Jurgens, D. A., Mohammad, S. M., Turney, P. D., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)* (pp. 356-364). Montreal, Canada: Association for Computational Linguistics.

SemEval-2012 Task 2: Measuring Degrees of Relational Similarity

David A. Jurgens
Department of Computer Science
University of California, Los Angeles
jurgens@cs.ucla.edu

Saif M. Mohammad
Emerging Technologies
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Peter D. Turney
Emerging Technologies
National Research Council Canada
peter.turney@nrc-cnrc.gc.ca

Keith J. Holyoak
Department of Psychology
University of California, Los Angeles
holyoak@lifesci.ucla.edu

Abstract

Up to now, work on semantic relations has focused on relation classification: recognizing whether a given instance (a word pair such as virus:flu) belongs to a specific relation class (such as CAUSE:EFFECT). However, instances of a single relation class may still have significant variability in how characteristic they are of that class. We present a new SemEval task based on identifying the degree of prototypicality for instances within a given class. As a part of the task, we have assembled the first dataset of graded relational similarity ratings across 79 relation categories. Three teams submitted six systems, which were evaluated using two methods.

1 Introduction

Relational similarity measures the degree of correspondence between two relations, where instance pairs that have high relational similarity are said to be analogous, i.e., to express the same relation (Turney, 2006). However, a class of analogous relations may still have significant variability in the degree of relational similarity of its members. Consider the four word pairs dog:bark, cat:meow, floor:squeak, and car:honk. We could say that these four X:Y pairs are all instances of the semantic relation ENTITY:SOUND; that is, X is an entity that characteristically makes the sound Y. Within a class of analogous pairs, certain pairs are more characteristic of the relation. For example, many would agree that dog:bark and cat:meow are better prototypes of the ENTITY:SOUND relation than floor:squeak. Our task requires automatic systems to quantify the degree of prototypicality of a target pair by measuring the relational similarity between it and pairs that are given as defining examples of a particular relation.

So far, most work in semantic relations has focused on differences *between* relation categories for classifying new relation instances. Past SemEval tasks that use relations have focused largely on discrete classification (Girju et al., 2007; Hendrickx et al., 2010) and paraphrasing the relations connecting noun compounds with a verb (Butnariu et al., 2010), which is also a form of discrete classification due to the lack of continuous degrees. However, there is some loss of information in any discrete classification of semantic relations. Furthermore, while some discrete classifiers provide a degree of confidence or probability for a relation classification, there is no *a priori* reason that such values would correspond to human prototypicality judgments. Our proposed task is distinct from these past tasks in that we focus on measuring the *degree* of relational similarity.[1] A graded measure of the degree of relational similarity would tell us that dog:bark is more similar to cat:meow than to floor:squeak. The discrete classification ENTITY:SOUND drops this information.

Systems that are successful at identifying degrees of relation similarity can have a significant impact where an application must choose between multiple instances of the same relation. We illustrate this with two examples. First, consider a relational search task (Cafarella et al., 2006). A user of a relational search engine might give the query,

[1] Task details and data are available at https://sites.google.com/site/semeval2012task2/ .

Subcategory	Relation name	Relation schema	Paradigms	Responses
8(e)	AGENT:GOAL	"Y is the goal of X"	pilgrim:shrine assassin:death climber:peak	patient:health runner:finish astronaut:space
5(e)	OBJECT:TYPICAL ACTION	"an X will typically Y"	glass:break soldier:fight juggernaut:crush	ice:melt lion:roar knife:stab
4(h)	DEFECTIVE	"an X is is a defect in Y"	fallacy:logic astigmatism:sight limp:walk	pimple:skin ignorance:learning tumor:body

Table 1: Examples of the three manually selected paradigms and the corresponding pairs generated by Turkers.

