A VSM-based Statistical Model for the Semantic Relation Interpretation of Noun-Modifier Pairs

Nitesh Surtani
IIIT Hyderabad
nitesh.surtani0606@gmail.com

Soma Paul
IIIT Hyderabad
soma@iiit.ac.in

Abstract

The paper addresses the task of automatic interpretation of semantic relation in noun compounds. The problem has been attempted with both Ontology-based and Statistical approaches, but both approaches having their own limitations. We present a novel VSM-based statistical model which represents each relation with a weighted vector of prepositional and verbal paraphrases. The model ranks the paraphrases on their relevance and assigns higher weights to more relevant paraphrases. The performance of the model is compared with the Ontology model and the results are quite encouraging. We finally propose a Hybrid of the two models which compares on par with the best performing systems on Nastase and Szpakowicz (2003) dataset.

1 Introduction

There has been an increased interest in discovering the semantics of Noun Compounds (NCs[1]). There are two reasons that make this task quite essential and interesting in text understanding: **(i) their implicit nature**, for instance the NC *'monday meeting'* is the *meeting scheduled on monday* (Temporal), *'teacher meeting'* is the *meeting organized for teachers* (Participant) and *'NLP meeting'* is the *meeting to discuss NLP topics* (Quality-Topic); and **(ii) their frequent and compounding behavior**. NCs are very frequent in english and comprise of 3.9% and 2.6% of all tokens in the Reuters corpus and the British National Corpus (BNC) respectively (Baldwin and Tanaka, 2004). New NCs are very frequently constructed *eg. website design, internet usage, orange juice* etc., and sometimes combine with other words to form longer compounds, e.g., *orange juice company, orange juice company homepage* etc. The frequency spectrum of NCs follows a Zipfian distribution (Séaghdha, 2008), where many NC tokens belong to a *long tail* of low-frequency types. Over half of the *two-type* compounds in BNC occur just once (Kim and Baldwin, 2006).

The research focusing on the semantic interpretation of NCs has followed two directions: (i) Identifying the underlying semantic relation (Girju et al., 2005; Tratz and Hovy, 2010); and (ii) Paraphrasing the NC (Nakov, 2008; Butnariu and Veale, 2008; Butnariu et al., 2010). Consider the text:

> "A **large student protest** was *carried out during* ***monday evening*** by various **engineering colleges** to raise funds for research. This **London protest** saw tremendous participation by students from 14 colleges, seeing to which R&D dept. agreed to increase the **college funds** to 10,000,000 GBP. "

The sequences marked in bold in the above example are Noun compounds (NCs). In the above text, some NCs are interpretable via paraphrasing: **protest** was *carried out during* **evening**, where *'during'* defines the temporality of the protest. On the other hand, some NCs are not explicit: **student protest** meaning that the *'protest was done by the students'* (Agent), **London protest** meaning *'protest was held in London'* (Spatial), **monday evening** meaning *'evening of monday'* (Part-Of), **engineering colleges** meaning the *'colleges that specialize in engineering course'* (Purpose), **college funds** are the *'funds allocated for the college'* (Beneficiary). The goal of this paper is to discover the underlying semantic relation of the NCs *via paraphrasing*. The knowledge of semantic relation in the above NCs can help in answering questions like: *Where was the protest held? Who led the protest? etc.* The tasks has applications in many subfields of NLP, including Question Answering (Girju et al., 2006), Knowledge Base acquisition (Hearst, 1998) and others.

The task of semantic relation classification of NCs has been attempted in two directions: (i) using

[1] A noun compound (NC) is a sequence of nouns which act as a single noun (Downing, 1977), *eg. sunday morning*

a *knowledge-intensive ontology* and (ii) *extracting paraphrases from a large corpus*. We discuss two existing WordNet-based ontology models: SemScat 1 (by Moldovan et al. (2004)) and SemScat 2 (by Beamer et al. (2008)), which uses the WordNet's noun Hypernym (IS-A) hierarchy to find semantic similarity between two Noun-Noun pairs. The main focus (and contribution) of this paper is towards developing a Statistical model which uses Prepositional (*eg. 'benefit for consumer'*), Verbal (*eg. 'benefit involving consumer'*) and Verb+Prep (*eg. 'benefit received by consumer'*) paraphrases of the NC (*eg. 'consumer benefit'*) for identifying its relation.

The paper is organized as follows: **Section 2 (Related Works)** describes previous works on Ontology and Statistical models; **Section 3 (Data Analysis and Specification)** describes the dataset used for experiments, **Section 4 (Ontology-Based Model)** and **Section 5 (Corpus-Based Model)** discusses, experiments and provide insights on these two models. **Section 6 (Integrated Model)** develops a hybrid of the two models and **Section 7** concludes the paper.

2 Related Works

Ontology-based Approach: Nastase et al. (2006) explores both WordNet and Roget's Theasaurus for forming the classification features and find WordNet ontology to be more suitable for the task. Girju et al. (2003), Moldovan et al. (2004) and Beamer et al. (2008) propose *Iterative semantic specialization (ISS), SemScat 1 and SemScat 2 models* respectively, which utilize WordNet's Hypernym hierarchy and specialize the synsets from general to specific level. *ISS* employs *Decision Tree (C4.5)* for modelling a single *Part-Whole* relation. *SemScat 1* and *SemScat 2* are designed as multi-class classifiers for modelling a set of 35 relations (Moldovan et al., 2004) and 7 relations (Girju et al., 2007) respectively.

