expected. To reduce it we exploit the one sense per collocation property (Yarowsky, 1995). Given a cluster \(I_i\), we compute the set \(S_i\) of snippets that contain at least one collocation of \(I_i\). Any clusters \(I_a\) and \(I_b\) are merged if \(S_a \subseteq S_b\).

**C. Comparing the induced senses.** We used two techniques to measure the distributional similarity of major uses of the MWE and its semantic head, both based on Jaccard coefficient (J). “Major use” denotes the cluster of collocations which tags the most snippets. Lee (1999) shows that \(J\) performs better than other symmetric similarity measures such as cosine, Jensen-Shannon divergence, etc. The first is \(J_c = J(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}\) where \(A, B\) are sets of collocations. The second, \(J_{sn}\), is based on the snippets that are tagged by the induced uses. Let \(K_i\) be the set of snippets in which at least one collocation of the use \(i\) occurs. \(J_{sn} = J(K_j, K_k)\), where \(j, k\) are the major uses of the MWE and its semantic head, respectively.

**D. Determining compositionality.** Given the major uses of a MWE and its semantic head, the MWE is considered as compositional, when the corresponding distributional similarity measure (\(J_c\) or \(J_{sn}\)) value is above a parameter threshold, \(\text{sim}\). Otherwise, it is considered as non-compositional.

**3 Test set of MWEs**

To the best of our knowledge there are no noun compound datasets accompanied with compositionality judgements available. Thus, we developed an algorithm to aid human annotation. For each of the 52, 217 MWEs of WordNet 3.0 (Miller, 1995) we collected:

\(^1\)Thus, for “red carpet”, corpora will be collected for “red carpet” and “carpet”. The synonyms of “red carpet” are “rug”, “carpet” and “carpeting”\(^2\)Chinese Whispers is not guaranteed to converge, thus 200 was adopted as the maximum number of iterations.
Non-compositional MWEs
agony aunt, black maria, dead end, Dutch oven.
fish finger, fool's paradise, goat's rue, green light.
high jump, joint chiefs, lip service, living rock.
monkey puzzle, motor pool, prince Albert,
stocking stuffer, sweet bay, teddy boy, think tank.

Compositional MWEs
box white oak, cartridge brass, common iguana,
closed chain, eastern pipistrelle, field mushroom,
hard candy, king snake, labor camp, lemon tree,
life form, parenthetical-free notation, parking brake,
petit juror, relational adjective, taxonomic category,
telephone service, tea table, upland cotton.

Table 1: Test set with compositionality annotation. MWEs whose compositionality was successfully detected by: (a) Iclword baseline are in bold font, (b) manual parameter selection are underlined and (c) average cluster coefficient are in italics.

1. all synonyms of the MWE
2. all hypernyms of the MWE
3. sister-synsets of the MWE, within distance 3
4. synsets that are in holonymy or meronymy relation to the MWE, within distance 3

If the semantic head of the MWE is also in the above collection then the MWE is likely to be compositional, otherwise it is likely that the MWE is non-compositional.

6, 287 MWEs were judged as potentially non-compositional. We randomly chose 19 and checked them manually. Those that were compositional were replaced by other randomly chosen ones. The process was repeated until we ended up with 19 non-compositional examples. Similarly, 19 negative examples that were judged as compositional were collected (Table 1).

4 Evaluation setting and results

The sense induction component of our algorithm depends upon 3 parameters: \( P_1 \) is the \( G^2 \) threshold below which noun are removed from corpora. \( P_2 \) thresholds collocation frequencies and \( P_3 \) collocation weights. We chose \( P_1 \in \{5, 10, 15\} \), \( P_2 \in \{10^2, 10^3, 10^4, 10^5\} \) and \( P_3 \in \{0.2, 0.3, 0.4\} \). For reference, \( P_1 \) values of 3.84, 6.63, 10.83 and 15.13 correspond to \( G^2 \) values for confidence levels of 95\%, 99\%, 99.9\% and 99.99\%, respectively.

To assess the performance of the proposed algorithm we compute accuracy, the percentage of MWEs whose compositionality was correctly determined against the gold standard.

We compared the system’s performance against a baseline, Iclword, that assigns the whole graph to a single cluster and no graph clustering is performed. Iclword corresponds to a relevant SemEval-2007 baseline (Agirre and Soroa, 2007) and helps in showing whether sense induction can assist determining compositionality.

Our method was evaluated for each \( \langle P_1, P_2, P_3 \rangle \) combination and similarity measures \( J_c \) and \( J_{sn} \), separately. We used our development set to determine if there are parameter values that verify our hypothesis. Given a \( sim \) value (see section 2, last paragraph), we chose the best performing parameter combination manually.

