A Multiword Expression Data Set: Annotating Non-Compositionality and Conventionalization for English Noun Compounds

Meghdad Farahmand
University of Geneva
Geneva, Switzerland

Aaron Smith
Uppsala University
Uppsala, Sweden

Joakim Nivre
Uppsala University
Uppsala, Sweden

firstname.lastname@unige.ch firstname.lastname@lingfil.uu.se

Abstract

Scarcity of multiword expression data sets raises a fundamental challenge to evaluating the systems that deal with these linguistic structures. In this work we attempt to address this problem for a subclass of multiword expressions by producing a large data set annotated by experts and validated by common statistical measures. We present a set of 1048 noun-noun compounds annotated as non-compositional, compositional, conventionalized and not conventionalized. We build this data set following common trends in previous work while trying to address some of the well known issues such as small number of annotated instances, quality of the annotations, and lack of availability of true negative instances.

1 Introduction

The lack of practical data sets that can be used in the training and evaluation of multiword expression (MWE) related systems is a notorious problem (McCarthy et al., 2003; Hermann et al., 2012). It is partly due to the heterogeneous nature of MWEs, partly due to their frequency, and partly due to the unclear boundaries between MWEs and regular phrases. These issues have made the compilation of useful MWE data sets challenging, and any effort to create them invaluable.

In this work we present a data set of two-word English noun-noun compounds which are annotated for two properties: non-compositionality and conventionalization. Although non-compositionality can apply at different levels, from syntactic to semantic, by non-compositionality we strictly mean semantic non-compositionality. Semantic non-compositionality in simple terms is the property of a compound whose meaning can not be readily interpreted from the meanings of its components.

Conventionalization meanwhile refers to the situation where a sequence of words that refer to a particular concept is commonly accepted in such a way that its constituents cannot be substituted for their near-synonyms, according to some cultural or historical convention. Conventionalization can also be referred to as institutionalization or statistical idiosyncrasy (Sag et al., 2002), and is closely related to the concept of collocation (Baldwin and Kim, 2010). Conventionalization is a very broad concept and can apply to a wide range of compounds. Although a large fraction of compounds are to some extent conventionalized, we are interested in and annotate only clear and well-known conventionalizations, which we refer to as "marked conventionalization". For instance, although *exit door* and *floor space* have some elements of conventionalization, this property is more conspicuous in *weather forecast*, *car wash*, and *traffic light*. We assume that non-compositional compounds are by definition conventionalized and annotate this property only when a compound is compositional.

Our data set comprises 1048 compounds which are annotated with binary decisions about whether they are (i) non-compositional and (ii) conventionalized. Although non-compositionality can be a grey area and a non-binary decision may be more precise, eventually this decision must be reduced to a binary

Proceedings of NAACL-HLT 2015, pages 29–33,
Denver, Colorado, May 31 – June 5, 2015.

one: whether or not a compound should be lexicalized due to its non-compositionality.

The main contributions of this work can be described as follows: coverage for two major properties of MWEs (non-compositionality and conventionalization); providing both positive and negative instances of non-compositional and conventionalized classes, allowing the evaluation of MWE identification/extraction systems in terms of both true positive and true negative rates; incorporating a larger number of annotated instances compared to related data sets.

2 Related work

The most important related work is that of Reddy et al. (2011), which provides 90 compounds with a mean compositionality score between 0 and 5. They acquired their annotations using Amazon Mechanical Turk from 30 turkers. They detect and discard poor annotations using Spearman Coefficient Correlation. The number of instances in their final data set, however, might not be enough for evaluation purposes. Moreover, it might not be a trivial task to adapt an identification/extraction system to produce a similar non-compositionality ranking. Korkontzelos and Manandhar (2009) present a data set that comprises 19 non-compositional and 19 compositional instances. In this work the size of the data set is small and compound selection process and the rationale behind decisions about non-compositionality is not expounded. Other related but slightly different works are Biemann and Giesbrecht (2011) who present a set of adjective-noun, verb-subject, and verb-object pairs and their non- compositionality judgments, and McCarthy et al. (2003) who present a set of 116 phrasal verbs and rank their non-compositionality between 0 and 9.

Data sets that incorporate conventionalization are rather difficult to come by. The closest are collocation sets which are also scarce in their own right. Most collocation sets that we could find were either commercial or not publicly available. Moreover, since collocation can refer to a wide range of MWEs and human agreement on statistical idiosyncrasy is not high enough, it is hard to find an annotated collocation set. Instead, extraction systems have been used to automatically produce such sets and the outcomes have been commonly evaluated by either manual evaluation (Smadja, 1993), or by ranking the collocation candidates and calculating precision and recall of the extraction system for the set of n highest ranking candidates (Evert, 2005).

