A Large Challenging Dataset for Noun Compound Interpretation

Anonymous ACL submission

Abstract

The ability to discern implicit semantic relations in noun-compounds is imperative for natural language understanding, and many datasets have been released for training and evaluating such methods. Recently, most focus has been on the Tratz (2011) dataset, which is currently the only large-scale dataset that facilitates training neural methods, but Dima (2016) showed that it is solvable by memorizing typical relations of single words. We release a new dataset which is explicitly constructed to contain challenging examples. Indeed, we show that as opposed to the Tratz (2011) dataset, baseline performance on the new dataset is mediocre. Finally, we demonstrate the dataset utility by using its gold labels to improve an RTE model.

1 Introduction

Noun-compound interpretation is a task concerned with automatically determining the semantic relation that holds between the head and the modifier of a noun-compound, as in *vanilla bean* (*bean* is the source of *vanilla*) and *vanilla pudding* (*vanilla* is part-of the *pudding*). The importance of this task for many NLP applications has led to the development of many datasets for training and evaluation of noun-compound interpretation methods.

In recent years, with the shift of the research community toward neural methods, research on the task focused on the Tratz and Hovy (2010) and Tratz (2011) datasets, which are the only publicly-available datasets that are large enough for training neural methods. However, it was recently shown by Dima (2016) that the state-of-the-art performance on these datasets (around 80% accuracy) is mostly a result of overfitting to the typical relation of a single word in the noun-compound (e.g. the head *guide* always indicates the TOPIC relation, as in *travel guide*, *fishing guide*).

In this paper we present a new large dataset, complementary to Tratz (2011), which is explicitly constructed such that each head and modifier can participate in a variety of relationships. Focusing on the more challenging noun-compounds requires methods to model the relation between the head and the modifier without relying on the relation distribution of each single word. Indeed, we show that the various existing noun-compound interpretation methods fail to do so, yielding low performance on the new dataset.

To estimate the dataset utility, we show that using the gold labels helps improving the performance of a model for recognizing textual entailment. In the future, we hope that the new dataset will be used for training better noun-compound interpretation methods that would benefit this and other downstream applications.

2 Existing Noun-Compound Datasets

Table 1 displays existing datasets for noun-compound classification, that differ by several parameters detailed below.

Size. Most datasets are relatively small, ranging from several hundred to around 2,000 instances. Many have a relatively large number of classes, often with only a few instances per class, making them difficult to use with data-intensive machine learning methods. This is especially true for neural methods, that typically require a large amount of data. The only publicly-available dataset which is large enough for that purpose is Tratz (2011) (and its previous version Tratz and Hovy (2010)).

Noun-Compounds. Dima (2016) achieved 80% accuracy on the Tratz (2011) dataset using a neural network whose only features are the word embed-

Dataset	Instances	Relations
Vanderwende (1994)	395	13
Lauer (1995)	385	8
Barker and Szpakowicz (1998)	505	21
Nastase and Szpakowicz (2003)	600	31 (5)
Rosario and Hearst (2001)	1,660	45
O'Séaghdha (2007)	1,443	35 (6)
Girju (2007)	2,031	22 (8)
Kim and Baldwin (2007)	2,169	20
Tratz and Hovy (2010)	17,509	43 (10)
Tratz (2011)	19,158	37 (12)

Table 1: Existing noun-compound datasets, along with their size and number of relations (brackets: number of coarse-grained relations, if applicable).

dings of the noun-compound's constituent words. Her analysis revealed that the model memorizes which words fit the head or the modifier slot of a specific relation, e.g. the *Mr.* modifier indicates a PERSONAL_TITLE relation, regardless of the head.

Whether the majority of noun-compounds can be classified based on the distributions of single words or not, there is no doubt that in many cases this distribution is misleading. For example, most compounds with the head *oil* hold the SOURCE relation (*olive oil*, *coconut oil*, *soybean oil*...), while *cooking oil* holds the PURPOSE relation. If the existing datasets do not exhibit enough of this phenomena, systems trained on them would not improve in recognizing these difficult cases.

Relations. Previous work disagrees on the vocabulary of relations to detect and their granularity level. Some datasets are labeled with few coarse-grained labels, often following Levi's (1978) suggestion that noun-compounds are derived by deleting one of 9 predicates (i.e., cause, make, use, be, in, for, from, about). Other datasets have a much richer relation inventory. Many datasets solve this disagreement by providing two levels of annotations, where the fine-grained relations are grouped into a small number of coarse-grained relations.