"List all things that are part of a car." SemEval-2007 Task 4 proposed that a relational search engine would use semantic relation classification to answer queries like this one. For this query, a classifier that was trained with the relation PART:WHOLE would be used. However, a system for measuring degrees of relational similarity would be better suited to relational search than a discrete classifier, because the relational search engine could then rank the output list in order of applicability. For the same query, the search engine could rank each item X in descending order of the degree of relational similarity between X:car and a training set of prototypical examples of the relation PART:WHOLE. This would be analogous to how standard search engines rank documents or web pages in descending order of relevance to the user's query.

As a second example, consider the role of relational similarity in analogical transfer. When faced with a new situation, we look for an analogous situation in our past experience, and we use analogical inference to transfer information from the past experience (the source domain) to the new situation (the target domain) (Gentner, 1983; Holyoak, 2012). Analogy is based on relational similarity (Gentner, 1983; Turney, 2008). The degree of relational similarity in an analogy is indicative of the likelihood that transferred knowledge will be applicable in the target domain. For example, past experience tells us that a dog barks to send a signal to other creatures. If we transfer this knowledge to a new experience with a cat meowing, we can predict that the cat is sending a signal, and we can act appropriately with that prediction. If we transfer this knowledge to a new experience with a floor squeaking, we might predict that the floor is sending a signal, which might lead us to act inappropriately. If we have a choice among several source analogies, usually the source pair with the highest degree of relational similarity to the target pair will prove to be the most useful analogy in the target domain, providing practical benefits beyond discrete relational classification.

2 Task Description

Here, we describe our task and the two-level hierarchy of semantic relation classes used for the task.

2.1 Objective

Our task is to rate word pairs by the degree to which they are prototypical members of a given relation class. The relation class is specified by a few paradigmatic (highly prototypical) examples of word pairs that belong to the class and also by a schematic representation of the relation class. The task requires comparing a word pair to the paradigmatic examples and/or the schematic representation. For example, suppose the relation class is REVERSE. We may specify this class by the paradigmatic examples attack:defend, buy:sell, love:hate, and the schematic representation "X is the reverse act of Y" or "X may be undone by Y." Given a pair such as repair:break, we compare this pair to the paradigmatic examples and/or the schematic representation, in order to estimate its degree of prototypicality. The challenges are (1) to infer the relation from the paradigmatic examples and identify what relational or featural attributes best characterize that relation, and (2) to identify the relation of the given pair and rate how similar it is to that shared by the paradigmatic examples.

2.2 Relation Categories

Researchers in psychology and linguistics have considered many different categorizations of semantic relations. The particular relation categorization is often driven by both the type of data and the intended application. Nastase and Szpakowicz (2003) propose a two-level hierarchy for noun-modifier relations, which has been widely used (Nakov and Hearst, 2008; Nastase et al., 2006; Turney and Littman, 2005; Turney, 2005). Others have used classifications based on the requirements for a specific task, such as Information Extraction (Pantel and Pennacchiotti, 2006) or biomedical applications (Stephens et al., 2001).

We adopt the relation classification scheme of Bejar et al. (1991), which includes ten high-level categories (e.g., CAUSE-PURPOSE and SPACE-TIME). Each category has between five and ten more refined subcategories (e.g., CAUSE-PURPOSE includes CAUSE:EFFECT and ACTION:GOAL), for a total of 79 distinct subcategories. Although these categories do not reflect all possible semantic relations, they greatly expand the coverage of relation types from those used in past relation-based SemEval tasks (Girju et al., 2007; Hendrickx et al., 2010), which used only seven and nine relation types, respectively. Furthermore, the classification includes many of the fundamental relations, e.g., TAXONOMIC and PART:WHOLE, while also including relations between a variety of parts of speech and less common relations, such as REFERENCE (e.g., SIGN:SIGNIFICANT) and NONATTRIBUTE (e.g., AGENT:ATYPICAL ACTION). Using such a large relation class inventory enables evaluating the generality of an approach, while still measuring performance on commonly used relations.

