Measuring Degrees of Relational Similarity: SemEval-3 Task Proposal

David A. Jurgens[1], Saif M. Mohammad[2], Peter D. Turney[3], and Keith J. Holyoak[4]

April 26, 2011

1. Department of Computer Science, University of California, Los Angeles, jurgens@cs.ucla.edu
2. Institute for Information Technology, National Research Council of Canada, saif.mohammad@nrc-cnrc.gc.ca
3. Institute for Information Technology, National Research Council of Canada, peter.turney@nrc-cnrc.gc.ca
4. Department of Psychology, University of California, Los Angeles, holyoak@lifesci.ucla.edu

1. Introduction

Consider the three word pairs dog:bark, cat:meow, and car:vroom. We could say that these three X:Y pairs are all instances of the semantic relation ENTITY:SOUND; that is, X is an entity that characteristically makes the sound Y. However, there is some loss of information in any discrete classification of semantic relations. A graded measure of the degree of relational similarity between word pairs would tell us that dog:bark is more similar to cat:meow than to car:vroom. The discrete classification ENTITY:SOUND drops this information.

Semantic relations have been the topic of three past SemEval tasks:

1. SemEval-1, Task 4: Classification of Semantic Relations between Nominals [1]
2. SemEval-2, Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals [2]
3. SemEval-2, Task 9: Noun Compound interpretation Using Paraphrasing Verbs [3]

Our proposed task is distinct from these past tasks in that we focus on measuring the degree of relational similarity. SemEval-1, Task 4 and SemEval-2, Task 8 both involved discrete classification of semantic relations, with no notion of degrees of similarity. SemEval-2, Task 9 involved labeling relations with verbs; for example, the three word pairs above could all be labeled with *makes*; X *makes* the sound Y. This is also a form of discrete classification, lacking continuous degrees.

When we are 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 & Thagard, 1995). 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 car vrooming, we might predict that the car is sending a signal, which might lead us to act inappropriately. If we have a choice among several source analogies, usually the source with the highest degree of relational similarity to the target will prove to be the most useful analogy in the target domain. Therefore systems that go beyond discrete relational classification to graded relational similarity will have the practical benefit that we can estimate the likelihood that their output will be applicable and appropriate.

For example, SemEval-1, Task 4 proposed *relational search* as a motivating application for semantic relation classification [1]. Cafarella et al. (2006) describe four types of relational search tasks. For example, a user of a relational search engine might give the query, “List all things that are part of a car.” SemEval-1, 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 query, “List all things that are part of a car,” 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.

Similarly, a user could provide the query, “What illnesses have been cured by drugs?,” where the system must respond with examples of the DRUG:ILLNESS relation, such as aspirin:fever, calamine:poison ivy, food:starvation, coffee:drowsiness, and penicillin:staphylococcus. It would be more useful for the relational search engine to rank the X:Y pairs in descending order of the degree of prototypicality for the DRUG:ILLNESS relation, rather than randomly listing examples, as a system based purely on discrete classification would do.

2. Description of the Task

We propose to create a dataset in which word pairs are manually classified into various categories, as in past SemEval tasks, but furthermore 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 prototypical members of the given semantic relation class. This dataset will be used in two tasks.

In the first task, systems will be given examples of pairs sampled from a given category, and the systems will have to answer questions that test their ability to match human performance on distinguishing pairs according to how well they represent the given category. For this task, the systems will have access to prototypical members of the given category and they will know the category from which the examples are sampled.

The second task will be a traditional supervised classification problem, using the same dataset and the same categories. The dataset will be divided into training and testing sets. The category labels will be visible in the training set but hidden in the testing set. The systems must guess the categories of the pairs in the testing set and indicate the degree to which each pair is a prototypical member of the guessed category. Systems will receive two scores, one for accurate classification (assigning pairs to the correct categories) and another for accurate prototypicality grades (estimating the degree to which each pair is a prototypical member of the category). For the second task, the systems will have access to prototypical members of every category, but they will not know the categories of the testing set pairs.