3 Dataset Creation

We present `WN-Compounds`, a new large dataset of noun-compound interpretations, which intentionally consists of non-trivial noun-compounds (§3.1). We empirically chose the relation inventory (§3.2) and used expert annotations (§3.3).

3.1 Noun Compounds Selection

We extracted 14,289 lexicalized noun-compounds from WordNet (Miller, 1995), where each noun-compound consists of exactly two words mh.[1]

We required that each word m and h has at least 3 WordNet synsets, assuming that polysemous words are more likely to interact with other words in various relations, making the dataset more challenging. For instance, a single-sense word like *intergalactic* would always serve as a location modifier. Conversely, thanks to the fine-grained nature of WordNet synsets, different aspects of the same concept are often split into multiple synsets. For example, *coffee bean* (where *coffee* belongs to the coffee bean synset) holds a SOURCE relation, while *coffee cup* (where *coffee* belongs to the coffee beverage synset) holds a PURPOSE relation.

3.2 Relation Inventory

Similarly to Tratz and Hovy (2010), we decided on the relation inventory empirically. The goal was a relatively coarse-grained relation inventory, but one which can separate the extracted noun-compounds well. To that end, a sample of 250 noun-compounds was annotated by each of the authors without pre-defining a relation inventory, and the two proposed relation inventories were then consolidated. Our final relation inventory consists of 12 relations, and is described in Table 2.

3.3 Annotation

The dataset was annotated by 5 trained linguists. The first 500 compounds were annotated by all the annotators, who suggested changes to the relation inventory. After incorporating their feedback, the rest of the noun-compounds were annotated by a single annotator each.

The annotation was performed using a dedicated tool in which compounds are assigned into their relation group, where each group contains several examples. The relation groups were ordered such that the INSTANCE-OF relation was last, and the annotators were instructed to label compounds to this relation only if they do not fit into any other relation, as this relation is somewhat more general and can include compounds that fit into other, more specific relations.

Since all noun-compounds in the dataset are lexicalized, we followed the strict definition of "non-compositional", instructing the annotators to choose this label only when there is no straightforward way to combine the word meanings, as in *horse radish*, which is unrelated to *horse*. This differs from Tratz (2011), whose "lexicalized" relation consists of both strictly non-compositional

[1] Multi-word expressions stored in the lexicon as a single lexical unit. The criteria to discern those from free combinations of words are discussed in Zgusta (1967).

relation	description	examples	instances
ATTRIBUTE-1	$[w_1]$ $[w_2]$ is a $[w_2]$ which has a property $[w_1]$ or resembles $[w_1]$	*thick skin, scorpion fish*	2951 (22.77%)
PURPOSE	$[w_1]$ $[w_2]$ is a $[w_2]$ which is used by or exists for $[w_1]$	*baby oil, service door*	2585 (19.95%)
INSTANCEOF	$[w_1]$ $[w_2]$ is a specific instance or a subclass of $[w_2]$, and the relation between $[w_1]$ and $[w_2]$ is none of the above	*cayenne pepper, legal age*	2371 (18.30%)
NONCOMPOSITIONAL	$[w_1]$ and $[w_2]$ do not combine in a straightforward way	*baby sitting, horse radish*	883 (6.81%)
PARTOF-1	$[w_1]$ $[w_2]$ is a $[w_2]$ is contained or located in $[w_1]$	*field game, breast cancer*	827 (6.38%)
PARTOF-2	$[w_1]$ $[w_2]$ is a $[w_2]$ that contains $[w_1]$	*water tank, olive family*	756 (5.83%)
SOURCE-1	$[w_1]$ $[w_2]$ is a $[w_2]$ whose source is $[w_1]$	*olive oil, roman law*	655 (5.05%)
MEANS	$[w_1]$ $[w_2]$ is a $[w_2]$ performed using $[w_1]$ as instrument/means	*video recording, combination lock*	517 (3.99%)
TOPIC	$[w_1]$ $[w_2]$ is a $[w_2]$ whose topic is $[w_1]$	*wanted poster, ethnic joke*	407 (3.14%)
SOURCE-2	$[w_1]$ $[w_2]$ is a $[w_2]$ that provides $[w_1]$	*flower girl, oil well*	367 (2.83%)
ATTRIBUTE-2	$[w_1]$ $[w_2]$ is an attribute related to $[w_1]$	*price level, blood type*	346 (2.67%)
OWNER	$[w_1]$ $[w_2]$ is a $[w_2]$ owned or experienced by $[w_1]$	*team spirit, national park*	293 (2.26%)

Table 2: WN-Compounds relations, along with their descriptions, examples, and frequency.

