Duluth : Measuring Degrees of Relational Similarity with the Gloss Vector Measure of Semantic Relatedness

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

Abstract

This paper describes the Duluth systems that participated in Task 2 of SemEval–2012. These systems were unsupervised and relied on variations of the Gloss Vector measure found in the freely available software package WordNet::Similarity. This method was moderately successful for the Class-Inclusion, Similar, Contrast, and Non-Attribute categories of semantic relations, but mimicked a random baseline for the other six categories.

1 Introduction

This paper describes the Duluth systems that participated in Task 2 of SemEval–2012, Measuring the Degree of Relational Similarity (Jurgens et al., 2012). The goal of the task was to rank sets of word pairs according to the degree to which they represented an underlying category of semantic relation. A highly ranked pair would be considered a good or prototypical example of the relation. For example, given the relation *Y functions as an X* the pair *weapon:knife* (X:Y) would likely be considered more representative of that relation than would be *tool:spoon*.

The task included word pairs from 10 different categories of relational similarity, each with a number of subcategories. In total the evaluation data consisted of 69 files, each containing a set of approximately 40 word pairs. While training examples were also provided, these were not used by the Duluth systems. The system–generated rankings were compared with gold standard data created via Amazon Mechanical Turk.

The Duluth systems relied on the Gloss Vector measure of semantic relatedness (Patwardhan and Pedersen, 2006) as implemented in WordNet::Similarity (Pedersen et al., 2004)[1]. This quantifies the degree of semantic relatedness between two word senses. It does not, however, discover or indicate the nature of the relation between the words. When given two words as input (as was the case in this task), it measures the relatedness of all possible combinations of word senses associated with this pair and reports the highest resulting score. Note that throughout this paper we use *word* and *word sense* somewhat interchangeably. In general it may be assumed that the term *word* or examples of words refers to a word sense.

A key characteristic of this task was that the word pairs in each of the 69 sets were scored assuming a particular specified underlying semantic relation. Given this, the limitation that the Gloss Vector measure does not discover the nature of relations was less of a concern, and led to the hypothesis that a word pair that was highly related would also be a prototypical example of the underlying category of semantic relation. Unfortunately the results from this task do not generally support this hypothesis, although for a few categories at least it appears to have some validity.

This paper continues with a review of the Gloss Vector measure, and explains its connections to the Adapted Lesk measure. The paper then summarizes the results of the three Duluth systems in this task, and concludes with some discussion and analysis of where this method had both successes and failures.

[1] wn-similarity.sourceforge.net

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 497–501,
Montréal, Canada, June 7-8, 2012.

2 Semantic Relatedness

Semantic relatedness is a more general notion than semantic similarity. We follow (Budanitsky and Hirst, 2006) and limit semantic similarity to those measures based on distances and perhaps depths in a hierarchy made up of *is–a* relations. For example, *car* and *motorcycle* are similar in that they are connected via an *is–a* relation with *vehicle*. Semantic similarity is most often applied to nouns, but can also be used with verbs.

Two word senses can be related in many ways, including similarity. *car* and *furnace* might be considered related because they are both made of steel, and *firefighter* and *hose* might be considered related because one uses the other, but neither pair is likely to be considered similar. Measures of relatedness generally do not specify the nature of the relationship between two word senses, but rather indicate that they are related to a certain degree in some unspecified way. As a result, measures of relatedness tend to be symmetric, so A is related to B to the same degree that B is related to A. It should be noted that some of the relations in Task 2 were not symmetric, which was no doubt a complicating factor for the Duluth systems.

3 Adapted Lesk Measure

The Gloss Vector measure was originally developed in an effort to generalize and improve upon the Adapted Lesk measure (Banerjee and Pedersen, 2003).[2] Both the Gloss Vector measure and the Adapted Lesk measure start with the idea of a *supergloss*. A supergloss is the definition (or gloss) of a word sense that is expanded by concatenating it with the glosses of other surrounding senses that are connected to it via some WordNet relation. For example, a supergloss for *car* might consist of the definition of *car*, the definition of *car's* hypernym (e.g., *vehicle*), and the definitions of the meronyms (part-of) of *car* (e.g., *wheel*, *brake*, *bumper*, etc.) Other relations as detailed later in this paper may also be used to expand a supergloss.