Statistical Approach: Nakov and Hearst (2006) suggests that the semantics of noun compounds is best expressible using multiple paraphrases involving verbs and prepositions. For example, *bronze statue* is a statue that is *made of, is composed of, consists of, contains, is of, is, is handcrafted from, is dipped in, looks like* bronze. Nastase et al. (2006) makes an assumption that senses of NCs can be derived through collocated words learned from large corpus and use a sparse vector of collocated words as features (approx 10,000 features). Their system performs with low accuracy and is outperformed by their WordNet model of sparse Hypernym synset feature vector. Nulty (2007) extracts 28 prepositional paraphrases by forming simple *'N2 prep N1' or 'N2 prep the Y'* templates and querying the web. He shows that the less frequent prepositions achieve higher accuracy than the more frequent ones in classifying the relation. This observation aligns with ours and we employ a TF/IDF (modified) scheme to assign higher weights to such paraphrases. Turney (2006b) introduces a *Latent Relational Analysis (or LRA)* model. The model extracts all possible synonyms for the modifier and the head using a thesaurus and uses a list of 64 joining terms, *J* such as *'of', 'for'* and *'to'* to form 128 phrases (i.e. *M J H* and *H J M*). From the set of extracted paraphrases, top few thousands selected paraphrases are used to build an incidence matrix, whose dimensionality is reduced using singular value decomposition (SVD). Nastase et al. (2006), Turney and Littman (2005), Turney (2006a), Turney (2006b) and Nulty (2007) compare their systems on Nastase dataset, where Turney (2006b) outperforms others achieving a accuracy of 58% and 54.6% macro-averaged f-score[2].

3 Data Specification

We work with two datasets: (i) Nastase and Szpakowicz (2003) dataset of noun-modifier pairs (referred as Nastase dataset in the paper); and (ii) Butnariu et al. (2013) SemEval-13 Task 4 gold-paraphrased dataset (referred as SemEval dataset). Nastase dataset uses a two-level taxonomy of 5 coarse-grained and 30 fined-grained relations and comprises of 600 Noun-Modifier pairs consisting of a head noun and a modifier which can either be noun, adjective or adverb. The data is annotated with *semantic relation* of the NC and *POS tag* and *WordNet senses* of the modifier & head. This data has some issues: there are 4 cases of repetition and 3 compounds contain multi-word modifier (eg.- *'test tube' baby*), which have been pruned out. In the remaining 593 NCs, there are 326 instances of noun (55%), 260 instances of adjectives (44%) and 7 instances of adverbs (1%) modifier. The SemEval dataset consists of 355 Noun-Noun compounds which are manually paraphrased by *approx.* 30 annotators, with a total of 12,471 paraphrases. Each paraphrase is assigned a frequency, which is number of annotators who have marked that paraphrase for the given NC. We have annotated the NCs with *semantic relations* and modifier & head *WordNet senses* following the guidelines from Nastase and Szpakowicz (2003). The experiments on

[2] Macroaveraged f-score is the overall mean of f-scores of individual classes.

RELATION	Nastase (2003)	SemEval (2013)
Causal	85 (14.33%)	95(26.9)
Participant	259 (43.67%)	108 (30.5)
Quality	144 (24.28%)	107 (30.2)
Spatial	54 (9.1%)	32 (9.1)
Temporal	51 (8.6%)	13 (3.3)
Total	593 (100%)	355 (100%)

Table 1: Distribution of Relations in Datasets

the SemEval dataset of gold paraphrases helps us in harnessing the full potential of the Statistical model which is not possible with Nastase dataset, as the quality of extracted paraphrases is nowhere close to manually annotated paraphrases. On the Nastase dataset, we compare the performance of our Hybrid model with other models evaluated on this dataset.

The ontology and corpus models are designed to handle only Noun-Noun compounds (Beamer et al., 2008; Turney, 2006b). We extend the ontology model to work with adjective and adverb modifiers but such adaptation is not possible for the corpus model. The ontology model uses the Noun Hypernymy hierarchy, which is extended to adjectives and adverbs by linking them to their corresponding noun synsets, through following WordNet relations: *derivationally_related_form, pertainym, attributes_to* and *similar_to* (eg. *electric#a#1* → *electricity#n#1*). The corpus model always yields similar paraphrases with adjective or adverb modifiers, making such paraphrases irrelevant for classification. Thus, the corpus model works with only 326 Noun-Noun compounds in Nastase dataset.

4 Ontology-Based Approach

We experiment with two WordNet-based models: *SemScat 1* by Moldovan et al. (2004) and *SemScat 2* by Beamer et al. (2008). The model works on the principle that two NCs having similar concepts in the Hypernym hierarchy encode same relation.

