Combining Heterogeneous Models for Measuring Relational Similarity

Alisa Zhila*
Instituto Politecnico Nacional
Mexico City, Mexico
alisa.zhila@gmail.com

Wen-tau Yih **Christopher Meek**
Microsoft Research
Redmond, WA 98052, USA
{scottyih,meek}@microsoft.com

Geoffrey Zweig
Microsoft Research
Redmond, WA 98052, USA
gzweig@microsoft.com

Tomas Mikolov*
BRNO University of Technology
BRNO, Czech Republic
tmikolov@gmail.com

Abstract

In this work, we study the problem of measuring relational similarity between two word pairs (e.g., *silverware:fork* and *clothing:shirt*). Due to the large number of possible relations, we argue that it is important to combine multiple models based on heterogeneous information sources. Our overall system consists of two novel general-purpose relational similarity models and three specific word relation models. When evaluated in the setting of a recently proposed SemEval-2012 task, our approach outperforms the previous best system substantially, achieving a 54.1% relative increase in Spearman's rank correlation.

1 Introduction

The problem of measuring relational similarity is to determine the degree of correspondence between two word pairs. For instance, the analogous word pairs *silverware:fork* and *clothing:shirt* both exemplify well a *Class-Inclusion:Singular_Collective* relation and thus have high relational similarity. Unlike the problem of *attributional similarity*, which measures whether two words share similar attributes and is addressed in extensive research work (Budanitsky and Hirst, 2006; Reisinger and Mooney, 2010; Radinsky et al., 2011; Agirre et al., 2009; Yih and Qazvinian, 2012), measuring relational similarity is a relatively new research direction pioneered by Turney (2006), but with many potential applications. For instance, problems of identifying specific relations between words, such as synonyms, antonyms or associations, can be reduced to measuring relational similarity compared to prototypical word pairs with the desired relation (Turney, 2008). In scenarios like information extraction or question answering, where identifying the existence of certain relations is often the core problem, measuring relational similarity provides a more flexible solution rather than creating relational classifiers for predefined or task-specific categories of relations (Turney, 2006; Jurgens et al., 2012).

In order to promote this research direction, Jurgens et al. (2012) proposed a new shared task of measuring relational similarity in SemEval-2012 recently. In this task, each submitted system is required to judge the degree of a target word pair having a particular relation, measured by its relational similarity compared to a few prototypical example word pairs. The system performance is evaluated by its correlation with the human judgments using two evaluation metrics, Spearman's rank correlation and MaxDiff accuracy (more details of the task and evaluation metrics will be given in Sec. 3). Although participating systems incorporated substantial amounts of information from lexical resources (e.g., WordNet) and contextual patterns from large corpora, only one system (Rink and Harabagiu, 2012) is able to outperform a simple baseline that uses PMI (pointwise mutual information) scoring, which demonstrates the difficulty of this task.

In this paper, we explore the problem of measuring relational similarity in the same task setting. We argue that due to the large number of possible relations, building an ensemble of relational simi-

*Work conducted while interning at Microsoft Research.

Proceedings of NAACL-HLT 2013, pages 1000–1009,
Atlanta, Georgia, 9–14 June 2013.

larity models based on heterogeneous information sources is the key to advance the state-of-the-art on this problem. By combining two general-purpose relational similarity models with three specific word-relation models covering relations like IsA and synonymy/antonymy, we improve the previous state-of-the-art substantially – having a relative gain of 54.1% in Spearman's rank correlation and 14.7% in the MaxDiff accuracy!

Our main contributions are threefold. First, we propose a novel directional similarity method based on the vector representation of words learned from a recurrent neural network language model. The relation of two words is captured by their vector offset in the latent semantic space. Similarity of relations can then be naturally measured by a distance function in the vector space. This method alone already performs better than all existing systems. Second, unlike the previous finding, where SVMs learn a much poorer model than naive Bayes (Rink and Harabagiu, 2012), we show that using a highly-regularized log-linear model on simple contextual pattern features collected from a document collection of 20GB, a discriminative approach can learn a strong model as well. Third, we demonstrate that by augmenting existing word-relation models, which cover only a small number of relations, the overall system can be further improved.

The rest of this paper is organized as follows. We first survey the related work in Sec. 2 and formally define the problem in Sec. 3. We describe the individual models in detail in Sec. 4. The combination approach is depicted in Sec. 5, along with experimental comparisons to individual models and existing systems. Finally, Sec. 6 concludes the paper.

