Towards a Standardized Dataset for Noun Compound Interpretation

Name of author

Address - Line 1
Address - Line 2
Address - Line 3

Abstract

Noun compounds are interesting constructs in Natural Language Processing (NLP). Interpretation of noun compounds is the task of uncovering a relationship between component nouns of a noun compound. There has not been much progress in this field due to lack of a standardized set of relation inventory and associated annotated dataset which can be used to evaluate suggested solutions. Available datasets in the literature suffer from two problems. Firstly, the approaches to creating some of the relation inventories and datasets are statistically motivated, rather than being linguistically motivated. Secondly, there is little overlap among the semantic relation inventories used by them. We attempt to bridge this gap through our paper. We present a dataset that is (a) linguistically grounded by using Levi (1978)’s theory, and (b) uses frame elements of FrameNet as its semantic relation inventory. The dataset consists of 2,600 examples created by an automated extraction from FrameNet annotated corpus, followed by manual investigation. These attributes makes our dataset useful for noun compound interpretation in general-purpose setting.

1. Introduction

Noun compounds are continuous sequences of nouns that act as a single semantic construct. They raise interesting challenges in Natural Language Processing. Without proper interpretation and/or paraphrasing of noun compounds, NLP methods may fail miserably at different tasks. The meaning of a noun compound is composed of the meanings of the individual constituents and the way they are semantically related.

Noun compound interpretation is the task of detecting this underlying semantic relation (e.g., student protest: student ← AGENT ← protest). It is an important submodule for various NLP tasks such as machine translation (Baldwin and Tanaka, 2004; Balyan and Chatterjee, 2014), question answering (Ahn, 2005) etc.

Noun compound interpretation can manifest itself in two settings: out-of-context interpretation and context-dependent interpretation. In out-of-context interpretation, given the noun compound, the task is either to annotate it with a relation from a semantic relation inventory (e.g., student protest: AGENT), or to produce a paraphrase (e.g., student protest: “protest carried out by student”).

Any automated approaches for noun compound interpretation need a semantic relation inventory of noun-noun relations, and an annotated dataset on which models can be trained. However, there is little agreement among researchers regarding the set of relations that can hold between the constituents of a noun compound. None of the proposed semantic relation inventories has been accepted by the community as complete and appropriate for general-purpose text. Some are coarse-grained, while others are fine grained. There is little overlap among them. Also, some of these inventories and their accompanying dataset are created from another application’s perspective, and not for the sake of creating a noun compound dataset. Thus they cannot be used for learning noun compound interpretation in a general-purpose setting.

A dataset that can be used in general-purpose setting needs to be linguistically grounded. One such work is that of Levi’s, who claims that noun compounds are created either through predicate deletion or through predicate nominalization. For example, student protest and student demonstration are examples of predicate nominalization with heads as verbal form and nominalized form, respectively. Orange juice is an example of predicate deletion as connecting predicate (like, made of) has been simply dropped while creating the compound. We ground our dataset on this theory.

FrameNet is a lexical resource based on the theory of frame semantics. Among other things, it captures predicate-argument interactions. Such information can be used for compounding. For instance, border camp with RESIDENCE:LOCATION and rescue attempt with ATTEMPT:GOAL are examples of predicate deletion and predicate nominalization, respectively, with corresponding frame and frame-element as labels. Intuitively, the frame elements are descriptive enough of the relation among the predicate and argument.

Thus, we propose a dataset that is linguistically grounded (by Levi’s theory) and uses frame elements as semantic relation inventory.

The rest of the paper is organized as follows: Section 2. covers some background needed for further discussion. Section 3. discusses other semantic relation inventories, highlighting their shortcomings. Section 4. discusses the creation and statistics of our dataset. Section 5. presents several observations during this activity, followed by the conclusion and future work.

2. Background

A noun compound can be of any length. A typical way for interpretation of longer (having more than two components) is parsing it to get a binary tree based on head-modifier pairs, and interpret each internal node of the tree with two children of the node as components. For example, parse trees (in bracketed form) for “plastic water bottle” and “water bottle cap” are as follows:

(plastic [water bottle])    ([water bottle] cap)

After parsing, the problem reduces to the interpretation of two components of each internal node. In literature,
most work focuses on the interpretation of noun-noun compounds, i.e., noun compounds composed of two nouns. In the rest of this paper, by noun compound, we mean noun-noun compounds.