Researchers in psychology and linguistics have considered many different categorizations of semantic relations. We will adopt the relation classification scheme of Bejar et al. (1991), which includes 10 high-level categories (e.g., Cause-Purpose and Space-Time). Each category has between 5 and 10 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 include many fundamental relations. Bejar et al. (1991) provide one to ten example pairs for each of their subcategories. We will use these examples as the prototypes of each subcategory.

3. Data Creation

We will build on the categories and prototypes of Bejar et al. (1991) in two phases. In the first phase, people will be given the prototypical examples of a subcategory and they will be asked to create new pairs that instantiate the same relation as the prototypes. In the second phase, people will be asked to distinguish the new pairs from the first phase according to the degree to which they are good representatives of the given subcategory.

For example, consider Bejar et al.'s (1991) subcategory 8(c):

> 8. Cause-Purpose
> 8 (c). Enabling Agent:Object — *match:candle, gasoline:car, mnemonic:memory*

In the first phase, people would see instructions like the following:

> Consider these three examples of pairs of words, where each pair is related in the same way: *match:candle, gasoline: car, mnemonic:memory*. That is, a match is used to light a candle, gasoline is used to drive a car, and a mnemonic is used to remember something that is difficult to remember. Please give three new examples of word pairs that illustrate this same relation.
> (1) ___________ : ___________
> (2) ___________ : ___________
> (3) ___________ : ___________

In the second phase, people would see MaxDiff questions (Louviere, 1991) concerning the new pairs from the first phase. MaxDiff is a choice procedure in which participants are given a target concept and then asked to select the best and worst items from a list of options [4]. A question for subcategory 8(c) might look like the following:

> Consider the relation between the words in the pair *match:candle*.
> Which of the following four pairs is the BEST example of the same relation? _____
> Which of the four pairs is the WORST example of the the same relation? _____
> (1) *key:door*
> (2) *hammer:nail*
> (3) *battery:flashlight*
> (4) *experiment:science*

MaxDiff [4] is a strong alternative to standard rating scales such as the Likert scale [5]. Like pairwise comparison [6], MaxDiff avoids problems with scale biases, yet it is more efficient for data gathering than pairwise comparison.

Training and testing data will be acquired from human subjects using the Amazon Mechanical Turk (MTurk) system [7]. MTurk is becoming a popular choice in computational linguistics for gathering large numbers of human responses to linguistic questions (Snow et al., 2008; Mohammad & Turney, 2010).

4. Evaluation Methodology and Criteria

We described two tasks in Section 2, (1) distinguishing pairs from known subcategories according to their degree of prototypicality for the given subcategory and (2) classifying pairs with hidden category labels and indicating the degree to which each pair is a prototypical member of the guessed category.

For the first task, systems will have access to Bejar et al.'s (1991) categories and the prototypical examples of pairs in each category. Systems will be given exactly the same MaxDiff questions as the people are given, as described in Section 3. Thus the systems will know the subcategory from which the sample pairs were taken. For each MaxDiff question, a system will get half a point if its answer to the BEST question matches the human answer and another half a point if its answer to the WORST question matches the human answer.

We will also apply a hierarchical Bayes model [8] to the MaxDiff questions and answers, to infer numerical ratings of the degree of prototypicality of each pair with respect to the prototype(s) of the pair's subcategory. (The hierarchical Bayes model is the recommended method for inferring numerical ratings from MaxDiff answers [4].) This will permit a second score: each system will be scored by the correlation between its numerical prototypicality ratings of each pair and the ratings produced by the hierarchical Bayes model. It will be interesting to compare these two different scoring methods, accuracy on MaxDiff questions versus correlation with (inferred) human prototypicality ratings.

The human MaxDiff questions and answers will be divided into a small development set and a larger evaluation set. The development set, with both questions and answers, will be released to the SemEval participants first. Later the evaluation set, with only questions, will be released. At some time after participants have submitted their systems' answers for the evaluation questions, system scores and evaluation set answers will be released. (For scoring the systems, note that the hierarchical Bayes model will have access to the human MaxDiff questions and answers for both the development and evaluation sets, whereas the participating systems will not have access to the human MaxDiff answers for the evaluation set.)