2 Related Work

Building a classifier to determine whether a relationship holds between a pair of words is a natural approach to the task of measuring relational similarity. While early work was mostly based on hand-crafted rules (Finin, 1980; Vanderwende, 1994), Rosario and Hearst (2001) introduced a machine learning approach to classify word pairs. They targeted classifying noun modifier pairs from the medical domain into 13 classes of semantic relations. Features for each noun modifier pair were constructed using large medical lexical resources and a multi-class classifier was trained using a feed-forward neural network with one hidden layer. This work was later extended by Nastase and Szpakowicz (2003) to classify general domain noun-modifier pairs into 30 semantic relations. In addition to extracting features using WordNet and Roget's Thesaurus, they also experimented with several different learners including decision trees, memory-based learning and inductive logic programming methods like RIPPER and FOIL. Using the same dataset as in (Nastase and Szpakowicz, 2003), Turney and Littman (2005) created a 128-dimentional feature vector for each word pair based on statistics of their co-occurrence patterns in Web documents and applied the k-NN method ($k = 1$ in their work).

Measuring relational similarity, which determines whether two word pairs share the same relation, can be viewed as an extension of classifying relations between two words. Treating a relational similarity measure as a distance metric, a testing pair of words can be judged by whether they have a relation that is *similar* to some prototypical word pairs having a particular relation. A multi-relation classifier can thus be built easily in this framework as demonstrated in (Turney, 2008), where the problems of identifying synonyms, antonyms and associated words are all reduced to finding good analogous word pairs. Measuring relational similarity has been advocated and pioneered by Turney (2006), who proposed a latent vector space model for answering SAT analogy questions (e.g., *mason*:*stone* vs. *carpenter*:*wood*). In contrast, we take a slightly different view when building a relational similarity measure. Existing classifiers for specific word relations (e.g., synonyms or Is-A) are combined with general relational similarity measures. Empirically, mixing heterogeneous models tends to make the final relational similarity measure more robust.

Although datasets for semantic relation classification or SAT analogous questions can be used to evaluate a relational similarity model, their labels are either binary or categorical, which makes the datasets suboptimal for determining the quality of a model when evaluated on instances of the same relation class. As a result, Jurgens et al. (2012) proposed a new task of "Measuring Degrees of Relational Similarity" at SemEval-2012, which includes 79 relation

categories exemplified by three or four prototypical word pairs and a schematic description. For example, for the *Class-Inclusion:Taxonomic* relation, the schematic description is *"Y is a kind/type/instance of X"*. Using Amazon Mechanical Turk[1], they collected word pairs for each relation, as well as their degrees of being a good representative of a particular relation when compared with defining examples. Participants of this shared task proposed various kinds of approaches that leverage both lexical resources and general corpora. For instance, the Duluth systems (Pedersen, 2012) created word vectors based on WordNet and estimated the degree of a relation using cosine similarity. The BUAP system (Tovar et al., 2012) represented each word pair as a whole by a vector of 4 different types of features: context, WordNet, POS tags and the average number of words separating the two words in text. The degree of relation was then determined by the cosine distance of the target pair from the prototypical examples of each relation. Although their models incorporated a significant amount of information of words or word pairs, unfortunately, the performance were not much better than a random baseline, which indicates the difficulty of this task. In comparison, a supervised learning approach seems more promising. The UTD system (Rink and Harabagiu, 2012), which mined lexical patterns between co-occurring words in the corpus and then used them as features to train a Naive Bayes classifier, achieved the best results. However, potentially due to the large feature space, this strategy did not work as well when switching the learning algorithm to SVMs.

3 Problem Definition & Task Description

Following the setting of SemEval-2012 Task 2 (Jurgens et al., 2012), the problem of measuring the degree of relational similarity is to rate word pairs by the degree to which they are prototypical members of a given relation class. For instance, comparing to the prototypical word pairs, {*cutlery:spoon*, *clothing:shirt*, *vermin:rat*} of the *Class-Inclusion:Singular_Collective* relation, we would like to know among the input word pairs {*dish:bowl*, *book:novel*, *furniture:desk*}, which one best demonstrates the relation.