4.1 Model Formulation

Let L be the set of all the hypernym entity types (or synsets). Let the training set of n instances $T = ((x_1 r_1)....(x_n r_n))$, where $x_1....x_n$ represent the NCs annotated with semantic relations $r_1....r_n$ respectively, where $r_i \in$ relation set R. The input x_k is represented in terms of modifier and head features $< f_i^m, f_j^h >$, where f_i^m, $f_j^h \in L$ represent synsets at level i and j in the hypernym hierarchy, combinedly represented as f_{ij}. Therefore, the goal is to model the prediction function $F : (L \times L) \rightarrow R$.

The SemScat models strives to learn generalized sets of Hypernym synsets, known as Boundary, G. For instance - $G_1 = \{entity\}$ and $G_2 = \{physical - entity, abstract - entity, thing\}$ are two boundaries where G_2 is hyponym of G_1. The algorithm starts by creating the most general boundary $G_1 = \{entity\}$ and all the training examples are mapped to this boundary by forming $< Modifier-Head >$ feature $f_{11} = \{entity-entity\}$. Then, the model computes the probability of each relation r for every feature formed in this new boundary. Next, the model identifies the most ambiguous feature (the one having the highest entropy) using the weighted entropy measure (Beamer et al., 2008) and specializes its modifier & head synsets by their hyponyms. The algorithm again computes the statistics on this new boundary and the process is repeated.

Key Differences- SemScat 1 and SemScat 2: The main difference between the two models is the manner in which they store their boundary. SemScat1 strives to discover a single optimal boundary G^*, at which all the features map uniquely to a relation. But practically, the boundary G^* is overspecialized and therefore, the model finds a boundary G_k which generalizes well over the test set by using a *development set*. The SemScat 1 terminates the further specialization of the boundary when the performance of the model drops on the development set (i.e. the model starts to over-specialize). It also uses a Threshold parameter (T) to restricts the over-specialization of the features f_{ij}, by treating it as disambiguated if the most probable relation corresponding to feature f_{ij} has the probability greater than T. On the other hand, the SemScat 2 model keeps track of all the boundaries ranging from the most general to most specific boundary (G^*), $G = \{G_1, ..., G^*\}$ and terminates the training after discovering the boundary G^*. Given an unseen instance (with feature f_{ij}), SemScat 1 searches for this feature in the stored boundary G_k, and if the feature is matched at this boundary, it assigns the most probable relation corresponding to the feature f_{ij}, otherwise the instance is considered as missed and no relation is assigned; while SemScat 2 starts the search for the feature f_{ij} from the most specific boundary G^* and moves towards more general boundaries, and assigns the relation corresponding to the most specific feature matched.

4.2 Experiments

We perform two progressive experiments with ontology model. The experimental setup, results and insights gained are presented for each experiment separately. This section discusses the results of experiments on ***ONLY*** Nastase dataset. The results on SemEval dataset are presented in Section 6.

Experiment I: Comparison of SemScat 1 and SemScat 2: In this experiment, we compare the performance of the two models on the Nastase dataset. We perform k-fold cross-validation to evaluate the performance of the model over the complete dataset. The value of k is varied as: k = 5, 7, 10, 15, 30, 50, $N-1$ (Leave-one-out [3]). The data is divided into training, development and testing set for SemScat 1, while into training and testing set for SemScat2. The development set used in SemScat 1 comprises of 20% data from the training set. The Threshold factor (T) is varied from 0.6 to 0.9 in steps of 0.05.

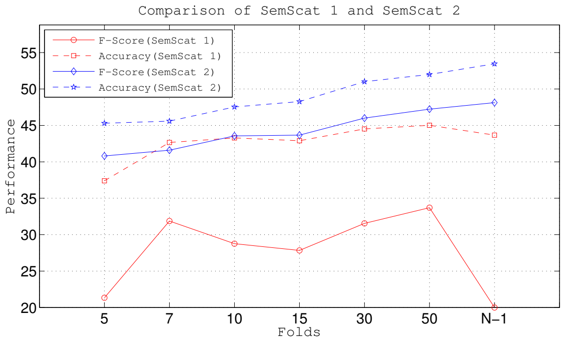

Figure 1: Performance comparison of SemScat 1 and SemScat 2 on Nastase dataset at varying k folds

SemScat 2 outperforms SemScat 1 on each fold achieving the optimal performance at $k = N - 1$ with the 53.46% accuracy (baseline 43.67%) and 48.13% f-score. SemScat 1 performs just above the baseline with accuracy and f-score of 45.02% and 33.70% respectively at $k = 50$ and $T = 0.7$, classifying most of the instances with the majority relation *Participant*. We find that the boundary G^* is quite specific (ranging from *level 6-8* on Nastase dataset) while the boundary generally selected by SemScat 1 ranges from *level 3-4* in the experiments. This reveals that SemScat 1 fails in achieving the goal of finding its optimal boundary that is the closest approximation of the boundary G^* and thus, misses out knowledge that would be useful for classification. The huge performance gap between the model using single boundary and the model storing multiple boundaries motivates us to investigate the authenticity of each boundary in attesting the relation.

Experiment II: Performance of Different Boundary Levels in SemScat 2: This experiment evaluates and compares the performance of multiple boundaries stored by the SemScat 2 model. The model is trained on optimal parameters $k = N - 1$ and the accuracy of each level is computed.