Dataset	Frequency	Distributional	Compositional	Feature-based
Tratz (2011) - fine	0.533	0.697	0.699	0.745
Tratz (2011) - coarse	0.595	0.751	0.743	0.759
WN-Compounds	0.339	0.439	0.421	0.418

Table 3: Baseline results (in terms of F_1) on the WN-Compounds and Tratz (2011) datasets.

compounds (*coach potato*) and highly-lexicalized but compositional noun-compounds (*pay phone*).

Finally, we removed relations with too few instances, remaining with 12,958 noun-compounds, distributed across relations as described in Table 2. The common annotations yielded moderate levels of agreement with averaged pairwise Kappa κ = 0.47 (Davies and Fleiss, 1982). We further elaborate on the agreement in Section 5.1.

4 Baseline Results

We report baseline results on WN-Compounds and on the Tratz (2011) dataset for comparison. v_m and v_h denote the word embeddings of a noun-compound mh, and k is the number of relations.

4.1 Baselines

Distributional (Dima, 2016). A neural network trained on the single embeddings concatenation: $\vec{x} = [v_m; v_h]$. The network contains a hidden layer $\vec{h} = \text{relu}(W_1 \cdot \vec{x})$ ($W_1 \in \mathcal{R}^{|x| \times d}$, d is tunable), outputting the highest scoring relation: $r = \text{argmax}_i \, \text{softmax}(\tan(W_2 \cdot \vec{h}))[i]$ ($W_2 \in \mathcal{R}^{d \times k}$).

Compositional (Dima, 2016). A set of neural networks: $r = \text{argmax}_i \, \text{softmax}(\tan(W \cdot \text{relu}(\vec{x})))[i]$ ($W \in \mathcal{R}^{|x| \times k}$), whose input x is some combination of v_h, v_m, and a vector representing mh, learned either with the Full-Additive (Zanzotto et al., 2010) or the Matrix compositional representations (Socher et al., 2012).

To learn the compositional representations, we obtain 300-dimensional distributional vectors for noun-compounds using the GloVe algorithm (Pennington et al., 2014) on the concatenation of English Wikipedia and the Gigaword corpus,[2] by replacing the occurrences of noun-compounds from Tratz (2011) and WN-Compounds with single tokens (e.g. *flower_girl*).[3]

Feature-based (Tratz and Hovy, 2010). A re-implementation of their classifier with a subset of the more successful features from WordNet, Roget's Thesaurus, and corpus n-grams.[4]

Frequency. We report the more successful baseline among assigning the most common relation in the train set for compounds with the same modifier (mh'), or for those with the same head ($m'h$).

4.2 Evaluation Settings

We split each dataset randomly to 70% train, 25% test and 5% validation, and use the validation set to tune hyper-parameters: d in the distributional approach, the composition function in the compositional approach, and the classifier (logistic regression or SVM), value of SVM's C parameter, and number of most useful features (according to the χ^2 measure) in the feature-based approach.

4.3 Results

Table 3 shows the baseline performance on the datasets. As opposed to the two versions of the Tratz (2011) dataset, where the baselines F_1 score are above 70, on the WN-Compounds dataset the best baseline achieves F_1 of 44, leaving plenty of room for improvement. We conjecture that this stems from the non-trivial noun-compounds in this dataset. Moreover, this reaffirms the conclusions of Dima (2016), that the methods mostly learn to rely on the head or the modifier, rather than recognizing the relation between them.

[2] https://catalog.ldc.upenn.edu/ldc2011t07

[3] We use the resulting *word* embeddings to represent also v_h and v_m in the classification models.

[4] Following the ablation tests in Tratz and Hovy (2010) and using English Wikipedia.

premise:	Obviously, if all members of the two boards were to pledge $1,000 per year, we would be most grateful and we would have a great story to tell.
hypothesis:	If every board member donated $1000 every year we would be grateful.
revised hypothesis:	If every member, *which is part of board*, donated $1000 every year we would be grateful.

Table 4: An example from the mismatched validation set, where explicitly specifying the relation in a noun-compound correctly changed the classification from contradiction to entailment.

Noun-compounds	Possible Relations
coin box, coal house	PARTOF-2, PURPOSE
water pore, oil rigger	SOURCE-2, PURPOSE
common louse, satinwood tree	INSTANCEOF, ATTRIBUTE-1

Table 5: Examples of noun-compounds on which annotators disagreed along with suggested labels.