3 Task Data

We constructed a new data set for the task, in which word pairs are manually classified into relation categories. Word pairs within a category are manually distinguished according to how well they represent the category; that is, the degree to which they are relationally similar to paradigmatic members of the given semantic relation class. Paradigmatic members of a class were taken from examples provided by Bejar et al. (1991). Due to the large number of annotations needed, we used Amazon Mechanical Turk (MTurk),[2] which is a popular choice in computational linguistics for gathering large numbers of human responses to linguistic questions (Snow et al., 2008; Mohammad and Turney, 2010). We refer to the MTurk workers as Turkers.

Question 1: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. What relation best describes these X:Y word pairs?

(1) "X worships/reveres Y"
(2) "X seeks/desires/aims for Y"
(3) "X harms/destroys Y"
(4) "X uses/exploits/employs Y"

Question 2: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. These X:Y pairs share a relation, "X R Y". Give four additional word pairs that illustrate the same relation, in the same order (X on the left, Y on the right). Please do not use phrases composed of two or more words in your examples (e.g., "racing car"). Please do not use names of people, places, or things in your examples (e.g., "Europe", "Kleenex").

(1) _____ : _____
(2) _____ : _____
(3) _____ : _____
(4) _____ : _____

Figure 1: An example of the two questions for Phase 1.

The data set was built in two phases. In the first phase, Turkers were given three paradigmatic examples of a subcategory and asked to create new pairs that instantiate the same relation as the paradigms. In the second phase, people were asked to distinguish the new pairs from the first phase according to the degree to which they are good representatives of the given subcategory.

Phase 1 In the first phase, we built upon the paradigmatic examples of Bejar et al. (1991), who provided one to ten examples for each subcategory. From these examples, we manually selected three instances to use as seeds for generating new examples, adding examples when a subcategory had less than three. The examples were selected to be balanced across topic domains so as not to bias the Turkers. For each subcategory, we manually created a schematic representation of the relation for the examples. Table 1 gives three examples.

[2] https://www.mturk.com/

To gather new examples of each subcategory, a two-part questionnaire was presented to Turkers (see Figure 1). In the first part, Turkers were shown the three paradigm word pairs for a subcategory along with a list of four relation descriptions (schematic representations of possible relations). One of the four schematic representations accurately described the three paradigm pairs and the other three schematics were distractors (confounding descriptions). Turkers were asked to select which of the four schematic representations best matched the paradigms. The first part of the questionnaire serves as quality control by ensuring that the Turker is capable of recognizing the relation. An incorrect answer to the question is used to recognize and eliminate confused or negligent responses, which were approximately 7% of the responses.

In the second part of the Phase 1 questionnaire, Turkers were shown the three prototypes again and asked to generate four word pairs that expressed the same relation. Turkers were directed to be mindful of the order of the words in each pair, as reversed orderings can have very different degrees of prototypicality in the case of directional relations.

The Turkers provided a total of 3160 additional examples for the 79 subcategories, 2905 of which were unique. We applied minor manual correction to remove spelling errors, which reduced the total number of examples to 2823. A median of 38 examples were found per subcategory with a maximum of 40 and minimum of 23. We note that Phase 1 gathers both high and low quality examples of the relation, which were all included to capture different degrees of prototypicality.

We included an additional 395 pairs by randomly sampling five instances of each subcategory and creating a new pair from the reversed arguments, i.e., adding pair Y:X to the subcategory containing X:Y. Adding reversals was inspired by an observation during Phase 1 that reversed pairs would occasionally be added by the Turkers themselves. We were curious to see what impact reversals would have on Turker responses and on the output of automatic systems. Reversals should reveal order sensitivity with a strongly directional relation, such as PART:WHOLE, but also perhaps there is order sensitivity with more symmetric relations, such as SYNONYMY. Phase 1 produced a total of 3218 pairs.

Question 1: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. What relation best describes these X:Y word pairs?

(1) "X worships/reveres Y"
(2) "X seeks/desires/aims for Y"
(3) "X harms/destroys Y"
(4) "X uses/exploits/employs Y"

Question 2: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. These X:Y pairs share a relation, "X R Y". Now consider the following word pairs:

(1) pig:mud
(2) politician:votes
(3) dog:bone
(4) bird:worm

Which of the above numbered word pairs is the MOST illustrative example of the same relation "X R Y"?