For representation of the semantic relation between the components of a noun compound, there are two major ways:

**Paraphrasing:** paraphrase a noun compound to show how the components are related (e.g., *orange juice: “juice made of orange”, “a drink consisting of the juice from oranges*, etc.) (Butnariu et al., 2009; Hendrickx et al., 2013). There can be multiple paraphrase of a noun compound.

**Labeling:** Assign a relation from predefined set of abstract relations (e.g., *orange juice: SUBSTANCE/MATERIAL/INGREDIENT*). (Levi, 1978; Warren, 1978; Tratz and Hovy, 2010)

Labeling is the most widely used representation in literature for noun compound. There are some attempts to paraphrase noun compounds. In between the two representation, researchers have also used scoring of template-based paraphrase for assigning abstract labels (Nakov, 2008; Nakov and Hearst, 2013).

### 2.1. FrameNet

FrameNet1 (Baker et al., 1998) is a lexical database that shows usage of words in actual text based on annotated examples. It is based on a theory of meaning called Frame Semantics (Fillmore, 1976). The theory claims that meanings of most words can be inferred from a semantic frame: a conceptual structure that denotes the type of event, relation, or entity and the involved participants. For example, the concept of walking typically involves a person walking (*SELF_MOVER*), the path on which walking occurs, the direction in which the walking occurs, and so on. In FrameNet, this information is represented by a frame called SELF_MOTION. SELF_MOVER, PATH, DIRECTION, etc. are called frame elements (FEs). Such frames are invoked in running text via words known as lexical units (LUs). Continuing the above example, some of the lexical units for the frame SELF_MOTION are *advance, crawl, dash, drive, march, run, walk, etc.* Most LUs are verbs. But, it can be a noun or an adjective, too.

An example sentence in FrameNet annotated data contains a target word along with linked LU, arguments of the target, and an FE for each of the targets. Following is an example of SELF_MOTION frame with *jog v* LU:

```
[SELF_MOVER] JOGGED [TARGET] [Path down the stairs] [Time while he began to argue with the builders].
```

In this work, we generate noun compounds from the FrameNet annotated sentences, and assign FEs as relations between them.

---

1https://framenet.icsi.berkeley.edu/

### 3. Related Work

For interpretation task, we need a representation of semantic relations (SRs) which is based on linguistic intuition and it can help the other NLP tasks. Scholars have proposed various inventories of abstract relations for the task. But, we found that each inventory lacks some something. We are trying to solve those shortcomings.

Levi (1978)’s study on noun compound generation is the most influential one. The study categorizes noun compounds based on the compounding process as (1) predicate deletion, where a predicate between the components is simply dropped to create a compound, and (2) predicate nominalization, where the head is nominalized form of a verb and modifier is an argument of the verb. They proposed a set of abstract predicates for the former category, but no labels for the latter category.


Barker and Szpakowicz (1998)’s proposed set of relations based on Levi (1978)’s theory and (Warren, 1978) inventory. They claim that SRs in their inventory are most widely used and can improve with time. Kim and Baldwin (2005) prepare a dataset for this inventory, but the dataset is highly imbalanced. For instance, out of 20 relations, TOPIC relation has 42% examples and PURPOSE relation has 23% examples in contrast to less than 10 examples of 3 SRs. Vanderwende (1994) used 13 relations based on the syntactical category and types of questions. Girju et al. (2005) provided another inventory of semantic relation based on Moldovan et al. (2004)’s semantic relation in noun phrases. But, most examples in the dataset uses prepositions as SRs. In addition, 14 of total 35 SRs has not any example in their dataset and 7 more SRs has less than 1% examples. For an inventory of SRs, if an SR has no example, then it raises a question on the base of the inventory.