5 Analysis

Is it possible that WN-Compounds is more challenging for the baseline methods simply because the labels are noisy? In this section, we investigate this possibility more deeply: we discuss the annotation limitations (§5.1) and then demonstrate that despite the limitations, the labels contain enough signal to benefit a downstream application (§5.2).

5.1 Limitations

As previously noted by Tratz and Hovy (2010), inter-annotator agreement on noun-compound relation annotations tends to be low. Our annotations yielded moderate agreement with averaged pairwise Kappa $\kappa = 0.47$ (Davies and Fleiss, 1982). For comparison, Tratz and Hovy (2010) reported relatively high κ values (0.57–0.67), using a weighted voting scheme and after conflating some relations. We attribute this gap to the difficulty of noun-compounds in WN-Compounds.

Interestingly, the annotators noted that many noun-compounds fit into more than one relation. This means that cases in which both annotators select valid relations may be counted as disagreements (see Table 5), while in classification, models may be over-penalized. In the future, we believe that noun-compound interpretation should be modeled as a multi-label classification task, where each noun-compound may have multiple relations.

5.2 Dataset Utility

To confirm the quality of WN-Compounds gold labels, we incorporate them directly into a model for Recognizing Textual Entailment (RTE) (Dagan et al., 2013). Given a sentence-pair, the goal is to determine whether the first sentence (*premise*) entails, contradicts, or is neutral to the second (*hypothesis*). The ability to interpret noun-compounds can potentially benefit RTE systems, when one sentence consists of a noun-compound ("*I made a phone call*"), and the other relies on its interpretation ("*I used the phone to make a call*").

Dataset	Original	With Noun Compounds
MultiNLI matched	0.716	**0.727**
MultiNLI mismatched	0.716	**0.731**

Table 6: ESIM performance on MultiNLI validation sets, with and without adding interpretations for noun-compounds in WN-Compounds.

We measure performance on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2017), with its two settings: matched and mismatched (where test and train examples are derived from different genres). We use the Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017), which uses a biLSTM to encode the words in each sentence, and an attention mechanism to align words across sentences.

We re-run the model on the original corpus, and, using the same settings, on a version of the corpus in which we replace *compositional* compounds from WN-Compounds in the hypothesis with explicit relations (e.g. *phone call*: "*call, which uses phone*"). Compounds in the premise are replaced with their heads, as the premise may provide additional unnecessary information.

Table 6 displays the performance on the validation sets.[5] Adding the interpretation of (a limited number of) noun-compounds slightly increases the performance, as exemplified in Table 4.

6 Conclusion

We release a new dataset for noun-compound interpretation which we explicitly constructed from polysemous constituents. This results in compounds whose relations are difficult to predict based on the head or the modifier alone, as is witnessed by the mediocre performance of the various baselines we explored. We believe that good performance would require integrating *joint* information about the head and modifier, and hope that this dataset can spur investigation in this direction.

We demonstrate the dataset utility by improving the performance of an RTE model using the gold noun-compound labels. Nonetheless, we argue that a better modeling for noun-compound interpretation would be a multi-label classification task. In the future, we plan to investigate this direction.

[5] A test script is available in Kaggle, but the test set is not available, so we couldn't make changes to it.

References

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 96–102.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1657–1668.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies* 6(4):1–220.

Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics* pages 1047–1051.

Corina Dima. 2016. *Proceedings of the 1st Workshop on Representation Learning for NLP*, Association for Computational Linguistics, chapter On the Compositionality and Semantic Interpretation of English Noun Compounds, pages 27–39. https://doi.org/10.18653/v1/W16-1604.

Roxana Girju. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 568–575. http://www.aclweb.org/anthology/P07-1072.

Su Nam Kim and Timothy Baldwin. 2007. Interpreting noun compounds using bootstrapping and sense collocation. In *Proceedings of Conference of the Pacific Association for Computational Linguistics*. pages 129–136.

Mark Lauer. 1995. Corpus statistics meet the compound noun. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*.

Judith N Levi. 1978. *The syntax and semantics of complex nominals*. Academic Press.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth international workshop on computational semantics (IWCS-5)*. pages 285–301.

Diarmuid O'Séaghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proc. Corpus Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01)*. pages 82–90.

Richard Socher, Brody Huval, D. Christopher Manning, and Y. Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1201–1211. http://aclweb.org/anthology/D12-1110.

Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. University of Southern California.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, pages 678–687. http://www.aclweb.org/anthology/P10-1070.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 782–788.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* .

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 1263–1271.

Ladislav Zgusta. 1967. Multi-word lexical units. *Word* pages 578–587.