Which of the above numbered word pairs is the LEAST illustrative example of the same relation "X R Y"?

Note: In some cases, a word pair might be in reverse order. For example, tree:forest is in reverse order for the relation "X is made from a collection of Y". The correct order would be forest:tree; a forest is made from a collection of trees. You should treat reversed pairs as BAD examples of the given relation.

Figure 2: An example of the two questions for Phase 2.

Phase 2 In the second phase, the response pairs from Phase 1 were ranked according to their prototypicality. We opted to create a ranking using MaxDiff questions (Louviere, 1991). MaxDiff is a choice procedure consisting of a question about a target concept and four or five alternatives. A participant must choose both the best and worse answers from the given alternatives.

MaxDiff is a strong alternative to creating a ranking from standard rating scales, such as the Likert scale, because it avoids scale biases. Furthermore MaxDiff is more efficient than other choice procedures such as pairwise comparison, because it does not require comparing all pairs.

Like Phase 1, Phase 2 was performed using a two-part questionnaire. The first question was identical to that of Phase 1: four examples of the same relation subcategory generated in Phase 1 were presented and the Turker was asked to select the correct relation from a list of four options. This first question served as a quality control measure for ensuring the Turker could properly identify the relation in question and it also served as a hint, guiding

the Turker toward the intended understanding of the shared relation underlying the three paradigms. In the second part, the Turker selects the most and least illustrative example of that relation from among the four examples of pairs generated by Turkers in Phase 1.

We aimed for five Turker responses for each MaxDiff question but averaged 4.73 responses for each MaxDiff question in a subcategory, with a minimum of 3.45 responses per MaxDiff question. Turkers answered a total of 48,846 questions over a period of five months, of which 6,536 (13%) were rejected due to a missing answer or an incorrect response to the first question.

3.1 Measuring Prototypicality

The MaxDiff responses were converted into the prototypicality scores using a counting procedure (Orme, 2009). For each word pair, the prototypicality is scored as the percentage of times it is chosen as most illustrative minus the percentage of times it is chosen as least illustrative (see Figure 2). While methods such as hierarchical Bayes models can be used to compute a numerical rank from the responses, we found the counting method to produce very reasonable results.

3.2 Data Sets

The 79 subcategories were divided into training and testing segments. Ten subcategories were provided as training with both the Turkers' MaxDiff responses and the computed prototypicality ratings. The ten training subcategories were randomly selected. The remaining 69 subcategories were used for testing. All data sets are now released on the task website under the Creative Commons 3.0 license.[3]

Participants were given the list of all pairs gathered in Phase 1 and the Phase 2 responses for the 10 training subcategories. Phase 2 responses for the 69 test categories were not made available. Participants also had access to the set of questionnaire materials provided to the Turkers, the full list of paradigmatic examples provided by Bejar et al. (1991), and the confounding schema relations from the initial questions in Phase 1 and Phase 2, which might serve as negative training examples.

[3] http://creativecommons.org/licenses/by/3.0/

4 Evaluation

Systems are given examples of pairs from a single category and asked to provide numeric ratings of the degree of relational similarity for each pair relative to the relation expressed in that category.

4.1 Scoring

Spearman's rank correlation coefficient, ρ, and a MaxDiff score were used to evaluate the systems. For Spearman's ρ, the prototypicality rating of each pair is used to build a ranking of all pairs in a subcategory. Spearman's ρ is then computed between the pair rankings of a system and the gold standard ranking. This evaluation abstracts away from comparing the numeric values so that only their relative ordering in prototypicality is measured.

In the second scoring procedure, we measure the accuracy of a system at answering the same set of MaxDiff questions as answered by the Turkers in Phase 2 (see Figure 2). Given the four word pairs, the system selects the pair with the lowest numerical rating as *least illustrative* and the pair with the highest numerical rating as *most illustrative*. Ties in prototypicality are broken arbitrarily. Accuracy is measured as the percentage of questions answered correctly. An answer is considered correct when it agrees with the majority of the Turkers. In some cases, two answers may be considered correct. For example, when five Turkers answer a given MaxDiff question, two Turkers might choose one pair as the most illustrative and two other Turkers might choose another pair as the most illustrative. In this case, both pairs would count as correct choices for the most illustrative pair.