Because our approaches are evaluated using the data provided in this SemEval-2012 task, we describe briefly below how the data was collected, as well as the metrics used to evaluate system performance. The dataset consists of 79 relation classes that are chosen according to (Bejar et al., 1991) and broadly fall into 10 main categories, including *Class-Inclusion*, *Part-Whole*, *Similar* and more. With the help of Amazon Mechanical Turk, Jurgens et al. (2012) used a two-phase approach to collect word pairs and their degrees. In the first phase, a lexical schema, such as *"a Y is one item in a collection/group of X"* for the aforementioned relation *Class-Inclusion:Singular_Collective*, and a few prototypical pairs for each class were given to the workers, who were asked to provide approximately a list of 40 word pairs representing the same relation class. Naturally, some of these pairs were better examples than the others. Therefore, in the second phase, the goal was to measure the degree of their similarity to the corresponding relation. This was done using the MaxDiff technique (Louviere and Woodworth, 1991). For each relation, about one hundred questions were first created. Each question consists of four different word pairs randomly sampled from the list. The worker was then asked to choose the most and least representative word pairs for the specific relation in each question.

The set of 79 word relations were randomly split into *training* and *testing* sets. The former contains 10 relations and the latter has 69. Word pairs in all 79 relations were given to the task participants in advance, but only the human judgments of the *training* set were available for system development. In this work, we treat the training set as the validation set – all the model exploration and refinement is done using this set of data, as well as the hyper-parameter tuning when learning the final model combination.

The quality of a relational similarity measure is estimated by its correlation to human judgments. This is evaluated using two metrics in the task: the MaxDiff accuracy and Spearman's rank correlation coefficient (ρ). A system is first asked to pick the most and least representative word pairs of each question in the MaxDiff setting. The average accuracy of the predictions compared to the human answers is then reported. In contrast, Spearman's ρ

[1] http://www.mturk.com

measures the correlation between the total orderings of all word pairs of a relation, where the total ordering is derived from the MaxDiff answers (see (Jurgens et al., 2012) for the exact procedure).

4 Models for Relational Similarity

We investigate three types of models for relational similarity. Operating in a word vector space, the *directional similarity* model compares the vector differences of target and prototypical word pairs to estimate their relational similarity. The *lexical pattern* method collects contextual information of pairs of words when they co-occur in large corpora, and learns a highly regularized log-linear model. Finally, the *word relation* models incorporate existing, specific word relation measures for general relational similarity.

4.1 Directional Similarity Model

Our first model for relational similarity extends previous work on semantic word vector representations to a directional similarity model for pairs of words. There are many different methods for creating real-valued semantic word vectors, such as the distributed representation derived from a word co-occurrence matrix and a low-rank approximation (Landauer et al., 1998), word clustering (Brown et al., 1992) and neural-network language modeling (Bengio et al., 2003; Mikolov et al., 2010). Each element in the vectors conceptually represents some latent topicality information of the word. The goal of these methods is that words with similar meanings will tend to be close to each other in the vector space.

Although the vector representation of *single* words has been successfully applied to problems like semantic word similarity and text classification (Turian et al., 2010), the issue of how to represent and compare pairs of words in a vector space remains unclear (Turney, 2012). In a companion paper (Mikolov et al., 2013), we present a vector offset method which performs consistently well in identifying both syntactic and semantic regularities. This method measures the degree of the analogy "a is to b as c is to d" using the cosine score of $(\vec{v}_b - \vec{v}_a + \vec{v}_c, \vec{v}_d)$, where a, b, c, d are the four given words and $\vec{v}_a, \vec{v}_b, \vec{v}_c, \vec{v}_d$ are the corresponding vectors. In this paper, we propose a variant called the *directional similarity* model, which performs better for semantic relations. Let $\omega_i = (w_{i_1}, w_{i_2})$ and $\omega_j = (w_{j_1}, w_{j_2})$ be the two word pairs being compared. Suppose $(\vec{v}_{i_1}, \vec{v}_{i_2})$ and $(\vec{v}_{j_1}, \vec{v}_{j_2})$ are the corresponding vectors of these words. The *directional vectors* of ω_i and ω_j are defined as $\vec{v}_i \equiv \vec{v}_{i_2} - \vec{v}_{i_1}$ and $\vec{v}_j \equiv \vec{v}_{j_2} - \vec{v}_{j_1}$, respectively. Relational similarity of these two word pairs can be measured by some distance function of v_i and v_j, such as the cosine function:

$$\frac{\vec{v}_i \cdot \vec{v}_j}{\|\vec{v}_i\| \|\vec{v}_j\|}$$

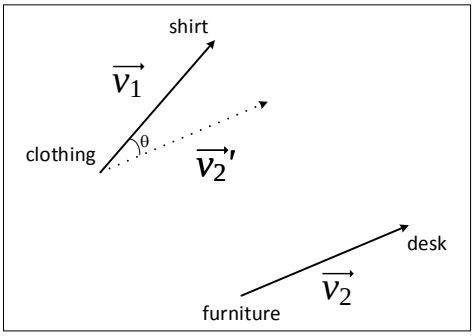

Figure 1: Directional vectors v_1 and v_2 capture the relations of *clothing:shirt* and *furniture:desk* respectively in this semantic vector space. The relational similarity of these two word pairs is estimated by the cosine of θ.

The rationale behind this variant is as follows. Because the difference of two word vectors reveals the change from one word to the other in terms of multiple topicality dimensions in the vector space, two word pairs having similar offsets (i.e., being relatively parallel) can be interpreted as they have similar relations. Fig. 1 further illustrates this method.

Compared to the original method, this variant places less emphasis on the similarity between words w_{j_1} and w_{j_2}. That similarity is necessary for syntactic relations where the words are often related by morphology, but not for semantic relations. On semantic relations studied in this paper, the *directional similarity* model performs about 18% relatively better in Spearman's ρ than the original one.

The quality of the directional similarity method depends heavily on the underlying word vector space model. We compared two choices with dif-

Word Embedding	Spearman's ρ	MaxDiff Acc. (%)
LSA-80	0.055	34.6
LSA-320	0.066	34.4
LSA-640	0.102	35.7
RNNLM-80	0.168	37.5
RNNLM-320	0.214	39.1
RNNLM-640	0.221	39.2
RNNLM-1600	0.234	41.2

Table 1: Results of measuring relational similarity using the directional similarity method, evaluated on the training set. The 1600-dimensional RNNLM vector space achieves the highest Spearman's ρ and MaxDiff accuracy.

ferent dimensionality settings: the word embedding learned from the recurrent neural network language model (RNNLM)[2] and the LSA vectors, both were trained using the same Broadcast News corpus of 320M words as described in (Mikolov et al., 2011). All the word vectors were first normalized to unit vectors before applying the directional similarity method. Given a target word pair, we computed its relational similarity compared with the prototypical word pairs of the same relation. The average of these measurements was taken as the final model score. Table 1 summarizes the results when evaluated on the training set. As shown in the table, the RNNLM vectors consistently outperform their LSA counterparts with the same dimensionality. In addition, more dimensions seem to preserve more information and lead to better performance. Therefore, we take the 1600-dimensional RNNLM vectors to construct our final directional similarity model.

4.2 Lexical Pattern Model

Our second model for measuring relational similarity is built based on lexical patterns. It is well-known that contexts in which two words co-occur often provide useful cues for identifying the word relation. For example, having observed frequent text fragments like *"X such as Y"*, it is likely that there is a *Class-Inclusion:Taxonomic* relation between *X* and *Y*; namely, *Y* is a type of *X*. Indeed, by mining lexical patterns from a large corpus, the UTD system (Rink and Harabagiu, 2012) managed to outperform other participants in the SemEval-2012 task of measuring relational similarity.

[2] http://www.fit.vutbr.cz/~imikolov/rnnlm

In order to find more co-occurrences of each pair of words, we used a large document set that consists of the Gigaword corpus (Parker et al., 2009), Wikipedia and LA Times articles[3], summing up to more than 20 Gigabytes of texts. For each word pair (w_1, w_2) that co-occur in a sentence, we collected the words in between as its *context* (or so-called "raw pattern"). For instance, *"such as"* would be the context extracted from *"X such as Y"* for the word pair (*X*, *Y*). To reduce noise, contexts with more than 9 words were dropped and 914,295 patterns were collected in total.

