Supervised Distributional Semantic Relatedness

Alistair Kennedy[1], Stan Szpakowicz[1,2]

[1] School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada
{akennedy,szpak}@eecs.uottawa.ca
[2] Institute of Computer Science
Polish Academy of Sciences, Warsaw, Poland

Abstract. Distributional measures of semantic relatedness determine word similarity based on how frequently a pair of words appear in the same contexts. A typical method is to construct a *word-context* matrix, then re-weight it using some measure of association, and finally take the vector distance as a measure of similarity. This has largely been an unsupervised process, but in recent years more work has been done devising methods of using known sets of synonyms to enhance relatedness measures. This paper examines and expands on one such measure, which learns a weighting of a *word-context* matrix by measuring associations between words appearing in a given context and sets of known synonyms. In doing so we propose a general method of learning weights for *word-context* matrices, and evaluate it on a word similarity task. This method works with a variety of measures of association and can be trained with synonyms from any resource.

1 Introduction

Measures of Semantic Relatedness (MSRs) are central to a variety of NLP tasks. In general, there are three methods of measuring semantic relatedness: resource-based methods, such as those using *WordNet* or *Roget's Thesaurus*; distributional methods, using large corpora; and hybrid methods, combining the two. Distributional MSRs rely on the hypothesis that the interchangeability of words is a strong indication of their relatedness (see [1] for an overview). If two words tend to appear in the same contexts regularly, they are more likely to be synonyms than those that do not. Usually some measure of association is used to determine the dependency between a word and the context in which it appears. This is an essentially unsupervised process. Recent work on MSRs that mix distributional and task-specific information includes [2–5]. We consider many of these methods partially supervised because they employ known sets of related words to train their system.

We describe an expansion of our supervised MSR first proposed in [6]. Our MSR reweighed a *word-context* matrix using Pointwise Mutual Information (PMI) to increase weight of contexts that tend to contain synonyms, while decreasing the weight of other contexts. We found the best results when combining supervised and unsupervised MSRs. Cosine similarity was used to measure vector distance. Our MSR resembles work where a function was learned to re-weight a matrix for measuring document simi-

$$\begin{array}{c} \\ x \in X \\ x \notin X \end{array} \begin{array}{c} \begin{array}{cc} y \in Y & y \notin Y \end{array} \\ \begin{bmatrix} O_{0,0} & O_{0,1} \\ O_{1,0} & O_{1,1} \end{bmatrix} \end{array}$$

Fig. 1. Confusion matrix of observed values.

$$\begin{array}{c} \\ x \in X \\ x \notin X \end{array} \begin{array}{c} \begin{array}{cc} y \in Y & y \notin Y \end{array} \\ \begin{bmatrix} E_{0,0} & E_{0,1} \\ E_{1,0} & E_{1,1} \end{bmatrix} \end{array}$$

Fig. 2. Confusion matrix of expected values.

larity [7, 8]. In [6] we used a tool called *SuperMatrix* [9], while now we use of our own implementation.[1] The following contributions add to our methodology from [6]:

- Evaluate several measures of association for *word-context* matrix re-weighting.
- Propose and evaluate an expansion to our supervised MSR.
- Evaluate the supervised MSR on verbs and adjectives, in addition to nouns.
- Explore training data from *WordNet* as well as *Roget's Thesaurus*.

Section 2 describes how we measure association, while Section 3 describes how these measures are applied to learning MSRs. Section 4 describes our experiments determining the best parameters for the MSRs and Section 5 concludes this work.

2 Measuring Association

A measure of association measures the dependency between two random variables, X and Y. Counts of co-occurring events $x \in X$, $y \in Y$, $x \notin X$ and $y \notin Y$ are recorded in a matrix of observed values, illustrated in Figure 1.[2] Using the observed counts the expected counts (Figure 2) are calculated with Equation 1.

$$E_{i,j} = \frac{\sum_y O_{i,y} \quad \sum_x O_{x,j}}{\sum_{x,y} O_{x,y}} \tag{1}$$

From the observed and expected counts, we calculate the dependency between X and Y using six measures of association: Pointwise Mutual Information (PMI) (Equation 3); Z-score (Equation 4); T-score (Equation 5); χ^2 (Equation 6); Log Likelihood (LL) (Equation 7); and Dice (Equation 2).