4.2 Baselines

We consider two baselines for evaluation: Random and PMI. The Random baseline rates each pair in a subcategory randomly. The expected Spearman correlation for Random ratings is zero. The expected MaxDiff score for Random ratings would be 25% (because there are four word pairs to choose from in Phase 2) if there were always a unique majority, but it is actually about 31%, due to cases where two pairs both get two votes from the Turkers.

Given a MaxDiff question, a Turker might select the pair whose words are most strongly associated

Team	Members	System	Description
Benemérita Universidad Autónoma de Puebla (México) (BUAP)	Mireya T. Vidal, Darnes V. Ayala, Jose A.R. Ortiz, Azucena M. Rendon, David Pinto, and Saul L. Silverio	BUAP	Each pair is represented as a vector over multiple features: lexical, intervening words, WordNet relations between the pair, and syntactic features such as part of speech and morphology. Prototypicality is based on cosine similarity with the class's pairs.
University of Texas at Dallas (UTD)	Bryan Rink and Sanda Harabagiu	NB	Unsupervised learning identifies intervening patterns between all word pairs. Each pattern is then ranked according to its subcategory specificity by learning a generative model from patterns to word pairs. Prototypicality ratings are based on confidence that the highest scoring pattern found for a pair belongs to the subcategory.
		SVM	Intervening patterns are found using the same method as UTD-NB. Word pairs are then represented as feature vectors of matching patterns. An SVM classifier is trained using a subcategory's pairs as positive training data and all other pairs as negative. Prototypicality ratings are based on SVM confidence of class inclusion.
University of Minnesota, Duluth (Duluth)	Ted Pedersen	V0	WordNet is used to build the set of concepts connected by WordNet relations to the pairs' words. Prototypicality is estimated using the vector similarity of the concatenated glosses.
		V1	Same procedure as V0, with one further expansion to related concepts.
		V2	Same procedure as V0, with two further expansions to related concepts.

Table 2: Descriptions of the participating teams and systems.

as the most illustrative and the least associated as the least illustrative. Therefore, we propose a second baseline where pairs are rated according to their Pointwise Mutual Information (PMI) (Church and Hanks, 1990), which measures the statistical association between two words. For this baseline, the prototypicality rating given to a word pair is simply the PMI score for the pair. For two terms x and y, PMI(x, y) is defined as $\log_2\left(\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)}\right)$ where $\mathrm{p}(\cdot)$ denotes the probability of a term or pair of terms. The PMI score was calculated using the method of Turney (2001) on a corpus of approximately 50 billion tokens, indexed by the Wumpus search engine.[4] To calculate $\mathrm{p}(x, y)$, we recorded all co-occurrences of both terms within a ten-word window.

5 Systems

Three teams submitted six systems for evaluation. Table 2 summarizes the teams and systems. Two teams (BUAP and UTD) based their approaches on discovering relation-specific patterns for each category, while the third team (Duluth) used vector space comparisons of the glosses related to the pairs.

[4]http://www.wumpus-search.org/

No single system was able to achieve superior performance on all subcategories. Table 3 reports the averages across all subcategories for Spearman's ρ and MaxDiff accuracy. Five systems were able to perform above the Random baseline, while only one system, UTD-NB, consistently performed above the PMI baseline.

However, the average performance masks superior performance on individual subcategories. Table 3 also reports the number of subcategories in which a system obtained a statistically significant Spearman's ρ with the gold standard ranking. Despite the low average performance, most models were able to obtain significant correlation in multiple subcategories. Furthermore, the significant correlations for different systems were not always obtained in the same subcategories. Across all subcategories, 43 had a significant correlation at $p < 0.05$ and 27 at $p < 0.01$. The broad coverage of significantly correlated subcategories spanned by the combination of all systems and the PMI baseline suggests that high performance on this task may be possible, but that adapting to each of the specific relation types may be very beneficial.