Treating each raw pattern as a feature where the value is the logarithm of the occurrence count, we then built a probabilistic classifier to determine the association of the context and relation. For each relation, we treated all its word pairs as positive examples and all the word pairs in other relations as negative examples[4]. 79 classifiers were trained in total, where each one was trained using 3,218 examples. The degree of relational similarity of each word pair can then be judged by the output of the corresponding classifier[5]. Although this seems like a standard supervised learning setting, the large number of features poses a challenge here. Using almost 1M features and 3,218 examples, the model could easily overfit if not regularized properly, which may explain why learning SVMs on pattern features performed poorly (Rink and Harabagiu, 2012). Instead of employing explicit feature selection methods, we used an efficient L_1 regularized log-linear model learner (Andrew and Gao, 2007) and chose the hyper-parameters based on model performance on the training data. The final models we chose were trained with $L_1 = 3$, where 28,065 features in average were selected automatically by the algo-

[3] We used a Nov-2010 dump of English Wikipedia, which contains approximately 917M words after pre-processing. The LA Times corpus consists of articles from 1985 to 2002 and has about 1.1B words.

[4] Given that not all word pairs belonging to the same relation category are equally good, removing those with low judgment scores may help improve the quality of the labeled data. We leave this study to future work.

[5] Training a separate classifier for each MaxDiff question using all words pairs except the four target pairs appears to be a better setting, as it would avoid including the target pairs in the training process. We did not use this setting because it is more complicated and performed roughly the same empirically.

rithm. The performance on the training data is 0.322 in Spearman's ρ and 41.8% in MaxDiff accuracy.

4.3 Word Relation Models

The directional similarity and lexical pattern models can be viewed as *general* purpose methods for relational similarity as they do not differentiate the specific relation categories. In contrast, for *specific* word relations, there exist several high-quality methods. Although they are designed for detecting specific relations between words, incorporating them could still improve the overall results. Next, we explore the use of some of these word relation models, including information encoded in the knowledge base and a lexical semantic model for synonymy and antonymy.

4.3.1 Knowledge Bases

Predetermined types of relations can often be found in existing lexical and knowledge databases, such as WordNet's Is-A taxonomy and the extensive relations stored in the NELL (Carlson et al., 2010) knowledge base. Although in theory, these resources can be directly used to solve the problem of relational similarity, such direct approaches often suffer from two practical issues. First, the word coverage of these databases is usually very limited and it is common that the relation of a given word pair is absent. Second, the degree of relation is often not included, which makes the task of measuring the degree of relational similarity difficult.

One counter example, however, is Probase (Wu et al., 2012), which is a knowledge base that establishes connections between more than 2.5 million concepts discovered automatically from the Web. For the *Is-A* and *Attribute* relations it encodes, Probase also returns the probability that two input words share the relation, based on the co-occurrence frequency. We used some relations in the training set to evaluate the quality of Probase. For instance, its Is-A model performs exceptionally well on the relation *Class-Inclusion:Taxonomic*, reaching a high Spearman's $\rho = 0.642$ and MaxDiff accuracy 55.8%. Similarly, its Attribute model performs better than our lexical pattern model on *Attribute:Agent_Attribute-State* with Spearman's $\rho = 0.290$ and MaxDiff accuracy 32.7%.

4.3.2 Lexical Semantics Measures

Most lexical semantics measures focus on the semantic similarity or relatedness of two words. Since our task focuses on distinguishing the difference between word pairs in the *same* relation category. The crude relatedness model does not seem to help in our preliminary experimental study. Instead, we leverage the recently proposed polarity-inducing latent semantic analysis (PILSA) model (Yih et al., 2012), which specifically estimates the degree of synonymy and antonymy. This method first forms a signed co-occurrence matrix using synonyms and antonyms in a thesaurus and then generalizes it using a low-rank approximation derived by SVD. Given two words, the cosine score of their PILSA vectors tend to be negative if they are antonymous and positive if synonymous. When tested on the *Similar:Synonymity* relation, it has a Spearman's $\rho = 0.242$ and MaxDiff accuracy 42.1%, both are better than those of our directional similarity and lexical pattern models.

5 Model Combination

In order to fully leverage the diverse models proposed in Sec. 4, we experiment with a model combination approach and conduct a model ablation study. Performance of the combined and individual models is evaluated using the *test* set and compared with existing systems.

