Vector Generation of an Explicitly-defined Multidimensional Semantic Space

Alex Grintsvayg (grinta@rpi.edu)
Vladislav D. Veksler (vekslv@rpi.edu)
Robert Lindsey (lindsr@rpi.edu)
Wayne D. Gray (grayw@rpi.edu)

Cognitive Science Department, 110 8th Street
Troy, NY 12180 USA

Measures of Semantic Relatedness (MSRs) are a recent breed of computational models of text comprehension. MSRs have been successfully used to model human web browsing behavior (Pirolli & Fu, 2003), language acquisition (Landauer & Dumais, 1997), and text comprehension (Lemaire, Denhiere, Bellissens, & Jhean-Iarose, 2006), among other things. MSRs have also been used in the applied domain for augmented search engine technology (Dumais, 2003), ETS essay grading (Landauer & Dumais, 1997), and many other applications.

The two most common types of measures of semantic relatedness are vector-based MSRs and probabilistic MSRs. Vector-based MSRs are complex, computationally expensive algorithms that represent words as vectors in a multidimensional semantic space. They work fairly well for small corpora, but the large amount of preprocessing they require makes them unusable for very large or dynamic corpora. Probabilistic MSRs are the opposite: simple metrics that can be used on an extremely large corpus. Their only downside is that they cannot compute the similarity between groups of words (something that vector-based MSRs can do easily).

In this paper we are proposing a new MSR that combines the best features of probabilistic and vector-based approaches, while adding flexibility and broadening the range of tasks that MSRs are capable of carrying out. Specifically, this technique allows non-vector-based MSRs to represent words in vector form. This representation gives probabilistic MSRs the ability to measure large multi-word terms without requiring them to perform computationally expensive preprocessing. In addition, the proposed MSR is incremental (allowing the addition of new terms to the corpus without the need for the large-scale recalculations performed by traditional vector-based measures) and has the ability to model domain-specific expertise by explicitly defining the dimensions of the semantic space that it uses. Preliminary results show that the proposed probabilistic-to-vector-based MSR conversion produces a measure that surpasses the performance of the original probabilistic MSR.

VGEM

In order to convert a probabilistic measure, *S*, into vector-based form, we use Vector Generation from Explicitly-defined Multidimensional semantic space (VGEM). VGEM's semantic space is explicitly defined by a set of words $d = \{d1, d2, ..., dn\}$, where each word defines a single dimension. To compute the vector for a word in this semantic space, VGEM uses *S* to calculate the semantic relatedness between the target word *w* and each dimension in *d*:

$$v(S,w,d) = [\, S(w,d1), S(w,d2), ..., S(w,dn) \,]$$

For example, if d = {"animal", "friend"} and the word in question is "dog", then the vector for "dog" would be [*S*("dog","animal"), *S*("dog","friend")]. If *S*("dog", "animal") is 0.81 and *S*("dog","friend") is 0.84, then the vector is v[0.81, 0.84]. See Table 1, Figure 1.

Table 1: Sample VGEM Computations

Words	Dimensions	
	Animal	Friend
Dog	0.81	0.84
Cat	0.81	0.67
Tiger	0.79	0.13
Robot	0.02	0.60

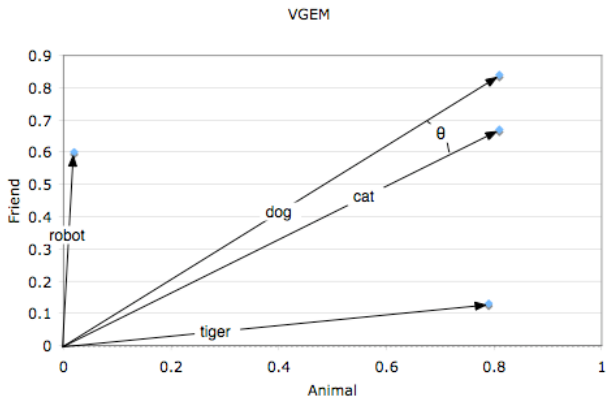

Figure 1: VGEM Semantic Space

Like all vector-based measures, VGEM defines similarity between two words to be the cosine of the angle between the vectors that represent those words. As the angle becomes smaller, and the cosine approaches 1, the words are considered more related. A value of 1 means that the two words are identical in meaning. For example, in Figure 1 the angle between “dog” and “cat” is relatively small, so the cosine of that angle will be close to 1 (.994), and the two words will be considered to be more related than any other pair of words shown.

