Explaining Similarity of Terms

Vishnu Vyas
USC Information Sciences Institute
Marina del Rey, CA
vishnu@isi.edu

Patrick Pantel
Yahoo! Inc.
Santa Clara, CA 95054
me@patrickpantel.com

Abstract

Computing the similarity between entities is a core component of many NLP tasks such as measuring the semantic similarity of terms for generating a distributional thesaurus. In this paper, we study the problem of explaining post-hoc why a set of terms are similar. Given a set of terms, our task is to generate a small set of explanations that best characterizes the similarity of those terms. Our contributions include: 1) an information-theoretic objective function for quantifying the utility of an explanation set; 2) a survey of psycholinguistics and philosophy for evidence of different sources of explanations such as descriptive properties and prototypes; 3) computational baseline models for automatically generating various types of explanations; and 4) a qualitative evaluation of our explanation generation engine.

1 Introduction

Computing similarity is at the core of many computer science tasks. Many have developed algorithms for computing the semantic similarity of words (Lee, 1999), of expressions to generate paraphrases (Lin and Pantel, 2001) and of documents (Salton and McGill, 1983). However, little investigation has been spent on automatically explaining why a particular set of elements are similar to one another.

Explaining similarity is an important part of various natural language applications such as question answering and building lexical ontologies such as WordNet (Fellbaum, 1998). Several questions must be addressed before one can begin to explore this topic. First, what constitutes a good explanation and what are the sources of these explanations? Second, how can we automatically generate these different types of explanations? Third, how do we empirically evaluate the quality of an explanation? In this paper, we propose a first analysis of these questions.

2 Related Work

The task of generating explanations has been studied in relation to Question Answering (Hirschman and Gaizauskas, 2001) and Knowledge Representation and Reasoning (Cohen et al., 1998). Within Question Answering, explanations have mostly been viewed from a deductive framework and have focused on proof trees and inference traces as sources of explanations (Moldovan and Rus, 2001). Summarization and text generation from proof trees have also been explored as explanations in QA systems (Barker et al., 2004). Lester (1997) proposed explanation design packages, a hybrid representation for discourse knowledge that generates multi-sentential explanations.

Detailed psycholinguistic studies into how people explain things suggests that people explain similarity using "feature complexes" (Fillenbaum, 1969), a bundle of features semantically related with a term. This suggests considering explanations of similarity as the shared features among a set of terms. Another competing idea from linguistic philosophy is the Prototype theory by Rosch (1975). It is argued that objects within a semantic category are represented by another, more commonly used or much simpler member of the same semantic category, called a prototype. And, within this view, explanations for similarity are prototypes from the same semantic category as the given terms. Deese (1966) investigated similarity in terms of stimulus-response word association experiments providing empirical evidence to consider other semantically similar words as explanations.

3 An Information Theoretic Framework for Explaining Similarity

In this section, we present an information-theoretic framework that defines a *good explanation* using the intuition that they are highly informative and reduce the uncertainty in the set of query terms. For example, consider the set of query terms {*Maybach, Maserati, Renault*}. One possible explanation of their similarity which is very informative is *they are all like a Ford* (i.e., a prototype explanation). Other possible explanations include *they can be driven using a steering wheel and they have wheels* (i.e., descriptive properties as explanations). Each of these explanations reduces the uncertainty regarding the semantics of the original set of terms. In information theory, the concept of reduction in uncertainty is related to information gain, and good explanation sets can be quantified in terms of information gain.

Formally, given a set of query terms Q, and a set of explanations E, we define the best explanation set as one which provides maximum information to the set Q, or in other words,

$$E = argmax_{E' \in \phi(\Xi)} I(Q; E') \quad (1)$$

where Ξ is the set of all explanations (discussed in detail in Section 4) and $\phi(X)$ represents the power set of X. The problem of choosing the best explanation set for a given query set is now reduced to a problem of optimization under I.

3.1 The Information Function

The information function I in Eq. (1) is a set function which defines the amount of information contributed by the set of explanations E' to the set of query words Q. There are many possible information functions, but we would like all of them to have some common properties.

Consistency

The information function should be consistent. For two sets, E and E', if $E \subseteq E'$ then $I(Q; E') \geq I(Q; E)$. In other words, given two explanation sets E and E', with E' containing extra explanations, not in E, the information function should assign larger values to E' with respect to Q than it assigns to E.

Explanation Set Cardinality

Another important requirement regarding I, is the size of the explanation sets. Any consistent information function would assign larger values to larger sets of explanations. This leads to a problem where the optimal solution is always the set of all explanations. We overcome this by fixing an upper bound for the size of explanation sets that are generated by the function I.

Redundancy and Joint Information

Many explanations in an explanation set might overlap semantically and the information function has to account for such overlaps. However, information functions which take such semantic overlap into account are computationally hard to optimize. One approach to this problem is to find approximate solutions using heuristic search techniques, however, relaxing this constraint lets us use common association measures such as mutual information (Cover and Thomas, 1991) as information functions.