Level	Total	Correct	Accuracy
2	17	5	29.41
3	162	65	40.12
4	198	107	54.04
5	152	104	68.42
6	36	17	47.22
7	23	18	78.26

Table 2: Performance of SemScat 2 at different levels

The results presented in Table 2 show that the confidence of the model in assigning the relation improves significantly with each level (except for *level 6*). The model performs with accuracy of only 29% at boundary *level 2* which shoots up to 78% at *level 7*. Most of the test instances are mapped at *level 4 and 5*, achieving accuracy of 54% and 68% respectively. This indicates that the NCs are classified accurately when matched with more specific knowledge. We capitalize of this useful insight in the Hybrid model. Further, we observe that the ontology model faces difficulty in disambiguating between certain set of relations, eg. *Student Protest* (Agent) and *Student Discount* (Beneficiary) are represented with very similar concepts in the Hypernym hierarchy and therefore, the model fails to classify the NCs correctly. On the other hand, the corpus model easily classifies these NCs, since *'protest* ***(led_by, organized_by)*** *students*' clearly points to **Agent** relation whereas *'discount* ***(for, given_to)*** *students'* suggest that modifier is the **Beneficiary** of the action. This complementing behavior of two models establishes the ground for integrating them.

5 Statistical Approach

The statistical model captures the meaning of the NC using *Prepositional, Verbal* and *Verb+Prepositional* paraphrases and uses them to identify the underlying semantic relation. For instance, *student protest* **(Participant)** is paraphrased as *'protest* ***(by, of, led_by, involving, started_by)*** *students*', *London protest* **(Spatial)** as *'protest* ***(in, at, of, held_at)*** *London*' and *evening protest* **(Temporal)** as *'protest* ***(during, of, held_during)*** *evening'*. In the above examples, the preposition ***'by'*** clearly points to **Participant** relation, ***'in'*** and ***'at'*** to **Spatial** relation and ***'during'*** to Temporal relation. Similarly, verbal paraphrases ***'involving'*** and ***'started_by'*** indicate **Participant** while paraphrases ***'held_at'*** and ***'held_during'*** indicate **Spatial** and **Temporal** relations respectively. Prepositions are polysemous in nature and the same preposition can indicate different semantic relation, as also observed by (Srikumar and Roth, 2013), *for eg.* the preposition ***'from'*** occurs in: *'death*

[3] In Leave-one-out, one instance is tested at a time while rest $N - 1$ instances are used for training

from** cancer'* (Causal-Cause), *'excerpt **from** the book'* (Participant-Source), *'protest **from** evening'* (Temporal) etc. But the degree of polysemity varies with prepositions, *for eg.* the preposition ***'of' in the above 3 NCs maps to 3 different relations but the prepositions ***'by', 'at'*** and ***'during'*** occur specifically with *Participant, Spatial* and *Temporal* relations respectively. Prepositions that map to a single or fewer relations are more relevant for the task than the ones which frequently occur with different relations and thus, are weighted higher. Furthermore, we observe that the verb+prep paraphrases are quite significant, as such verbs are mostly accompanied with relevant prepositions, *for eg.* the paraphrase *'Protest held during evening'* is plausible but *'Protest held of evening'* is not. Therefore, the preposition & verb in such paraphrases are given more relevance using a Strength parameter.

The statistical model represents each NC as a pair of vector of prepositional and verbal paraphrases. With the relation of the NC known (*i.e. supervised learning*), we transform the NC vectors into Relation Vectors, which represent the complete semantic class with a single pair of prepositional and verbal vector. The *Vector Space Model (VSM)* with *Nearest Neighbour* classifier employed by the model computes the cosine similarity of the test vector with each Relation vector and assigns it the relation with the highest similarity. The next sections describe the two most important modules of this model: Paraphrase Extraction and Vector Formation module.

5.1 Paraphrase Extraction Module

The goal of this system is to take an NC as input and provide the set of *prepositional, verbal and verb+prep* paraphrases for it. It consists of three submodules: former dealing with extraction while latter two perform cleaning of paraphrases.

Module 1: Paraphrase Extraction: We have relied mainly on the Google N-gram Corpus for extracting the paraphrases. Google has publicly released their web data as n-grams, also known as Web-1T corpus (Brants and Franz, 2006). The corpus contains 2-, 3-, 4- and 5-grams sequences and returns n-gram matches that occur more than 40 times. The templates for extraction with few (*correct and incorrect*) selected paraphrases for NC *Copper Coin* are presented in Table 3 and 4 respectively. Among incorrect paraphrases, the first two are syntactically illegitimate while the last two are syntactically sound but semantically illegitimate. *'coins are copper'* is part of *'one cent **coins are copper** or not'* while *'coins in copper'* is part of **'coins in copper** *bowl'*.

coin [s|p] <*>copper [s|p]	coin of copper
coin [s|p] <*><*>copper [s|p]	coins made from copper
coin [s|p] <*><*><*>copper [s|p]	coin is made of copper

Table 3: Extraction Templates with Examples

Correct Paraphrase	**Incorrect Paraphrase**
coin of copper 63	coin : copper 91
coins made from copper 108	coin jewelry copper 51
coins made of copper 49	coins are copper 91
coin is made of copper 146	coins in copper 55

Table 4: Paraphrases Extracted from Google N-Gram

Module 2: Syntactic Cleaning: To handle the syntactically ill-formed paraphrases, we prepare a set of plausible syntactic templates. The paraphrases for 60 NCs (*with total of 5716 paraphrases*) are manually marked as incorrect or correct (0 or 1 respectively) by two annotators, with high agreement of annotation, since the complexity of the task is **EASY**. The correct paraphrases are POS tagged using the *CMU ark-tweet POS-tagger* (more efficient in tagging 3- & 4-grams than the Stanford POS-tagger) and POS templates are extracted. The data is divided equally into training and testing sets of 30 NCs each. Table 5 shows that the syntactic templates, although learnt from considerably small training data, are exhaustive and achieve good coverage of 91.4% on the test set, but low precision of 56.7% as many semantically illegitimate paraphrases are matched by these templates.