Schneider et al. (2014) is another related work in which generic MWEs are annotated in a 55K-word English web corpus. Their work covers a broad range of "multiword phenomena" with emphasis on heterogeneity, gappy grouping and expression strength which represents the level of idiomaticity of a MWE. They build a corpus of MWEs without restricting themselves to any syntactic categories and they argue that this can to some extent address the problem of heterogeneity of MWEs.

3 Data Preparation

We downloaded English Wikipedia, removed the tags and segmented it into sentences. We then filtered very short and very long sentences, sentences which were not in English, and sentences which contained only numbers and non-alphanumeric characters. This resulted in a clean corpus with 24 million sentences (512 million words). We tagged this corpus using Stanford POS tagger and extracted a list of distinct contiguous noun-noun pairs ($\approx$ 2.6 million) and their frequencies. We filtered out low frequency pairs by removing the pairs whose frequency of occurrence in the corpus was below 10. This led to a set of 169,000 pairs (`filtered_list` hereafter). We divided this set into 5 frequency classes and randomly extracted 250 pairs from each of those frequency classes (`selected_list` hereafter) in line with McCarthy et al. (2003). Frequency classes were chosen in a way that each class holds approximately the same number of pairs.

Compositional compounds tend to be much more frequent than non-compositional ones: this might lead the data set to be inundated with compositional compounds. To mitigate this problem we asked two experts with backgrounds in corpus linguistics to each provide us with 50[1] examples that they thought were partly or fully non-compositional. These examples were mainly extracted from two non-overlapping random divisions

[1]The choice of this number was made taking into account our time and financial constraints.

of `filtered_list`, whilst also ensuring that there was no overlap with `selected_list`. Furthermore, the experts were provided with, and allowed to extract the examples from a set of frequent adjective-noun pairs which incorporate a relatively large number of non-compositional compounds such as *hard disk* and *big shot*. These 100 examples were then added to `selected_list`. The linguists who performed this selection did not participate in the annotation task.

Finally, we manually removed pairs with foreign or inappropriate/offensive words, those with incorrect POS tags, and the few pairs used to help describing the task to the annotators (see Section 4), from `selected_list`. We also removed those pairs for which a unified form was more common in the corpus (e.g. ice berg, paper work and life style for which iceberg, paperwork and lifestyle were more frequently occurring).

4 Annotations

We assigned the annotation task to three native and two non-native but fluent speakers of English. We chose to hire experts to perform the annotation task rather than using crowd-sourcing systems such as Amazon Mechanical Turk, where the results can be flawed for various reasons including scammers and low quality of the annotations (Biemann and Giesbrecht, 2011; Reddy et al., 2011). All of our annotators had advanced knowledge of English grammar and the majority had a background in linguistics. We provided the annotators with a detailed set of instructions about non-compositionality and conventionalization. The instructions were extensively exemplified by examples from Reddy et al. (2011), Hermann et al. (2012) and Baldwin and Kim (2010).

For each compound, we asked the annotators to make binary decisions about non-compositionality and marked conventionalization. We explained non-compositionality as being the property of compounds whose meanings cannot be readily interpreted from the meaning of their constituents. The annotators were asked to use the label 0 when they thought a compound was more compositional than non-compositional, and 1 when they thought the compound was more non-compositional than compositional.

Conventionalization, meanwhile, was defined as the main property of compounds that are collocational and whose constituents co-occur more than expected by chance. We introduced the annotators to the non-substitutability test which can help to decide if a compound is conventionalized: if neither of the constituents of the word pair can be substituted for their near synonyms (Manning and Schütze, 1999) we have a conventionalization. Taking *weather forecast* as an example, although *weather prediction* and *climate forecast* are syntactically correct and semantically plausible alternatives, they are not considered proper English compounds. The non-substitutability test often fails in compounds with less noticeable conventionalization; for instance we can still say *exit gate* instead of *exit door* or *floor area* instead of *floor space*. Identifying conventionalization is not a trivial task and human agreement on this property can be relatively low (Krenn et al., 2001). Therefore, we emphasized that we were only interested in marked conventionalization and that this property should be annotated only when the annotator was certain about its presence.

We asked the annotators to make decisions about marked conventionalization only when they annotated a compound as compositional: we assumed that non-compositional compounds are by definition conventionalized. In practice however, in order to avoid overestimated scores and loose overall judgements, we do not regard conventionalization based on non-compositionality and conventionalization annotated on a compositional compound as equal. Instead we define a third label X and assign it to the marked conventionalization field whenever a compound is annotated as non-compositional. This means the marked conventionalization field in fact has three possible labels (0, 1, and X). Throughout the paper, the scores and data set statistics for marked conventionalization are calculated based on these three labels. Nevertheless, the user of the data set retains the option of merging X and 1 and benefiting from a larger set of markedly conventionalized instances for particular tasks.