3.2 Marginal Formulation of the Information Function

Another equivalent formulation of Eq. (1) is to use marginal information gains. This formulation also gives a simple greedy algorithm to the optimization problem when the size of explanation set is fixed. Let us define the marginal gain in information to the set Q, when the explanation e is added to the set of explanations E as:

$$IG_{Q;E}(e) = I(Q; E \cup \{e\}) - I(Q; E)$$

Then, the best set of explanations of size k can be recursively defined as

$$E_0 = \{\}$$

$$E_n = E_{n-1} \cup \{e\}$$

such that

$$e = argmax_{e' \in \Xi} IG_{Q;E_{n-1}}(e')$$

and

$$|E_n| \leq k$$

If our marginal information gain is independent of the set of explanations to which it is added, we can rank explanations by their marginal information gains as added to the empty set. Then, choosing the top k explanations gives us the k-best explanation set for the query.

4 Sources for Similarity Explanations

In Section 3, we presented a framework for quantifying a good explanation set. In this section we present two sources of explanations, using *descriptive properties* and using *prototypes*.

4.1 Explanations from Descriptive Properties

The concept of *essence* as discussed by early empiricists was the first study of using descriptive properties to explain the similarity of a set of terms. Descriptive properties are the shared essential attributes of a set of similar terms and one way of explaining the similarity of a set of terms is to generate descriptive properties.

Within our framework in Section 3, let the query set Q be a set of similar words, and let Ξ, the set of all explanations be the set of all properties

that are shared by all the words within the query set. Using mutual information as our measure of association between properties and terms we can rewrite our information function I as:

$$I(Q;E) = \sum_{q \in Q} p(q) \sum_{e \in E} p(e \mid q) \log \frac{p(e \mid q)}{p(e)}$$

The marginal information gain for a single explanation e is:

$$IG_{Q;E}(e) = \sum_{q \in Q} p(q) \cdot p(e \mid q) \log \frac{p(e \mid q)}{p(e)}$$

Since the information gain is independent of the explanation set E, we can find the best set of size k by greedily choosing explanations until our explanation set reaches the desired size.

4.2 Explanations from Prototypes

As discussed in Section 2 given a set of query terms, people can represent their meaning using other common members from the same semantic category, called prototypes. Within the framework of Section 3, let Q be our set of query terms. To generate the set of all explanations Ξ, we use clusters in the CBC resource (Pantel and Lin, 2002) as an approximation to semantic categories and we collect all possible words that belong to that cluster which then becomes our candidate set.

Let C_q denote the cluster to which the query term q belongs to. Also let the set $C(Q)$ be the set of all clusters to which the query terms of Q belong to. Then

$$\Xi = \{w | C_w \in C(Q)\}$$

Now our information function can be written as:

$$I(Q;E) = \sum_{q \in Q} p(C_q) \sum_{e \in E} p(e \mid C_q) \log \frac{p(e \mid C_q)}{p(e)}$$

The marginal formulation of the above function is:

$$I_{Q;E}(e) = \sum_{q \in Q} p(C_q) \cdot p(e \mid C_q) \log \frac{p(e \mid C_q)}{p(e)}$$

We can find the optimal set of explanations of size k using a greedy algorithm as in Section 4.1.

5 Experimental Results

5.1 Experimental Setup

For each source of explanation discussed in Section 4, we estimated the model probabilities using corpus statistics extracted from the 1999 AP newswire collection (part of the TREC-2002 Aquaint collection).

In order to obtain a representative set of similar terms as queries to our systems, we randomly chose 100 concepts from the CBC collection (Pantel and Lin, 2002) consisting of 1628 clusters of nouns. For each of these concepts, we randomly chose a set of cluster instances (nouns), where the size of each set was randomly chosen to consist of two to five nouns.

Each of these samples forms a query. For each explanation source described in Section 4, we generated explanation sets for the random samples and in the next section we show a random selection of these system outputs.

5.2 Examples of Explanations using Descriptive Properties

For the algorithm discussed in Section 4.1, we derived our descriptive properties using the output of the dependency analysis generated by the Minipar (Lin, 1994) dependency parser. We use syntactic dependencies between words to model their semantic properties. The assumption here is that some grammatical relations, such as subject and object can yield semantic properties of terms. For example, from a phrase like "students eat many apples", we can infer the properties *can-be-eaten* for *apples* and *can-eat* for *students*. In this paper, we use a combination of corpus statistics and manual filters for grammatical relations to uncover candidate semantic properties.

Table 1: Explanations generated using descriptive properties.

Query Sets	**Explanations**
Palestinian-Israeli, India-Pakistan	talks(NN), conflict(NN), dialogue(NN), relation(NN), peace(NN).
TV, television-station	cable(NN), watch(obj), see(ON), channel(NN), local(ADJ-MOD)
Britney Spears, Janet Jackson	like(OBJ), concert(NN), video(NN), fan(NN), album(GEN)
Crisis, Uncertainty, Difficulty	face(OBJ), resolve(OBJ), overcome(OBJ), financial(ADJ-MOD), political(ADJ-MOD)

Intuitively, one would prefer adjectival modifiers and verbal propositions as good descriptive properties for explanations, and from the examples, we can see our algorithm generates such descriptive properties because of the high information contribution of such properties to the query set. However, our algorithm does not try to reduce the redundancy within the sets of explanations. We can see redundant explanations for examples in Table 1. The reason is that each explanation added to the set is independent of the ones already present in the set. In Pantel and Vyas (2008) we propose a joint information model to overcome this problem.