The best results for manual parameter selection were obtained for \( sim = 95\% \) giving an accuracy of 68.42\% for detecting non-compositional MWEs. In all experiments, \( J_{sn} \) outperforms \( J_c \).

With manually selected parameters, our system’s accuracy is higher than Iclword for all \( sim \) values (5% points) (fig. 2, table 1). The initial hypothesis holds; sense induction improves MWE compositionality detection.

5 Unsupervised parameter tuning

We followed Korkontzelos et al. (2009) to select the “best” parameters \( \langle P_1, P_2, P_3 \rangle \) for the collocational graph of each MWE or head word. We applied 8 graph connectivity measures (weighted and unweighted versions of average degree, cluster coefficient, graph entropy and edge density) separately on each of the clusters (resulting from the application of the Chinese whispers algorithm).

Each graph connectivity measure assigns a score to each cluster. We averaged the scores over
the clusters from the same graph. For each connectivity measure, we chose the parameter combination \( \langle P_1, P_2, P_3 \rangle \) that gave the highest score.

While manual parameter tuning chooses a single globally best set of parameters (see section 4), the graph connectivity measures generate different values of \( \langle P_1, P_2, P_3 \rangle \) for each graph.

### 5.1 Evaluation results

The best performing distributional similarity measure is \( J_{sn} \). Unweighted versions of graph connectivity measures perform better than weighted ones. Figures 3 and 4 present a comparison between the unweighted and weighted versions of all graph connectivity measures, respectively, for all \( sim \) values. Average cluster coefficient performs better or equally well to the other graph connectivity measures for all \( sim \) values (except for \( sim \in [90\%, 100\%] \)). The accuracy of average cluster coefficient is equal (68.42\%) to that of manual parameter selection (section 4, table 1). The second best performing unweighted graph connectivity measures is average graph entropy. For weighted graph connectivity measures, average graph entropy performs best, followed by average weighted clustering coefficient.

### 6 Conclusion and Future Work

We hypothesized that sense induction can assist in identifying compositional MWEs. We introduced an unsupervised system to experimentally explore the hypothesis, and showed that it holds. We proposed a semi-supervised way to extract non-compositional MWEs from WordNet. We showed that graph connectivity measures can be successfully employed to perform unsupervised parameter tuning of our system. It would be interesting to explore ways to substitute querying Yahoo! so as to make the system quicker. Experimentation with more sophisticated graph connectivity measures could possibly improve accuracy.

### References

E. Agirre and A. Soroa. 2007. Semeval-2007, task 02: Evaluating WSI and discrimination systems. In proceedings of SemEval-2007. ACL.

E. Agirre, D. Martinez, O. de Lacalle, and A. Soroa. 2006. Two graph-based algorithms for state-of-the-art WSD. In proceedings of EMNLP-2006. ACL.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of MWE decomposability. In proceedings of the MWE workshop. ACL.

T. Baldwin. 2006. Compositionality and MWEs: Six of one, half a dozen of the other? In proceedings of the MWE workshop. ACL.

C. Bannard, T. Baldwin, and A. Lascarides. 2003. A statistical approach to the semantics of verb-particles. In proceedings of the MWE workshop. ACL.

C. Biemann. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to NLP problems. In proceedings of TextGraphs. ACL.

T. Brants and A. Franz. 2006. Web 1 5-gram corpus, version 1. Technical report, Google Research.

P. Cook, A. Fazly, and S. Stevenson. 2008. The VNC-Token Dataset. In proceedings of the MWE workshop. ACL.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional MWEs using latent semantic analysis. In proceedings of the MWE workshop. ACL.

I. P. Klapaftis and S. Manandhar. 2008. WSI using graphs of collocations. In proceedings of ECAI-2008.

D. Klein and C. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In proceedings of NIPS 15. MIT Press.

I. Korkontzelos, I. Klapaftis, and S. Manandhar. 2009. Graph connectivity measures for unsupervised parameter tuning of graph-based sense induction systems. In proceedings of the UMSLLS Workshop, NAACL HLT 2009.

L. Lee. 1999. Measures of distributional similarity. In proceedings of ACL.

D. McCarthy, B. Keller, and J. Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In proceedings of the MWE workshop. ACL.

G. A. Miller. 1995. WordNet: a lexical database for English. ACM, 38(11):39–41.

D. Yarowsky. 1995. Unsupervised WSD rivaling supervised methods. In proceedings of ACL.