Ó Séaghdha and Copestake (2009) proposed an inventory of SRs based on RDP (recoverable deleted predicates) of Levi (1978). In addition 5 SRs for compositional NCs – the meaning of the compound is composed of the meaning of the components – they used 5 more SRs for other categories like lexicalized compounds, wrongly tagged compounds. The 5 compositional SRs has been further categories in total 11 categories. They have also prepared a dataset with 1443 examples for the 5 course and 11 fine relations.

In addition to the above-mentioned dataset for the general domain, Rosario et al. (2002) proposed an inventory of SRs for medical domain.

Tratz and Hovy (2010) claims that they have created a new inventory of semantic relations by comparing and consolidating the existing inventories. But, in contract, their inventory creation process is an iterative process to improve inter-annotator agreement. Ponkiya et al. (2016) reports many problems with this inventory.

Even after these inventory of semantic relation, when it comes to showing usage of the semantic relation, people used a custom set of semantic relation. For instance Balyan and Chatterjee (2014) shows how the interpretation of NCs can help automatic machine translation (MT). But, they
4. Proposed Dataset

4.1. Dataset Fields

Our dataset contains the following fields:

1. $w_1$: The first word of the noun compound.
2. $w_2$: The second word of the noun compound.
3. Frame: ID and name of the frame from which the example was created.
4. FE: ID and name of the frame element from which the example was created.
5. KB05: Label in Kim and Baldwin (2005)’s dataset (hereafter, KB05) (NA if not found).
6. OS08: Label in Ó Séaghdha and Copestake (2009)’s dataset (hereafter, OS08) (NA if not found).
7. TH10: Label in Tratz and Hovy (2010)’s dataset (hereafter, TH10) (NA if not found).
8. Type: Type of noun compound according to Levi’s theory.

4.2. Dataset Creation

The dataset was created in two phases: an automated phase where candidate noun compounds are extracted from FrameNet annotated corpus, followed by a manual phase where we annotate each candidate on the basis of Levi’s theory.

Automated Phase

In this phase, we take the examples from FrameNet for each frame $F$. Each example is processed as follows:

1. Find the target word $T$
2. Let $TSet$ be the set of all possible verbal forms and nominalized forms of $T$
3. For each chunk $C$ annotated with frame element $E$
   a. Let $H$ be the head word of the dependency parse of the chunk $C$
   b. If $\langle H, W \rangle$ occur in either KB05, OS08, or TH10 (where $W \in TSet$)
      - Output $\langle H, W \rangle$ as candidate NC, along with $F$, $E$, and labels in KB05, OS08, or TH10

Consider the example sentence from the frame PROTEST:

The civil war that began in February with [Degree mass] PROTESTS against Kadafi’s rule has paralyzed the industry

Here, the target word is protest. The chunk mass is annotated in FrameNet with frame element DEGREE, and has mass as the head word. $\langle$ mass protest $\rangle$ is present only in TH10, thus the process outputs $\{\text{mass, protest, PROTEST, Degree, NA, NA, TH10Label}\}$.

Similarly, the chunk against Kadafi’s rule is annotated in FrameNet with frame element ISSUE, and has rule as the head word. However $\langle$ rule protest $\rangle$ is not present in either KB05, OS08, or TH10. Thus, we do not consider it as a candidate noun compound.

Note that a candidate noun compound can be generated from more than one frames, thereby having multiple $\{\text{frame, frame_element,}\}$ labels. In that case we repeat the candidate noun compound, once for each label.

Manual Phase

In this phase, we check the correctness of labels (specially, frame and FE) manually. For example, consider the following annotated sentence:

If the scientists are right, then a major clue about how $\lfloor$Entity cancer $\rfloor$ DEVELOPS $\lfloor$Target in children $\rfloor$ has been found.

The previous automated phase generates children development as a candidate noun compound with the label $\{\text{coming_to_be, place}\}$. But, in general, usage of children development means “development of children”, and not “development in children”. We simply drop such noun compounds.

Then, each of these candidate NCs were manually annotated as follows:

- PD: The candidate is an NC created through predicate deletion. Examples: orange juice (juice made of orange), cricket bat (bat made for cricket), etc.
- PN: The candidate is an NC created through predicate nominalization. Examples: student protest, gender segregation, etc.