Team	System	Spearman's ρ	# of Subcategories		MaxDiff
			$p < 0.05$	$p < 0.01$	
BUAP	BUAP	0.014	2	0	31.7
UTD	NB	**0.229**	22	16	**39.4**
	SVM	0.116	11	5	34.7
Duluth	V0	0.050	9	3	32.4
	V1	0.039	10	4	31.5
	V2	0.038	7	3	31.1
Baselines	Random	0.018	4	0	31.2
	PMI	0.112	15	7	33.9

Table 3: Average Spearman's ρ and MaxDiff scores for all system across all 69 test subcategories. Columns 4 and 5 denote the number of subcategories with a Spearman's ρ that is statistically significant at the noted level of confidence.

Relation Class	Random	PMI	BUAP	UTD-NB	UTD-SVM	Duluth-V0	Duluth-V1	Duluth-V2
Class-Inclusion	0.057	0.221	0.064	**0.233**	0.093	0.045	0.178	0.168
Part-Whole	0.012	0.144	0.066	**0.252**	0.142	-0.061	-0.084	-0.054
Similar	0.026	0.094	-0.036	**0.214**	0.131	0.183	0.208	0.198
Contrast	-0.049	0.032	0.000	**0.206**	0.162	0.142	0.120	0.051
Attribute	0.037	-0.032	-0.095	**0.158**	0.052	0.044	-0.003	0.008
Non-Attribute	-0.070	**0.191**	0.009	0.098	0.094	0.079	0.066	0.074
Case Relations	0.090	0.168	-0.037	**0.241**	0.187	-0.011	-0.068	-0.115
Cause-Purpose	-0.011	0.130	0.114	**0.183**	0.060	0.021	0.022	0.042
Space-Time	0.013	0.084	0.035	**0.375**	0.139	0.055	-0.004	0.040
Reference	0.142	0.125	-0.001	**0.346**	0.082	0.028	0.074	0.067

Table 4: Average Spearman's ρ correlation with the Turker rankings in each of the high-level relation categories, with the highest average correlation for each subcategory shown in bold.

6 Discussion

Sensitivity to Pair Association The PMI baseline performed much better than anticipated, outperforming all systems but UTD-NB on many of the subcategories, despite treating all relations as directionless. Performance was highest in subcategories where the X:Y pair might reasonably be expected to occur together, e.g., FUNCTIONAL or CONTRADICTORY. However, PMI benefits from the design of our task, which focuses on rating pairs within a given subcategory. In a different task that mixed pairs from a variety of subcategories, PMI would perform poorly, because it would assign high scores to pairs of strongly associated words, regardless of whether they belong to a given subcategory.

Difficulty of Specific Subcategories Performance across the high-level categories was highly varied between approaches. The category-level summary shown in Table 4 reveals high-level trends in difficulty across all submitted systems. The submitted systems performed best for subcategories under the Similar category, while the systems performed worst for Non-Attribute subcategories.

As a further possibility of explaining performance differences between subcategories, we considered the hypothesis that the difficulty of a subcategory is inversely proportional to the range of prototypicality scores, i.e., subcategories with restricted ranges are more difficult. However, we found that the difficulty was uncorrelated with both the size of the interval spanned by prototypicality scores and the standard deviation of the scores.

Sensitivity to Argument Reversal The directionality of a relation can significantly impact the rated prototypicality of a pair whose arguments have been reversed. As an approximate measure of the effect on prototypicality when a pairs' arguments are reversed, we calculated the expected drop in rank

Team	System	Spearman's ρ No Reversals	Spearman's ρ With Reversals
BUAP	BUAP	-0.003	0.014
UTD	NB	0.190	0.229
	SVM	0.104	0.116
Duluth	V0	0.062	0.050
	V1	0.040	0.039
	V2	0.046	0.038
Baselines	Random	0.004	0.018
	PMI	0.143	0.112

Table 5: Average pair ranking correlation for all subcategories when reversed pairs are included and excluded.

between a pair and its reversed form. Based on the Turker rankings, the SEQUENCE (e.g., pregnancy:birth) and FUNCTIONAL (e.g., weapon:knife) subcategories exhibited the strongest sensitivity to argument reversal, while ATTRIBUTE SIMILARITY (e.g., rake:fork) and CONTRARY (e.g., happy:sad) exhibited the least.