	Recall	**Precision**	**F-Score**
[Without Constraints]	91.4	56.68	69.97
[With Constraints]	91.4	72.9	81.11

Table 5: Comparison of Syntactic Templates *before* and *after* applying Semantic Constraints

Module 3: Semantic Cleaning: The syntactic templates are unable to filter out semantically illegitimate paraphrases. Such paraphrases are cleaned by looking at their context, extracted from extended paraphrases: *coins in copper* $< * >< * >$.

> **Constraint:** *If the modifier of a given NC is part of a NP chunk having another noun as head, then it is not a legitimate paraphrase.*

eg: (NP (NNS **coins***)) (PP (IN* **in***) (NP (NN* **copper***) (NN bowl)))*

which means *'coins kept in bowl made of copper'*. Applying this constraint shows significant improvement in precision, with f-score reaching ~81%. There are still few paraphrases which are not filtered out by this module. For eg. ***'party after class** gets over'* for NC *'class party'*. The verb+prep paraphrases (eg. *'make_of: 242'*) are splitted into verb (eg. *'make: 242'*) and preposition (eg. *'of: 242'*) and contribute to respective vectors with Strength parameter S_p and S_v, as discussed in Experiment II.

5.2 Model Formulation

Let the training set of n instances $T = ((x_1 r_1)...(x_n r_n))$, where $x_1...x_n$ are the NCs and $r_1...r_n \in R$ are their corresponding relations. Each instance x_i is represented by two vectors: a prepositional and a verbal vector. The prepositional vector consists of $m = 30$ prepositions, $P =< p_1, ..., p_m >$ and the verbal vector consisting of top-k frequent verbs represented as, $V =< v_1, ..., v_k >$. The input x_i is mapped to the prepositional vector, $x_i^p =< p_1^i, ..., p_m^i >$ and verb vector $x_i^v =< v_1^i, ..., v_k^i >$, where p_j^i represents the weight of the feature j in prepositional vector. The NC vectors are transformed into Relation vectors, where each relation $r_i \in R$ is represented by a single pair of prepositional and verbal vectors, $R_i^p =< p_1^i, ..., p_m^i >$ and $R_i^v =< v_1^i, ..., v_k^i >$ respectively. The VSM computes the cosine similarity between the two vectors, where higher value of cosine similarity means that two vectors are more similar to each other.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} \vec{r_1}_i . \vec{r_2}_i}{\sqrt{\sum_{i=1}^{n} (\vec{r_1}_i)^2 . \sum_{j=1}^{n} (\vec{r_2}_j)^2}} = \frac{\vec{r_1} . \vec{r_2}}{\parallel \vec{r_1} . \vec{r_2} \parallel} \quad (1)$$

where $\vec{r_1}$ is the training vector and $\vec{r_2}$ is the test vector. We modify the VSM algorithm in case of computing similarity with Relation vectors, in order to allow them to handle the distribution of relations. Therefore, the Relation vectors are not converted to unit vector and thus, the VSM computes $\vec{r_1} \cos(\theta)$.

5.3 Vector Formation

In this module, we discuss the transformation of NC vector to Relation vector and describe the (modified) TF/IDF scheme used for weighting the vectors.

Forming Relation vector: A relation vector is a single pair of prepositional and verbal vector that captures the behavior of the entire relation and also incorporates the distribution of each relation in training data. The Relation vector (of relation r) is formed by the vector addition of all NC vectors in the training set that belong to relation r:

$$\langle\, R^r \,\rangle = \sum_{x \in T} \langle x^r \rangle \quad (2)$$

where T is the training set and x^r are the NCs in T with relation r.

Weighting Scheme: By weighting the vectors, we want to assign higher weights to more relevant paraphrase features. For our model, the paraphrases that map to a single or fewer relations are more relevant than the ones mapping to many relations. We use the *TF/IDF weighting function* but modify it with necessary variations. First, our *TF* function takes usual logarithmically scaled frequency but is normalized to ensure the equality in document length, since the frequency of paraphrases extracted for different NCs vary significantly. For calculating the *IDF*, we take into account the relative weights of each paraphrases (or features) rather than their occurrence (0 or 1) with the NC. This modification is essential, since the Vocabulary size $|V|$ = *Number of prepositions (or verbs)* in our model is relatively very small, and doing this ensures that the noisy extracted paraphrases (with low frequencies) do not harm the model.

$$TF_i^x = \frac{log(f_i)^x}{\sum_i log(f_i^x)};\ IDF_j = \frac{1}{\sum_{x \in T}(TF_j^x)} \quad (3)$$

5.4 Integration of Prep and Verb models

The employ two strategies to integrate the models using prepositional and the verb vector:

(i) Concatenation Model concatenates the features of preposition and verb vectors to form a single Prep+Verb vector of $m + k$ features. The relevance of verb and preposition vector features are weighted with a contribution factor f.

$$\langle\, Verb{+}Prep \,\rangle = \langle\, Prep \,\rangle \oplus f * \langle\, Verb \,\rangle \quad (4)$$

where $\oplus$ denotes concatenation of two vectors.