4.3. Statistics

The final dataset contains 2,600 noun compounds, formed through the combination of 818 modifiers and 806 heads. The set of unique modifiers and heads contains 1401 words, with 223 words appearing both as modifiers and as heads. Total number of unique frames is 409. Total number of unique frame + frame element combinations is 893.

5. Discussion

Here we discuss some of the advantages that one obtains from creating the dataset in the manner we have proposed.

- In FrameNet, hierarchy of FE (defined in addition to FE relations) can help generalization of the relations. For instance, our dataset has student protest: PROTESTER as an example. Using FE relations (e.g., PROTESTER of PROTEST –ISA→ AGENT of INTENTIONALLY_ACT), we can infer student protest: AGENT.
• As each noun compound has been annotated with frame and FE, details of corresponding frame and FE helps in paraphrasing the NC.

• On an average, number of frames that can be invoked by a head of an NC is not too high (3.27 for our dataset). This narrow downes degree of ambiguity. So, even thought we have thousands of FEs in FrameNet, search space for an NC is less than 100 most of the times.

• Example of FEs from FrameNet annotated data can help in disambiguation of FEs for a given NC. For example, while labeling fee-hike protest with FEs of PROTEST frame, there may be confusion between ISSUE and PURPOSE. In such case, examples of those FEs from FrameNet annotated data can help in disambiguation.

• The main intention of this exercise is to create a dataset that can be used as a gold standard for further research in noun compound interpretation. Therefore, we tried to reduce false positives as much as possible, i.e. ensure that a noun compound included in the dataset is labeled with correct frame and frame element. In the process, we decided to remove certain examples where the assigned label seemed to be a corner case. For example, consider the following example sentence:

\[
_{Entity} \text{The theater} \quad \text{PRESENTS}_\text{Target} \quad _{Phenomenon} \text{sky shows and IMAX films}.
\]

Our automated phase generates \textit{theatre presentation} as a candidate noun compound, with the label \{\textit{CAUSE_TO\_PRECIVE}, \textit{ENTITY}\}, implying that the presentation is by the theatre. However, this seems like a corner case, as a presentation is more likely to be at the theatre. Thus, we do not include such examples in our dataset.

• This exercise also lead to fixing errors present in other datasets. For instance, Ő Séaghdha and Copestake (2009) states that their dataset contains noun compounds created only by predicate deletion. However, we observed that out of the 145 noun compounds that matched with their dataset, 25 were of the type predicate nominalization. For example, they label \textit{questionnaire reply} as predicate deletion, but it is an example of predicate nominalization.

6. Potential Applications

Enriching Dependency Parsing

A direct application of this dataset that can benefit the community is the addition of this dataset and its information to a dependency parser. We believe that a dependency parser can be modified to include noun compound detection, and then use the appropriate frame element to improve erroneously labeled arcs. For instance, stanford parser currently labels student protest as student \textit{\text{comp}und} protest. The knowledge that student is a Protesor in student protest can help the parser to correctly parse it as student \textit{\text{nsubj}} protest

7. Conclusion and Future Work

In this paper, we proposed a dataset for noun compound interpretation. The dataset is linguistically grounded using Levi’s theory. It uses frames and frame elements of FrameNet as the semantic relation inventory. We took this steps with the goal of creating a standardized dataset, the lack of which is severely affecting research in noun compound interpretation. Our dataset contains 2,600 examples. Each noun compound is annotated according to the type of noun compound (predicate deletion vs. predicate nominalization), the frame and frame element through which the noun compound was created in the first place, and its label in three other datasets. We also discussed how this dataset can be useful to improve dependency parsers.

In the future, we will extend this dataset to include more noun compounds. Currently, we severely restricted our dataset size by considering only those noun compounds that occur in other datasets as valid noun compounds. For example, out of the 259 candidates generated automatically, a manual investigation suggested 58 valid noun compounds. However, our restrictions led to inclusion of only 6 noun compounds. Thus, there is a scope for including many more noun compounds. We will also investigate whether the set of frame elements applicable to noun compounds is a proper subset of the entire set of frame elements.

8. Bibliographical References


task 4: Free paraphrases of noun compounds. *Atlanta, Georgia, USA*, page 138.