5.3 Explanations using Prototypes

The algorithm discussed in Section 4.2 uses words that share the semantic category with words within the query set as the set of candidate explanations.

We can approximate the notion of semantic categories using clusters of semantically similar words. For this we used the CBC collection (Pantel and Lin 2002) of nouns. Using these clusters as semantic categories, the candidate set of all explanations is the set of all the words that belong to the same cluster. Table 2 shows some system outputs.

Table 2: Explanations generated using prototypes.

Query Sets	**Explanations**
TV, television station	station, network, radio, channel, television
Budweiser, Coors Light	Anheuser-Busch, Heineken, Coors, San Miguel, Lion Nathan
atom, electron,photon	particle, molecule, proton, Ion, isotope
Temple University,Michigan State University	University of Texas, University of Massachusetts, University of North Carolina,University of Virginia, University of Minnesota

6 Conclusions and Future Work

Computing the similarity between entities forms the basis of many computer science algorithms. However, we have little understanding of what constitutes the underlying similarity. In this paper, we investigated the problem of explaining why a set of terms are similar. We proposed an information-theoretic objective function for quantifying the utility of an explanation set, by capturing the intuition that the best explanation will be the one that is highly informative to the original query terms. We also explored various sources of explanations such as descriptive properties and prototypes. We then proposed baseline algorithms to automatically generate these types of explanations and we presented a qualitative evaluation of the baselines.

However, many other explanation sources were not addressed. Hypernyms and other hierarchical relations among words also form good explanation sources; for example the similarity of the terms $\{$*Ford, Toyota*$\}$ can be explained using the term *car*, a hypernym. Also our current explanation types would fail for query sets consisting of related terms such as $\{$*bus, road*$\}$. More appropriate for these queries would be identifying the relation linking the terms or giving analogies such as $\{$*boat, water*$\}$. We are working on algorithms to generate these explanation types within our information-theoretic framework. We are also investigating application-level quantitative evaluation methodologies. Candidate applications include providing answer support by explaining the answers generated by a QA system and explaining why a document was returned in an IR system.

References

Barker, K., Chaw, S., Fan, J., Porter, B., Tecuci, D., Yeh, P. Z., Chaudhri, V., Israel, D., Mishra, S., Romero, P., and Clark, P. 2004. *A Question-Answering System for AP Chemistry: Assessing KR&R Technologies.* In Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2004). Whistler, 488-497

Cohen, P., Schrag, R., Jones, E., Pease, A., Lin, A., Start, B., Gunning, D., and Burke, M. 1998. *The DARPA High Performance Knowledge Bases Project.* AI Magazine 19(4): 5-49.

Cover, T. M. and Thomas, J. A. 1991. *Elements of Information Theory.* Wiley Interscience, New York.

Deese, J. 1966. *The Structure of Associations in Language and Thought.* John Hopkins Press, Oxford, England.

Fellbaum, C. 1998. *WordNet: An electronic lexical database.* MIT Press.

Fillenbaum, S. 1969. *Words as feature complexes : false recognition of antonyms and synonyms* Journal of Exp.Psychology, 1969.

L Hirschman, R Gaizauskas - Natural Language Engineering 2001. *Natural language question answering: the view from here.* Natural Language Engineering.

Lee, Lillian. 1999. *Measures of Distributional Similarity.* In Proceedings of ACL-93. pp. 25-32. College Park, MD

Lester, J. C., and Porter, B. W. 1997. *Developing and Empirically Evaluating Robust Explanation Generators: The KNIGHT Experiments* Computational Linguistics, v.23(1) p.65-101.

Lin, D. and Pantel, P. 2001. *Discovery of Inference Rules for Question Answering.* Natural Language Engineering 7(4):343-360.

Lin, D. 1994. *Principar - an efficient, broad-coverage, principle-based parser.* In Proceedings of COLING-94. pp. 4248. Kyoto, Japan.

Moldovan, D. I., and Rus, V. 2001. *Logic Form Transformation of WordNet and its Applicability to Question Answering* Meeting of the Association for Computational Linguistics, p. 394-401.

Pantel, P. and Lin, D. 2002. *Discovering Word Senses from Text.* In Proceedings of SIGKDD-02. pp. 613619. Edmonton, Canada.

Pantel, P. and Vyas, V. 2008 *A Joint Information Model for n-best Ranking* In Procesedings of COLING-2008. Manchester, UK.

Rosch, E. 1975. *Cognitive representations of semantic categories.* Journal of Exp.Psychology: General, 104, 192-233.

Salton, G. and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* McGraw Hill.