(ii) Best Selection Model employs Best-Selection strategy by selecting the more confident of two models for classification in a given situation. This model separately evaluates for preposition and verb model the performance (i.e. $f - score$) of classifying each relation. Given a unseen instance, the two models predict the relation of NC independently but the model which assigns the relation with higher f-score is ultimately selected for classification.

5.5 Experiments

We perform *three* progressive experiments on the Statistical model on the SemEval dataset:

Experiment I: Comparing models on different parameters: In this experiment, we introduce 6 models on *three* varying parameters and compare their performance: NC vector (**-R**) vs Relation vector (**+R**), Weighted vector (**+W**) vs Unweighted vector (**-W**), Prior Probability (**+P**) vs Unit vector (**-P**). The experiments are conducted separately on prepositional and verbal vectors. The data is divided into training and testing set with *k-fold cross-validation*, k varying as, $k = 5, 7, 10, 15, 30, 50$ and $N - 1$.

Model	F	A	Model	F	A
1: [-R -W -P]	35.23	36.21	**4: [+R +W -P]**	35.77	38.54
2: [-R +W -P]	36.33	38.18	**5: [+R -W +P]**	37.97	39.34
3: [+R -W -P]	24.41	35.9	**6: [+R +W +P]**	38.82	40.72

Table 6: Performance of Models on Prepositional vector

The description of the models with their performance on prepositional vector is presented in Table 6. The **Model 6 [+R +W +P]** (i.e. *Model using weighted prior probability Relation vectors*) outperforms other models on both preposition and verb vectors. This model achieves an accuracy of 40.72% (baseline 30.55%) and f-score of 38.82% with prepositional vector and (Acc, F) of (34.93%, 35.78%) with verb vector at $k = N - 1$. Therefore, Model 6 is selected for the next two experiments.

Experiment II: Investigating the relevance of Verb+Prep paraphrases: This experiment investigates the relevance of verb+prep paraphrases (e.g. *'held during'*) over preposition and verb paraphrases. We have discussed that these paraphrases are splitted into preposition (i.e. *'during'*) and verb (*'hold'*) and contribute to the frequencies of corresponding features in preposition and verb vector, with respective weighting factors S_p and S_v, referred as the Strength parameters. Thus, a higher value of S_p and S_v means greater contribution of verb+preposition paraphrases in the classification model. The Strength parameters in Experiment I were fixed to $S_p = 1$ and $S_v = 1$ but are varied in this experiment from values 1 to 15.

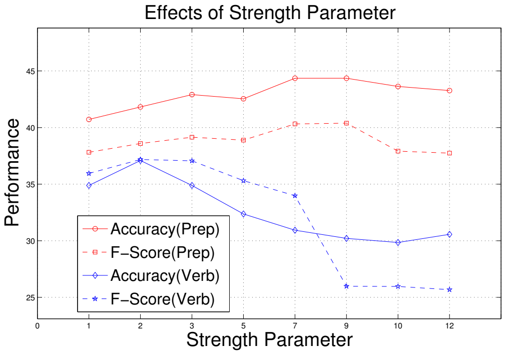

Figure 2: Performance of Preposition and Verb models on varying the Strength parameters S_p and S_v

The effects of Strength parameter on the preposition and the verb models on SemEval dataset are shown in Figure 2. The performance of prepositional model improves drastically (Acc, F) from (40.72%, 38.82%) to (44.36%, 40.39%) between values 1 to 9 (~4% improvement in accuracy and ~2.5% in f-score) and then drops down. The verb vector achieves best results at $S_v = 2$. This proves two things: First, verb+prep paraphrases have crucial contribution in the model and thus, finding such paraphrases in corpus is important, and secondly, the high value of $S_p = 9$ reveals that prepositions in verb+prep paraphrases are in fact quite relevant.

Experiment III: Integrating the Preposition and Verb models: In this experiment, we compare the Concatenation model and Best-Selection model for integrating the Prepositional and Verbal models. The experiment is performed on optimal parameters learnt from previous experiments, i.e. *Model 6* with $S_p = 9$ and $S_v = 2$ at $k = N - 1$. The Concatenation model concatenates the preposition and the verb feature vectors to form a single vector, with Contribution factor f varying from 0 to 2 in steps of 0.2. The Best-Selection model evaluates the performance of each relation on both the models and given a unseen instance, selects the model which classifies the relation with higher $f - score$.