In the Adapted Lesk measure, the relatedness between two word senses is a function of the number and length of their matching overlaps in their superglosses. Consecutive words that match are scored more highly than single words, and a higher score for a pair of words indicates a stronger relation. The Adapted Lesk measure was developed to overcome the fact that most dictionary definitions are relatively short, which was a concern noted by (Lesk, 1986) when he introduced the idea of using definition overlaps for word sense disambiguation. While the Adapted Lesk measure expands the size of the definitions, there are still difficulties. In particular, the matches between words in superglosses must be exact, so morphological variants (*run* versus *ran*), synonyms (*gas* versus *petrol*), and closely related words (*tree* versus *shrub*) won't be considered overlaps and will be treated the same as words with no apparent connection (e.g., *goat* and *vase*).

4 Gloss Vector Measure

The Gloss Vector measure[3] is inspired by a 2nd order word sense discrimination approach (Schütze, 1998) which is in turn related to Latent Semantic Indexing or Analysis (Deerwester et al., 1990). The basic idea is to replace each word in a written context with a vector of co-occurring words as observed in some corpus. In this task, the contexts are definitions (and example text) from WordNet. A supergloss is formed exactly as described for Adapted Lesk, and then each word in the supergloss is replaced by a vector of co–occurring words. Then, all the vectors in the supergloss are averaged together to create a new high dimensional representation of that word sense. The semantic relatedness between two word senses is measured by taking the cosine between their two averaged vectors. The end result is that rather than finding overlaps in definitions based on exact matches, a word in a definition is matched to whatever degree its co-occurrences match with the co-occurrences of the words in the other supergloss. This results in a more subtle and fine grained measure of relatedness than Adapted Lesk.

The three Duluth systems only differ in the relations used to create the superglosses, otherwise they are identical. The corpus used to collect co-occurrence information was the complete collection of glosses and examples from WordNet 3.0, which consists of about 1.46 million word tokens and almost 118,000 glosses. Words that appeared in a

[2] WordNet::Similarity::lesk

[3] WordNet::Similarity::vector

stop list of about 200 common words were excluded as co-occurrences, as were words that occurred less than 5 times or more than 50 times in the WordNet corpus. Two words are considered to co-occur if they occur in the same definition (including the example) and are adjacent to each other. These are the default settings as used in WordNet::Similarity.

5 Creating the Duluth Systems

There were three Duluth systems, V0, V1, and V2. These all used the Gloss Vector measure, and differ only in how their superglosses were created. The supergloss is defined using a set of relations that indicate which additional definitions should be included in the definition for a sense. All systems start with a gloss and example for each sense in a pair, which is then augmented with definitions from additional senses as defined for each system.

5.1 Duluth-V0

V0 is identical to the default configuration of the Gloss Vector measure in WordNet::Similarity. This consists of the following relations:

hypernym (hype) : class that includes a member, e.g., a car is a kind of vehicle (hypernym).

hyponym (hypo) : the member of a class, e.g., a car (hyponym) is a kind of vehicle.

holonym (holo) : whole that includes the part, e.g., a ship (holonym) includes a mast.

meronym (mero) : part included in a whole, e.g., a mast (meronym) is a part of a ship.

see also (also) : related adjectives, e.g., egocentric see also selfish.

similar to (sim) : similar adjectives, satanic is similar to evil.

is attribute of (attr) : adjective related to a noun, e.g., measurable is an attribute of magnitude.

synset words (syns) : synonyms of a word, e.g., car and auto are synonyms.[4]

For V0 the definition and example of a noun is augmented with its synonyms and the definitions and examples of any hypernyms, hyponyms, meronyms, and holonyms to which it is directly connected. If the word is a verb it is augmented with its synonyms and any hypernyms/troponyms and hyponyms to which it is directly connected. If the word is an adjective then its definition and example are augmented with those of adjectives directly connected via see also, similar to, and is attribute of relations.

[4]Since synonyms have the same definition, this relation augments the supergloss with the synonyms themselves.

5.2 Duluth-V1

V1 uses the relations in V0, plus the holonyms, hypernyms, hyponyms, and meronyms (X) of the see also, holonym, hypernym, hyponym, and meronym relations (Y). This leads to an additional 20 relations that bring in definitions "2 steps" away from the original word. These take the form of *the holonym of the hypernym of the word sense*, or more generally *the X of the Y* of the word sense, where X and Y are as noted above.

5.3 Duluth-V2

V2 uses the relations in V0 and V1, and then adds the holonym, hypernyms, hyponyms, and meronyms of the 20 relations added for V1. This leads to an additional 80 relations of the form *the hypernyms of the meronym of the hyponym*, or more generally *the X of the X of the Y* of the word.

For example, if the word is *weapon*, then a hypernym of the meronym of the hyponym (of *weapon*) would add the definitions and example of *bow* (hyponym), *bowstring* (meronym of the hyponym), and *cord* (hypernym of the meronym of the hyponym) to the gloss of *weapon* to create the supergloss.