We seek an optimal linear combination of all the individual models by treating their output as *features* and use a logistic regression learner to learn the weights[6]. The training setting is essentially the same as the one used to learn the lexical pattern model (Sec. 4.2). For each relation, we treat all the word pairs in this relation group as positive examples and all other word pairs as negative ones. Consequently, 79 sets of weights for model combination are learned in total. The average Spearman's ρ of the 10 training relations is used for selecting the values of the L_1 and L_2 regularizers[7]. Evaluated on the remaining 69 relations (i.e., the test set), the average results of each main relation group and the overall

[6]Nonlinear methods, such as MART (Friedman, 2001), do not perform better in our experiments (not reported here).

[7]We tested 15 combinations, where $L_1 \in \{0, 0.01, 0.1\}$ and $L_2 \in \{0, 0.001, 0.01, 1, 10\}$. The parameter setting that gave the highest Spearman rank correlation coefficient score on the training set was selected.

Relation Group	Rand.	BUAP	Duluth$_{V0}$	UTD$_{NB}$	DS	Pat.	IsA	Attr.	PILSA	Com.
Class-Inclusion	0.057	0.064	0.045	0.233	0.350	0.422	**0.619**	-0.137	0.029	0.519
Part-Whole	0.012	0.066	-0.061	0.252	0.317	0.244	-0.014	0.026	-0.010	**0.329**
Similar	0.026	-0.036	0.183	0.214	0.254	0.245	-0.020	0.133	0.058	**0.303**
Contrast	-0.049	0.000	0.142	0.206	0.063	**0.298**	-0.012	-0.032	-0.079	0.268
Attribute	0.037	-0.095	0.044	0.158	**0.431**	0.198	-0.008	0.016	-0.052	0.406
Non-Attribute	-0.070	0.009	0.079	0.098	0.195	0.117	0.036	0.078	-0.093	**0.296**
Case Relations	0.090	-0.037	-0.011	0.241	**0.503**	0.288	0.076	-0.075	0.059	0.473
Cause-Purpose	-0.011	0.114	0.021	0.183	**0.362**	0.234	0.044	-0.059	0.038	0.296
Space-Time	0.013	0.035	0.055	0.375	0.439	0.248	0.064	-0.002	-0.018	**0.443**
Reference	0.142	-0.001	0.028	**0.346**	0.301	0.119	0.033	-0.123	0.021	0.208
Average	0.018	0.014	0.050	0.229	0.324†	0.235	0.058‡	-0.010‡	-0.009‡	**0.353**‡

Relation Group	Rand.	BUAP	Duluth$_{V0}$	UTD$_{NB}$	DS	Pat.	IsA	Attr.	PILSA	Com.
Class-Inclusion	30.1	29.0	26.7	39.1	46.7	43.4	**59.6**	24.7	32.3	51.2
Part-Whole	31.9	35.1	29.4	40.9	**43.9**	38.1	31.3	29.5	31.0	42.9
Similar	31.5	29.1	37.1	39.8	38.5	38.4	30.8	36.3	34.2	**43.3**
Contrast	30.4	32.4	38.3	40.9	33.6	42.2	32.3	31.8	30.1	**42.8**
Attribute	30.2	29.2	31.9	36.5	47.9	38.3	30.7	31.0	28.8	**48.3**
Non-Attribute	28.9	30.4	36.0	36.8	38.7	36.7	32.3	32.8	27.7	**42.6**
Case Relations	32.8	29.5	28.2	40.6	**54.3**	42.2	32.8	25.7	31.0	50.6
Cause-Purpose	30.8	35.4	29.5	36.3	**45.3**	38.0	30.3	28.1	32.0	41.7
Space-Time	30.6	32.5	31.9	43.2	**50.0**	39.2	33.2	29.3	30.6	47.7
Reference	35.1	30.0	31.9	41.2	**45.7**	36.9	30.4	27.2	30.2	42.5
Average	31.2	31.7	32.4	39.4	44.5‡	39.2	33.3‡	29.8‡	30.7‡	**45.2**‡

Table 2: Average Spearman's ρ (Top) and MaxDiff accuracy (%) (Bottom) of each major relation group and all 69 testing relations. The best result in each row is highlighted in boldface font. Statistical significance tests are conducted by comparing each of our systems with the previous best performing system, UTD$_{NB}$. † and ‡ indicate the difference in the average results is statistically significant with 95% or 99% confidence level, respectively.

results are presented in Table 2. For comparison, we also show the performance of a random baseline and the best performing system of each participant in the SemEval-2012 task.