The inclusion of reversed pairs potentially adds a small amount of noise to the relation identification process for subcategories with directional relations. Two teams, BUAP and UTD, accounted for relation directionality, while Duluth did not, which resulted in the Duluth systems ranking reversed pairs the same. Therefore, we conducted a post-hoc analysis of the impact of reversals by removing the reversed pairs from the computed prototypicality rankings. Table 5 reports the resulting Spearman's ρ. With Spearman's ρ, we can easily evaluate the impact of the reversals, because we can delete a reversed pair without affecting anything else. For the MaxDiff questions, if there is one reversal in a group of four choices, then we need to delete the whole MaxDiff question. Therefore we do not include the MaxDiff score in Table 5.

Removing reversals decreased performance in the three systems that were sensitive to pair ordering (BUAP, UTD-NB, and UTD-SVM), while only marginally increasing performance in the three systems that ignored the ordering. The performance decrease in systems that use ordering suggests that the reversed pairs are easily identified and ranked appropriately low. As a further estimate of the models' ability to correctly order reversals, we compared the difference in a reversal's rank for both a system's ranking and the ranking computed from Turker Responses. Table 6 reports the Root Mean Squared Error (RMSE) in ranking difference for the three systems that took argument order into account. Although not the best performing system, Table 6 indicates that the UTD-SVM system was most able to appropriately weight reversals' prototypicality. In contrast, the UTD-NB system often had many pairs tied for the lowest rank, which either resulted in pair and its reversal being tied or having a much smaller rank difference, thereby increasing its RMSE.

Team	System	RMSE
BUAP	BUAP	256.07
UT Dallas	NB	257.15
	SVM	209.95
Baseline	Random	227.25

Table 6: RMSE in estimating the difference in rank between a pair and its reversal in the gold standard.

7 Conclusions

We have introduced a new task focused on rating the degrees of prototypicality for word pairs sharing the same relation. Participants first identify the relation shared between example pairs and then rate the degree to which each pair expresses that relation. As a part of the task, we constructed a dataset of prototypicality ratings for 3218 word pairs in 79 different relation categories.

Participating systems used combinations of corpus-based, syntactic, and WordNet features, with varying degrees of success. The task also included a competitive baseline, PMI, which surpassed all but one system. Several models obtained moderate performance in select relation subcategories, but no one approach succeeded in general, which introduces much opportunity for future improvement. We also hope that both the example pairs and their prototypicality ratings will be a valuable data set for future research in Linguistics as well as Cognitive Psychology. All data sets for this task have been made publicly available on the task website.

Acknowledgements

This research was supported by ONR grant N000140810186.

References

Isaac I. Bejar, Roger Chaffin, and Susan E. Embretson. 1991. *Cognitive and Psychometric Analysis of Analogical Problem Solving*. Springer-Verlag.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Dairmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 39–44. Association for Computational Linguistics.

Michael J. Cafarella, Michele Banko, and Oren Etzioni. 2006. Relational web search. In *WWW Conference*.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluation (SemEval-2007)*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 33–38. Association for Computational Linguistics.

Keith J. Holyoak. 2012. Analogy and relational reasoning. In *Oxford handbook of thinking and reasoning*, pages 234–259. Oxford University Press.

Jordan J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics.

Preslav Nakov and Marti A. Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL*, volume 8.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 285–301. ACL Press 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 AAAI*, volume 21, page 781.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and hb.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, J. Mostafa, et al. 2001. Detecting gene relations from medline abstracts. In *Pacific Symposium on Biocomputing*, volume 6, pages 483–495. Citeseer.

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. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502.

Peter D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of IJCAI*, pages 1136–1141.

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

Peter D. Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33(1):615–655.