6 Results

There were two evaluation scores reported for the participating systems, Spearman's Rank Correlation Coefficient, and a score based on Maximum Difference Scaling. Since the Gloss Vector measure is based on WordNet, there was a concern that a lack of WordNet coverage might negatively impact the results. However, of the 2,791 pairs used in the evaluation, there were only 3 that contained words unknown to WordNet.

6.1 Spearman's Rank Correlation

The ranking of word pairs in each of the 69 files were evaluated relative to the gold standard using Spearman's Rank Correlation Coefficient. The average of these results over all 10 categories of se-

Table 1: Selected Spearman's Values

Category	rand	v0	v1	v2
SIMILAR	.026	.183	**.206**	.198
CLASS-INCLUSION	.057	.045	**.178**	.168
CONTRAST	-.049	.142	.120	**.198**
average (of all 10)	.018	**.050**	.039	.038

Table 2: Selected MaxDiff Values

Category	rand	v0	v1	v2
SIMILAR	31.5	37.1	**39.2**	37.4
CLASS-INCLUSION	31.0	29.2	**35.6**	33.1
CONTRAST	30.4	**38.3**	36.0	33.8
NON-ATTRIBUTE	28.9	**36.0**	33.0	33.5
average (of all 10)	31.2	**32.4**	31.5	31.1

mantic relations was quite low. Random guessing achieved an averaged Spearman's value 0.018, while Duluth-V0 scored 0.050, Duluth-V1 scored 0.039, and Duluth-V2 scored 0.038.

However, there were specific categories where the Duluth systems fared somewhat better. In particular, results for category 1 (CLASS-INCLUSION), category 3 (SIMILAR) and category 4 (CONTRAST) represent improvements on the random baseline (shown in Table 1) and at least some modest agreement with the gold standard.

The results from the other categories were generally equivalent to what would be obtained with random selection.

6.2 Maximum Difference Scaling

Maximum Difference Scaling is based on identifying the least and most prototypical pair for a given relation from among a set of four pairs. A random baseline scores 31.2%, meaning that it got approximately 1 in 3 of the MaxDiff questions correct. None of the Duluth systems improved upon random to any significant degree : Duluth-V0 scored 32.4, Duluth-V1 scored 31.5, and Duluth-V2 scored 31.1. However, the same categories that did well with Spearman's also did well with MaxDiff (see Table 2). In addition, there is some improvement in category 6 (NON-ATTRIBUTE) at least with MaxDiff scoring.

7 Discussion and Conclusions

The Gloss Vector measure was able to perform reasonably well in measuring the degree of relatedness for the following four categories (where the definitions come from (Bejar et al., 1991)):

CLASS-INCLUSION : one word names a class that includes the entity named by the other word

SIMILAR : one word represents a different degree or form of the ... other

CONTRAST : one word names an opposite or incompatible of the other word

NON-ATTRIBUTE : one word names a quality, property or action that is characteristically not an attribute of the other word

Of these, CLASS-INCLUSION and SIMILAR are well represented by the hypernym/hyponym relations present in WordNet and used by the Gloss Vector measure. WordNet's greatest strength lies in its hypernym tree for nouns, and that was most likely the basis for the success of the CLASS-INCLUSION and SIMILAR categories. While the success with CONTRAST may seem unrelated, in fact it may be that pairs of opposites are often quite similar, for example *happy* and *sad* are both emotions and are similar except for their polarity.

A number of the relations used in Task 2 are not well represented in WordNet. For example, there was a CASE RELATION which could benefit from information about selectional restrictions or case frames that just isn't available in WordNet. The same is true of the CAUSE-PURPOSE relation as there is relatively little information about casual relations in WordNet. While there are part-of relations in WordNet (meronyms/holonyms), these did not prove to be common enough to be a significant benefit for the PART-WHOLE relations in the task.

For many of the relations in the task the Gloss Vector measure was most likely relying primarily on hypernym and hyponym relations, which explains the bias towards categories that featured similarity-based relations. We are however optimistic that a Gloss Vector approach could be more successful given a richer set of relations from which to draw information for superglosses.

References

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, August.

I. Bejar, R. Chaffin, and S. Embretson. 1991. *Cognitive and Psychometric Analysis of Analogical Problem Solving*. Springer–Verlag, New York, NY.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

D. Jurgens, S. Mohammad, P. Turney, and K. Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, June.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM Press.

S. Patwardhan and T. Pedersen. 2006. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, April.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. Wordnet::Similarity - Measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pages 1024–1025, San Jose.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.