We draw two conclusions from this table. First, both of our general relational similarity models, the directional similarity (DS) and lexical pattern (Pat) models are fairly strong. The former outperforms the previous best system UTD$_{NB}$ in both Spearman's ρ and MaxDiff accuracy, where the differences are statistically significant[8]; the latter has comparable performance, where the differences are not statistically significant. In contrast, while the IsA relation from Probase is exceptionally good in identifying *Class-Inclusion* relations, with high Spearman's $\rho = 0.619$ and MaxDiff accuracy 59.6%, it does not have high correlations with human judgments in other relations. Like in the case of Probase Attribute and PILSA, specific word-relation models individually are not good measures for general relational similarity. Second, as expected, combining multiple diverse models (Com) is a robust strategy, which provides the best overall performance. It achieves superior results in both evaluation metrics compared to UTD$_{NB}$ and only a lower Spearman's ρ value in one of the ten relation groups (namely, *Reference*). The differences are statistically significant with p-value less than 10^{-3}.

In order to understand the interaction among different component models, we conducted an ablation study by iteratively removing one model from the final combination. The weights are re-trained using the same procedure that finds the best regularization parameters with the help of training data. Table 3 summarizes the results and compares them with the

[8]We conducted a paired-t test on the results of each of the 69 relation. The difference is considered statistically significant if the p-value is less than 0.05.

	Spearman's ρ						MaxDiff Accuracy (%)					
Relation Group	Com.	-Attr	-IsA	-PILSA	-DS	-Pat	Com.	-Attr	-IsA	-PILSA	-DS	-Pat
Class-Inclusion	0.519	0.557	0.467	**0.593**	0.490	0.570	51.2	53.7	49.2	54.6	49.3	**56.2**
Part-Whole	**0.329**	0.326	0.335	0.331	0.277	0.285	42.9	42.1	42.6	41.8	38.5	**42.9**
Similar	**0.303**	0.269	0.302	0.281	0.256	0.144	**43.3**	41.2	42.7	40.5	40.2	38.9
Contrast	0.268	0.234	0.267	**0.289**	0.260	0.156	**42.8**	42.0	42.4	41.5	42.7	38.1
Attribute	0.406	0.409	0.405	0.433	0.164	**0.447**	48.3	47.8	48.2	**49.1**	36.9	49.0
Non-Attribute	0.296	0.287	**0.296**	0.276	0.123	0.283	42.6	42.9	42.6	41.8	36.0	**43.0**
Case Relations	0.473	0.497	0.470	0.484	0.309	**0.498**	50.6	52.5	50.2	50.9	42.9	**53.2**
Cause-Purpose	0.296	0.282	0.299	**0.301**	0.205	0.296	41.7	41.6	41.6	41.2	36.6	**44.1**
Space-Time	**0.443**	0.425	0.443	0.420	0.269	0.431	47.7	47.2	47.7	46.9	40.5	**49.5**
Reference	0.208	**0.238**	0.205	0.168	0.102	0.210	42.5	42.3	**42.6**	41.8	36.1	41.4
Average	0.353	0.348	0.350	**0.354**	0.238‡	0.329	45.2	45.0	44.9‡	44.7	39.6‡	**45.4**

Table 3: Average Spearman's ρ and MaxDiff accuracy results of different model combinations. *Com* indicates combining all models, where other columns show the results when the specified model is removed. The best result in each row is highlighted in boldface font. Statistical significance tests are conducted by comparing each ablation configuration with *Com*. ‡ indicates the difference in the average results is statistically significant with 99% confidence level.

original combination model.

Overall, it is clear that the directional similarity method based on RNNLM vectors is the most critical component model. Removing it from the final combination decreases both the Spearman's ρ and MaxDiff accuracy by a large margin; both differences (Com vs. -DS) are statistically significant. The Probase IsA model also has an important impact on the performance on the *Class-Inclusion* relation group. Eliminating the IsA model makes the overall MaxDiff accuracy statistically significantly lower (Com vs. -IsA). Again, the benefits of incorporating Probase Attribute and PILSA models are not clear. Removing them from the final combination lowers the MaxDiff accuracy, but neither the difference in Spearman's ρ nor MaxDiff accuracy is statistically significant. Compared to the RNNLM directional similarity model, the lexical pattern model seems less critical. Removing it lowers the *Similar* and *Contrast* relation groups, but improves some other relation groups like *Class-Inclusion* and *Case Relations*. The final MaxDiff accuracy becomes slightly higher but the Spearman's ρ drops a little (Com vs. -Pat); neither is statistically significant.