$$Dice = \frac{2 * O_{0,0}}{\sum_j O_{0,j} + \sum_i O_{i,0}} \tag{2}$$

$$PMI = \log \frac{O_{0,0}}{E_{0,0}} \tag{3}$$

$$Z\text{-}score = \frac{O_{0,0} - E_{0,0}}{\sqrt{E_{0,0}}} \tag{4}$$

$$T\text{-}score = \frac{O_{0,0} - E_{0,0}}{\sqrt{O_{0,0}}} \tag{5}$$

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{6}$$

$$LL = 2 \sum_{i,j} O_{i,j} \log \frac{O_{i,j}}{E_{i,j}} \tag{7}$$

[1] The code used in these experiments is available as a Java package called *Generalized Term Semantics* (GenTS) ⟨http://eecs.uottawa.ca/∼akennedy/Site/Resources.html⟩.

[2] The notation we use to describe this process is derived from that in [10].

Table 1. Counts of unique words, contexts and non-zero entries in the *word-context* matrices.

POS	Terms	Contexts	Non-zero Entries
Nouns	43 834	1 050 178	28 296 890
Verbs	7 141	1 423 665	25 239 485
Adjectives	17 160	360 436	8 379 637

These measures can be divided into three groups. LL and χ^2 use all observed and expected values from Figures 1 and 2. PMI, T-score and Z-score use only $O_{0,0}$ and $E_{0,0}$. Dice measures vector overlap.

3 Measures of Semantic Relatedness

This section describes how the Measures of Semantic Relatedness (MSRs) are implemented using the measures of association from Section 2. In all cases, we use cosine similarity to measure distance; the difference is how the *word-context* matrix is re-weighted. Before we can evaluate the MSRs, we must first build a *word-context* matrix.

We build a *word-context* matrix using a common procedure [11]. We use a Wikipedia dump as a corpus,[3] and parse it with *Minipar* [11] to create a set of dependency triples. An example of a triple is $\langle settle, obj, question \rangle$: the noun "question" appears as the object of the verb "settle". A dependency triple $\langle w_1, r, w_2 \rangle$ generates *word-context* pairs $(w_1, \langle r, w_2 \rangle)$ and $(w_2, \langle w_1, r \rangle)$. When the words w_1 and w_2 are used as part of a context, they can be of any part-of-speech, and all relations r are allowed. When w_1 and w_2 are the words, they must be single words with no upper case letters, digits or symbols. From these triples, we built three matrices for nouns, verbs and adjectives/adverbs.[4] One problem is that some words and contexts appear very infrequently. To remedy this, we only use nouns and adjectives that appear 35 times or more, and verbs 10 times or more. Likewise a context had to be used twice to be included.[5] We report the sizes of our matrices in Table 1.

3.1 Unsupervised Learning of Context Weights

When measuring semantic relatedness in an unsupervised fashion, we take $x \in X$ to be the appearance of a word, while $y \in Y$ is the appearance of a context. We count the following observed values:

- $O_{0,0}$ $[x \in X \wedge y \in Y]$: w_i is found in context c_j;
- $O_{0,1}$ $[x \in X \wedge y \notin Y]$: w_i is found in a context other than c_j;
- $O_{1,0}$ $[x \notin X \wedge y \in Y]$: a word other than w_i is found in context c_j;
- $O_{1,1}$ $[x \notin X \wedge y \notin Y]$: a word other than w_i is found in a context other than c_j.

The unsupervised MSR uses these counts to create a unique score for every *word-context* pair. The matrix is then re-weighted with these scores.

[3] Downloaded in August 2010.

[4] *Minipar* uses the symbol "A" for adjectives and adverbs, so we placed them in the same matrix.

[5] We found numbers by experimenting (selecting random words and generating lists of synonyms) and found that they make matrices fairly reliable.