The concatenation of prepositional and verbal features in the Concatenation model degrades the performance at every contribution factor f, achieving the best accuracy of only 36.72% with 35.16% f-score when both vectors are equally weighted at $f = 1$, shown in Figure 3. This shows that the significance of preposition features is diluted by the less significant verb features. On the other hand, the performance with Best-Selection model shoots up, which achieves accuracy of 46% with drastic improvement of ~ 7 in f-score, reaching 47.19%.

5.6 Observations

The Best-Selection model integrating the prepositional model and verbal model is selected as the best Statistical model, with optimal parameters $S_p = 9$ and $S_v = 2$ at fold $k = N - 1$. It uses Weighted Relation vector incorporating Prior probability [+R +W +P] for both preposition and verb feature vectors. This model achieves 46.04% accuracy (baseline 30.5%) and 47.19% f-score on SemEval dataset and hugely outperforms the ontology model, which performs just above the baseline achieving accuracy of only 34.53% and f-score of 36.56%.

The corpus model performs below the expectations on Nastase dataset, with performance subpar to the ontology model, achieving accuracy of 52.3% (~2% less than ontology model) with low f-score of 35.1%. The main reason for this is the insuffi-

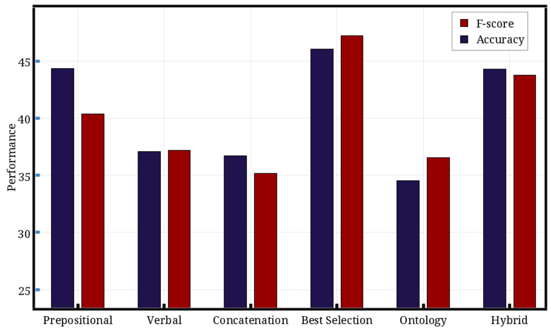

Figure 3: Comparison of All Models on SemEval Data

cient and poor quality of paraphrases obtained from the corpus, mainly verb and verb+preposition paraphrases, which are nowhere close to the human annotated paraphrases. We have selected only those NCs for the experiments for which atleast 3 paraphrases are found. Thus, experiment is performed with 241 NCs (out of 326 Noun-Noun pairs) for which this criteria is satisfied. An interesting property of the Relation vector is that it maps the lexical terms (*i.e. prepositions and verbs*) to semantic relations, and ranks them in decreasing order of their co-occurence with the relation. Table 7 presents the *five* top weighted prepositions for each relation.

Relation	Examples	Top-5 Prepositions
Causal	*advertisement agency, cancer death*	for, with, against, from, on
Quality	*trade statistics, wafer buscuit*	like, about, as, of, on
Spatial	*garden party, village school*	towards, near, at, in, around
Temporal	*spring weather, summer meeting*	during, after, in, at, from
Participant	*army coup, class party*	by, from, of, in, for

Table 7: Top-5 Relevant prepositions for each relation

6 Hybrid Model

The goal of the Hybrid model is to integrate knowledge of two very different models: one using the knowledge from a ontology while other deriving it from a corpus. The model employs a Best-Selection strategy which does nothing more than selecting the more suitable model for classification for any given test instance. Therefore, for the model to be efficient, it must satisfy two conditions:

a) The constituent models must be complimenting.

b) The model must have a selection criteria that works efficient in different circumstances. We find the two models to be complementing as the statistical model identifies some relations more accurately than ontology model and vice-versa, as discussed in Section 4.2. Further, we find that the performance of ontology model improves with each level of specialization (in Table 2). This insight is useful in implementing the selection criteria. The model computes a *Preference Score, P* for each model and selects the model with higher score for classifying the unseen instance. For ontology model, the *f-score* of each relation, r_i at each boundary level G_k is evaluated. Similarly, the *f-score* of each relation, r_i is evaluated for corpus model. Now, given a unseen instance, the following decision is taken:

$$P^{Ont}(r_1)^X > P^{Cor}(r_2), \quad then\ R^* = r_1; \\ else\ R^* = r_2 \qquad (5)$$

where $P^{Ont}(r_1)^{G_k}$ is the f-score of relation r_1 at boundary level G_k and R^* is the assigned relation.

The Hybrid model on the data of 241 NCs (on which corpus model is evaluated) performs quite well and outperforms the ontology and corpus models by 4.5% and 6.5% respectively, as shown in Table 8. These results are slightly better than the state-of-the-art system tested on this dataset (Turney, 2006b) but are below when compared on complete dataset of 593 NCs (out of which 352 NCs use only ontology model). The overall performance on Nastase dataset of 593 NCs achieves 55.31% accuracy with 49.47% f-score. On SemEval dataset, the performance of statistical model drops by ~2% when integrated with ontology model, which performs poorly on this dataset, as shown in Figure 3.

Relation	Ontology (593 NC)	Corpus (241 NC)	Hybrid (241 NC)	Hybrid (593 NC)
Quality	45	39.18	49.59	49.59
Temporal	78.26	55.81	75	75
Spatial	29.41	14.81	35.71	35.71
Participant	64.6	65.72	67.78	67.78
Causal	29.09	0	34.78	34.78
Macro-Avg F	49.27	35.11	52.57	52.57
Accuracy	54.35	52.28	58.92	55.31

Table 8: Comparison of Models on Nastase Dataset

7 Conclusion

This paper presents a Statistical VSM-based model which represents each relation with a vector of prepositional and verbal paraphrases. The statistical model needs to solve two problems: **(i)** Identifying which paraphrases are relevant in disambiguating the relations, which is challenging (Nastase et al., 2006; Nulty, 2007); and **(ii)** Finding those paraphrases in corpus for given NC is hard (Surtani et al., 2013). We work extensively to improve on the first part of the problem, but we fail in finding good set of paraphrases from the corpus. The statistical model has shown huge potential over the ontology model (which also requires WordNet senses of modifier & head, a challenging task (WSD)). The future task is to achieve a better Paraphrase Extraction system.