For the second task, the data will be divided into training and testing sets. SemEval participants will apply supervised learning algorithms to the training set and generate guesses for the subcategory and prototypicality ratings for the pairs in the testing set. The systems will be scored by their accuracy on guessing the subcategory labels and also by the correlation of their prototypicality ratings with prototypicality numbers from the hierarchical Bayes model. At some time after participants have submitted their systems' answers for the testing set, system scores and testing set categories and ratings will be released.

Bejar et al. (1991) have 10 high-level categories and 79 subcategories. There will be two variations of the second task, one with the 10 high-level categories and one with the 79 subcategories. Participating teams will have the option to do any or all of these sub-tasks, (1) the first task, (2) the second task with 10 categories, and/or (3) the second task with 79 subcategories.

5. Discussion

As with all tasks that ask the annotators to judge degrees, there are possible concerns about inter-annotator agreement (IAA) and the meaningfulness of the annotation results. For instance, with the dog:bark, cat:meow, car:vroom example, it might not be clear why car:vroom is different from the other two pairs without an explanation (e.g., dogs bark and cats meow to send a signal, but the vroom of a car is typically not a signal). In some cases, the degree of similarity may be a function of the problem to be solved (i.e., the problem that triggered the analogical inference), thus the degree of similarity may not be easy to define without regard to the application context.

We expect that the MaxDiff methodology will prove to give high IAA, but the true test of this hypothesis will be the actual IAA results that we obtain when we create the data. We plan to measure IAA in a variety of ways (e.g., Spearman's rank correlation coefficient [9], Pearson's product-moment correlation coefficient [10], and Kendall's coefficient of concordance [11]) and report the outcome of these measurements.

6. Copyright

Our datasets, including all annotations, will be released under a Creative Commons License [12].

7. Resources Required

The data will be created using the MTurk. The process will have the following stages:

1. Carefully construct the questionnaire for the MTurk.
2. Run a pilot task on MTurk. Make sure all is in order. Possibly refine the questionnaire.
3. Upload the complete, refined questionnaire on MTurk.
4. Postprocess the MTurk results, including data cleaning and data analysis.

We expect the full process will take about two months and cost under $1000. We will pay the costs.

References

Bejar, I.I., Chaffin, R., and Embretson, S.E. (1991). *Cognitive and Psychometric Analysis of Analogical Problem Solving*. New York: Springer-Verlag.

Cafarella, M.J., Banko, M., and Etzioni, O. (2006). *Relational Web Search.* University of Washington, Department of Computer Science and Engineering, Technical Report 2006-04-02.

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

Holyoak, K., and Thagard, P. (1995). *Mental Leaps*. MIT Press.

Louviere, J.J. (1991). *Best-Worst Scaling: A Model for the Largest Difference Judgments*. Working Paper, University of Alberta.

Mohammad, S.M., and Turney, P.D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, June 2010, LA, California, 26-34.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* Association for Computational Linguistics, Honolulu, Hawaii, 254-263.

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

Notes

[1] SemEval-1, Task 4, https://docs.google.com/View?docid=d2jm3f3_98kcwd4
[2] SemEval-2, Task 8, https://docs.google.com/View?docid=dfvxd49s_36c28v9pmw
[3] SemEval-2, Task 9, https://docs.google.com/View?docid=dfvxd49s_35hkprbcpt
[4] MaxDiff, http://en.wikipedia.org/wiki/MaxDiff
[5] Likert Scale, http://en.wikipedia.org/wiki/Likert_scale
[6] Pairwise Comparison, http://en.wikipedia.org/wiki/Pairwise_comparison
[7] Amazon Mechanical Turk, https://www.mturk.com/mturk/welcome
[8] Hierarchical Bayes Model, http://en.wikipedia.org/wiki/Hierarchical_Bayes_model
[9] Spearman's Rank Correlation Coefficient, http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
[10] Pearson's Product-moment Correlation Coefficient, http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
[11] Kendall's Coefficient of Concordance, http://en.wikipedia.org/wiki/Kendall%27s_W
[12] Creative Commons, http://creativecommons.org/