Using this vector-based approach allows VGEM to represent a group of words as a vector sum of the words that

make up the group. For example, to compute the vector for this paragraph, VGEM would create a vector representation for each word in the paragraph and add those vectors (component by component). This vector sum will represent the meaning of the whole paragraph, and its relatedness to other vectors may be measured as the cosine of the angle between those vectors. Continuing with the example in Table 1/Figure 1, the vector to represent the words "dog cat tiger" would be the sum of first three vectors in Table 1, v[2.41, 1.64].

Advantages

The main advantage of VGEM over probabilistic MSRs is that it can compute relatedness between multi-word terms. A probabilistic MSR cannot find the similarity between two paragraphs because the probability of any two paragraphs co-occurring (word for word) in any context is virtually zero. VGEM, like other vector-based measures, can simply represent a paragraph or even a whole document as a vector, and then compare that vector to other vectors within its semantic space.

Moreover, VGEM is incremental, and does not need to pre-compute all semantic relatedness scores within the corpus before it can be used to make comparisons. Among other advantages, this lack of need for extensive preprocessing affords VGEM a larger dynamic lexicon. Other MSRs cannot handle corpora that are very large or corpora that change often (adding even a single word may require reprocessing the whole corpus).

Performance

In addition to granting probabilistic MSRs the ability to process multi-word terms by converting them into vector form, it is important to note that this conversion preserves, or possibly improves, the representative accuracy of the original measure. Here we examine the conversion of a popular probabilistic MSR, Pointwise Mutual Information (PMI), into vector-based form called VGEM-PMI (VGEM that uses PMI as its similarity metric). PMI is a computationally inexpensive technique, and it does reasonably well on most tests of language comprehension (Turney, 2001).

For the purposes of this preliminary comparison we chose 199 random words as the dimensions for VGEM, and the World Wide Web (indexed by Google) as the corpus for both measures. To evaluate MSR performance, we compared each measure (PMI and VGEM-PMI) to human word association norms (Nelson, McEvoy, & Schreiber, 1998). The association norms database that we used contains 5017 ***cue*** words that were presented to human subjects, along with the top ***target*** words that the subjects responded with for each cue. For each of the 5017 cue words, the MSR was presented with a list containing *n* target words that are related to the cue (based on the human data) and *n* random words (distractors). The *2n* words were sorted based on their semantic similarity to the cue word (as measured by the MSR). Then, the top *n* words were compared to the original *n* cue words to see how many of them matched. The score on each trial was *c/n*, where *c* is the number of words that correctly matched the originals targets. The final score for each MSR was the average of the scores across all trials.

Our preliminary results show that VGEM-PMI (M=58.04%, SE=.28%) performed better than PMI (M=52.50%, SE=.28%), $t_{\text{two-tail}}$=14.66, p<.001.

Summary and Future Work

VGEM-PMI performed better than PMI on the human word association norms test. While this result is promising, we believe that VGEM can do a lot better. In our test, we crudely defined VGEM's dimensions using 199 random words. Clearly, there are much better ways of doing this. Our future research will focus on different ways of selecting dimensions to best capture the relationships between all the words in the corpus.

Explicitly selecting VGEM's dimensions may even allow us to model domain-specific expertise. To do this, the words that constitute the dimensions could be chosen from a specific domain (e.g., politics, meteorology, or early Renaissance art). This would create an MSR that can discern the nuances of the meanings of words from the chosen domain. A modeler might create a dozen such MSRs, each proficient in a different area of expertise.

VGEM is a powerful tool for any task that could use an MSR. It is fast enough to work on any corpus, yet powerful enough to compare the meanings of whole pages of text at once. Its versatility allows it to model domain-specific expertise and learning, which might shed new light on the way in which humans acquire language.

Acknowledgments

Many thanks to Stephane Gamard for his many contributions to this project. The work was supported in part by the Disruptive Technology Office, ARIVA contract N61339-06-C-0139 issued by PEO STRI. The views and conclusions are those of the authors, not of the US Government or its agencies.

References

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science, 27*(3), 491-524.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Lemaire, B., Denhiere, G., Bellissens, C., & Jhean-Larose, S. (2006). A computational model for simulating text comprehension. *Behavior Research Methods, 38*(4), 628-637.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. : http://www.usf.edu/FreeAssociation/.

Pirolli, P., & Fu, W.-T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. *Lecture Notes in Computer Science, 2702*, 45-54.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)* (pp. 491-502). Freiburg, Germany.