References

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).

Brandon Beamer, Alla Rozovskaya and Roxana Girju. 2008. *Automatic Semantic Relation Extraction with Multiple Boundary Generation*. In AAAI (pp 824-829).

Cristina Butnariu and Tony Veale. 2008. *A concept-centered approach to noun-compound interpretation*. Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid O Séaghdha, Stan Szpakowicz and Tony Veale. 2010. *Semeval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions*. In Proceedings of the 5th SIGLEX Workshop on Semantic Evaluation.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid O Séaghdha, Stan Szpakowicz and Tony Veale. 2013. *Semeval-13 task 4: Free Paraphrases of Noun Compounds*. In Proceedings of the International Workshop on Semantic Evaluation, Atlanta, Georgia.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe and Roxana Girju. 2004. *Models for the Semantic Classification of Noun Phrases*. In the proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics . Boston, MA.

Diarmuid O Séaghdha. 2008. *Learning compound noun semantics*. University of Cambridge, Cambridge, UK.

Marti Hearst. 1998. *Automated discovery of WordNet relations*. WordNet: an electronic lexical database, 131-151.

Nitesh Surtani, Arpita Batra, Urmi Ghosh and Soma Paul. 2013. *IIITH: A Corpus-Driven Co-occurrence Based Probabilistic Model for Noun Compound Paraphrasing*. In Proceedings of the International Workshop on Semantic Evaluation, Atlanta, Georgia.

Olutobi Owoputi, Brendan O Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith 2013. *Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters*. HLT-NAACL.

Pamela Downing. 1977. On the creation and use of English noun compounds. *Language*, 53(4): 810-842.

Paul Nulty. 2007. *Semantic classification of noun phrases using web counts and learning algorithms*. In Proceedings of the ACL 2007 Student Research Workshop (ACL-07), pages 79-84.

Peter D Turney and Michael L. Littman. 2005. *Corpus-based learning of analogies and semantic relations*. Machine Learning, 60(1-3):251-278.

Peter D Turney. 2006. *Expressing implicit semantic relations without supervision*. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL-06), Sydney, Australia.

Peter D Turney. 2006. *Similarity of semantic relations*. Computational Linguistics, 32 (3), 379-416.

Preslav Nakov and and Marti Hearst. 2006. *Using verbs to characterize noun-noun relations*. Artificial Intelligence: Methodology, Systems, and Applications. Springer Berlin Heidelberg.

Preslav Nakov. 2008. *Noun compound interpretation using paraphrasing verbs: Feasibility study*. Artificial Intelligence: Methodology, Systems, and Applications. Springer Berlin Heidelberg, 2008. 103-117.

Roxana Girju, Adriana Badulescu and Dan Moldovan. 2003. *Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics.

Roxana Girju, Dan Moldovan, Marta Tatu and Daniel Antohe. 2005. *On the semantics of noun compounds*. Computer, Speech and Language 19(4):479-496.

Roxana Girju, Adriana Badulescu and Dan Moldovan. 2006. *Automatic Discovery of Part-Whole Relations*. In Computational Linguistics.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter D. Turney, Deniz Yuret. 2007. *Semeval-2007 task 04: Classification of semantic relations between nominals.*. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 13-18). Association for Computational Linguistics.

Satanjeev Banerjee, Ted Pedersen. 2002. *An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet - In the Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, pp.* 136-145, Mexico City.

Su Nam Kim and Timothy Baldwin. 2006. *Interpreting semantic relations in noun compounds via verb semantics*. Proc. ACL-06 Main Conference Poster Session, Sydney, Australia, 491-498.

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.

Thorsten Brants, Alex Franz. 2006. *Web 1T 5-gram Version1*. Linguistic Data Consortium.

Timothy Baldwin and Takaaki Tanaka. 2004. *Translation by machine of compound nominals: Getting it* *right*. In Proceedings of ACL-2004 Workshop on Multiword Expressions: Integrating Processing.

Vivek Srikumar and Dan Roth. 2013. *Modeling semantic relations expressed by prepositions*. Transactions of Association of Computational Linguistics.

Vivi Nastase and Stan Szpakowicz. 2003. *Exploring noun-modifier semantic relations*. In Fifth International Workshop on Computational Semantics (IWCS-5), pages 285-301, Tilburg, The Netherlands.

Vivi Nastase, Jelber Sayyad Shirabad, Marina Sokolova and Stan Szpakowicz. 2006. *Learning noun-modifier semantic relations with corpus-based and Wordnet-based features*. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), pages 781-787.

Vivi Nastase, Preslav Nakov, Diarmuid O Séaghdha and Stan Szpakowicz. 2014. *Semantic Relations between Nominals*.