5 Validation of the Annotations

To ensure that the annotations are sound and in order to eradicate possible problems caused by human er-

ror, we calculated Spearman Correlation Coefficient (ρ) between all the annotations and took the average Spearman ρ for each annotator. This was done separately for non-compositionality and marked conventionalization. The results are shown in Table 1.

	average ρ (non-comp.)	average ρ (marked conv.)
annotator1	0.58	0.60
annotator2	0.34	0.46
annotator3	0.52	0.57
annotator4	0.54	0.63
annotator5	0.59	0.64

Table 1: The average Spearman ρ for non-compositionality and conventionalization.

We used Spearman ρ as a means to filter the less reliable annotations (Reddy et al., 2011) by discarding the annotations that had an average Spearman ρ of below 0.50. This left us with four sets of annotations for each property.

6 Inter-Annotator Agreement

We calculated inter-annotator agreement in terms of Fleiss' kappa between the four remaining annotations. A summary of Fleiss' kappa scores and their interpretation according to Landis and Koch (1977) is presented in Table 2.

	non-comp.	marked conv.
Fleiss's kappa	0.62	0.55
kappa error	0.012	0.009
interpretation	substantial agreement	moderate agreement

Table 2: Inter-annotator agreement in terms of Fleiss' kappa for non-compositionality and conventionalization.

The observered moderate agreement on conventionalization is consistent with the findings of Krenn et al. (2001), and in accordance with our claim that conventionalization can be more difficult than non-compositionality for humans to distinguish.

7 Results

Our final data set contains a list of 1048 compounds and, for each compound, four judgements about non-compositionality and four judgments about marked conventionalization. Essentially, our data set consists of three classes of compounds: (i) non-compositional (ii) compositional but markedly conventionalized (iii) compositional and non-conventionalized. These three classes can be described as follows in the context of training and evaluation tasks: (i) positive instances of non-compositional compounds (ii) negative instances of non-compositional but positive instances of conventionalized compounds, and (iii) negative instances of both previous types. We make the data set available as a set of compounds and (2×4) judgments for each (`raw_dataset` hereafter). `raw_dataset` can be used in various formats. For instance we generated a set of compounds that were judged to be non-compositional and conventionalized based on the decision of the majority (3 or more out of 4) and extracted several examples of different classes which are presented in Table 3.

non-compositional	**compositional but conventionalized**	**compositional and not conventionalized**
battle cry	bulletin board	area director
flag stop	cable car	art collection
gun dog	car chase	ankle injury
jet lag	food court	animal life
lead time	wish list	bus service
face value	speed limit	computer usage
mind map	background check	wrestling fan

Table 3: Examples of different classes of compounds that were classified based on the decision of the majority.

One can also generate a set of judgements based on the unanimous decision of the annotators. In each of these two formats, however, some good examples of MWEs are missed due to the fact that half of the annotators marked them as conventionalized due to non-compositionality (label X) while the other half marked them as conventionalized compositional nouns (label 1). One can therefore generate another format that covers such compounds. In such cases it is up to the user of the data set to decide whether they want to regard such instances as non-compositional, as solely conventionalized, or simply as an instance of an MWE. Table 4 presents the key statistics relating to the data set.

Non-Compositionality	
Annotated as non-comp. by the majority	140 (out of which 82 are unanimous)
Annotated as comp. by the majority	840 (out of which 763 are unanimous)
Annotated as comp. by half and non-comp. by the other half	62
Marked Conventionalization	
Annotated as comp. but conv. by the majority	155 (55 of which are unanimous)
Annotated as comp. and non conv. by the majority	570 (467 of which are unanimous)
Annotated as conv. by half and non-conv. by the other half	76
Other[2]	241

Table 4: Data set statistics.

8 Conclusion

We presented a data set of English noun-noun compounds which are judged for two major properties of MWEs: non-compositionality and conventionalization (statistical idiosyncrasy). The data set consists of both positive and negative instances of non-compositional and conventionalized MWEs and can effectively be used in evaluation and training of MWE identification and extraction systems. We recruited expert annotators and validated the reliability of their judgments using common statistical measures. We calculated inter-annotator agreement in terms of Fleiss' kappa, showing moderate and substantial agreements between the annotators for the two properties. The strengths of this data set are its granularity, incorporating both positive and negative instances of MWEs, and the credibility of the judgements as a result of recruiting expert annotators and using statistical validations.

[2] As mentioned before, non conv. in practice has three labels (0, 1, X). "Other" means either the compound was annotated as conv. (1) by half and non-comp. (X) by the other half, or the majority annotated these instances as non-compositional (X), however a minority annotated them as something else (0, 1), or the annotation for these instances includes all labels (0, 1, X) so that none of the labels are the majority.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool*.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 21–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefan Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 132–141. Association for Computational Linguistics.

Ioannis Korkontzelos and Suresh Manandhar. 2009. Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 65–68. Association for Computational Linguistics.

Brigitte Krenn, Stefan Evert, et al. 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Christopher D Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acqusation and Treatment*, pages 73–80.

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

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T Mordowanec, Henrietta Conrad, and Noah A Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC. Reykjavík, Iceland*.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.