Notice that the main purpose of the ablation study is to verify the importance of an individual component model when a significant performance drop is observed after removing it. However, occasionally the overall performance may go up slightly. Typically this is due to the fact that some models do not provide useful signals to a particular relation, but instead introduce more noise. Such effects can often be alleviated when there are enough quality training data, which is unfortunately not the case here.

6 Conclusions

In this paper, we presented a system that combines heterogeneous models based on different information sources for measuring relational similarity. Our two individual general-purpose relational similarity models, *directional similarity* and *lexical pattern* methods, perform strongly when compared to existing systems. After incorporating specific word-relation models, the final system sets a new state-of-the-art on the SemEval-2012 task 2 test set, achieving Spearman's $\rho = 0.353$ and MaxDiff accuracy 45.4% – resulting in 54.1% and 14.7% relative improvement in these two metrics, respectively.

Despite its simplicity, our directional similarity approach provides a robust model for relational similarity and is a critical component in the final system. When the lexical pattern model is included, our overall model combination method can be viewed as a two-stage learning system. As demonstrated in our work, with an appropriate regularization strategy, high-quality models can be learned in both stages. Finally, as we observe from the positive effect of adding the Probase IsA model, specific word-relation models can further help improve the system

although they tend to cover only a small number of relations. Incorporating more such models could be a steady path to enhance the final system.

In the future, we plan to pursue several research directions. First, as shown in our experimental results, the model combination approach does not always outperform individual models. Investigating how to select models to combine for each specific relation or relation group individually will be our next step for improving this work. Second, because the labeling process of relational similarity comparisons is inherently noisy, it is unrealistic to request a system to correlate human judgments perfectly. Conducting some user study to estimate the performance ceiling in each relation category may help us focus on the weaknesses of the final system to enhance it. Third, it is intriguing to see that the directional similarity model based on the RNNLM vectors performs strongly, even though the RNNLM training process is not related to the task of relational similarity. Investigating the effects of different vector space models and proposing some theoretical justifications are certainly interesting research topics. Finally, we would like to evaluate the utility our approach in other applications, such as the SAT analogy problems proposed by Turney (2006) and question answering.

Acknowledgments

We thank Richard Socher for valuable discussions, Misha Bilenko for his technical advice and anonymous reviewers for their comments.

References

E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca and A. Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL '09*, pages 19–27.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *ICML '07*.

I.I. Bejar, R. Chaffin, and S.E. Embretson. 1991. *Cognitive and psychometric analysis of analogical problem solving*. Recent research in psychology. Springer-Verlag.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32:13–47, March.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Timothy W. Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

J.H. Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Ann. Statist*, 29(5):1189–1232.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes, 25*, pages 259–284.

Jordan J. Louviere and G. G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048.

Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan Cernocky. 2011. Strategies for training large scale neural network language models. In *ASRU*.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword fourth edition. Technical report, Linguistic Data Consortium, Philadelphia.

Ted Pedersen. 2012. Duluth: Measuring degrees of relational similarity with the gloss vector measure of semantic relatedness. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval*

2012), pages 497–501, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *WWW '11*, pages 337–346.

J. Reisinger and R. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *NAACL '10*.

Bryan Rink and Sanda Harabagiu. 2012. UTD: Determining relational similarity using lexical patterns. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 413–418, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

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.

Mireya Tovar, J. Alejandro Reyes, Azucena Montes, Darnes Vilariño, David Pinto, and Saul León. 2012. BUAP: A first approximation to relational similarity measuring. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 502–505, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of Association for Computational Linguistics (ACL 2010)*.

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

P. D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Peter Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *International Conference on Computational Linguistics (COLING)*.

Peter D. Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research (JAIR)*, 44:533–585.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of COLING-94*, pages 782–788.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492, May.

Wen-tau Yih and Vahed Qazvinian. 2012. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of NAACL-HLT*, pages 616–620, Montréal, Canada, June.

Wen-tau Yih, Geoffrey Zweig, and John Platt. 2012. Polarity inducing latent semantic analysis. In *Proceedings of NAACL-HLT*, pages 1212–1222, Jeju Island, Korea, July.