3.2 Supervised Learning of Context Weights

A supervised MSR would use measures of association not just between words and contexts, but between pairs of words co-occurring in a context and pairs of words from our training data known to be synonyms. We calculate an association score for every context c_k. In this case, $x \in X$ represents a word pair's co-occurrence in context, $y \in Y$ – a pair of synonymous words. We explore three sources of training data coming from the 1911 and 1987 editions of *Roget's Thesaurus* and *WordNet* 3.0. We identify synonyms by selecting words from the same synset in *WordNet* or from the same Semicolon Group in *Roget's*.[6] A few examples of synonyms from the 1911 *Roget's Thesaurus*: $\langle calculator, algebraist, mathematician \rangle$ and $\langle boating, yachting \rangle$.

To calculate the association, we count pairs of words $\langle w_i, w_j \rangle$ for each context c_k:

- $O_{0,0}$ [$x \in X \wedge y \in Y$]: $\langle w_i, w_j \rangle$ are synonyms and both appear in c_k;
- $O_{0,1}$ [$x \in X \wedge y \notin Y$]: $\langle w_i, w_j \rangle$ are synonyms and only one appears in c_k;
- $O_{1,0}$ [$x \notin X \wedge y \in Y$]: $\langle w_i, w_j \rangle$ are not synonyms and both appear in c_k;
- $O_{1,1}$ [$x \notin X \wedge y \notin Y$]: $\langle w_i, w_j \rangle$ are not synonyms and only one appears in c_k.

When taking these counts, a pair can be counted multiple times if both its words appear more than once in a given context. This takes care of situations when context c_k contains a large set of unrelated words with low counts, and a small set of related words but with high counts. Now $score(c_k)$ can be calculated for every context c_k using one of the measures of association. (Negative scores are rounded up to 0.) The scores are normalized so that their average is 1.0. We then multiply the count of each word in c_k by $score(c_k)$. Some contexts contain no words from the training data, so a weight cannot be calculated. We give such contexts a score of 1.0.

We propose a second version of this training methodology. Our version finds a unique weight for every relationship r and then applies that weight to all contexts $\langle r, w_i \rangle$ .. $\langle r, w_j \rangle$. We use the same method as described above, but we combine the counts for contexts that share a common relation r. In this experiment, rather than learning contexts most appropriate for measuring semantic relatedness, we are learning which syntactic relationships best indicate semantic relatedness. The hypothesis behind this method is that the syntactic relationship is more important than the word in any given context. These two training methodologies will be distinguish by referring to them as learning at the "context" level and the "relation" level.

3.3 Combined Learning of Context Weights

In [6] we found that the best results came when mixing supervised and unsupervised learning. Supervised weighing is first performed on the matrix and then unsupervised weighting is run to reweight the matrix a second time. One problem with this methodology is identifying optimal parameters – measures of association and training type – before building a combined method. We run experiments first to identify those parameters and then construct and test this combined method.

[6] A Semicolon Group in *Roget's* contains near-synonyms, just like synset members in *WordNet*.

4 Evaluating the Measures

We first evaluate the individual supervised and unsupervised systems on a tuning set and then use them to build the combined method to be evaluated on a test set. Our evaluation task is to determine whether two words appears in the same Head in *Roget's Thesaurus*. *Roget's Thesaurus* divides the English lexicon into approximately 1000 broad categories named Heads, which represent such broad concepts as *Existence, Nonexistence, Materiality, Immateriality, Advice, Council, Reward* and *Punishment*. Each Head can contain nouns, verbs, adjectives and adverbs. To create the tuning set and the test set, we randomly create two sets of 1000 nouns, two sets of 600 verbs and two sets of 600 adjectives that appear in the 1987 *Roget's Thesaurus*. All words from the tuning set or test set are removed from the training data selected from *WordNet*, or *Roget's Thesaurus* for the supervised MSRs. We generate a long lists of all nearest neighbours for each of these words, with each measure. For example the four most related words to *psychology*, with their scores are: *sociology (0.720), anthropology (0.707), linguistics (0.582), economics (0.572)*. We evaluate these measures at a variety of recall points: the top 1, 5, 10, 20, 50 and 100 nearest neighbours. We also include an unweighted matrix as a baseline to these experiments.

4.1 Tuning Our Measure of Semantic Relatedness

We perform experiments with six different measures of association applied to three parts of speech. We use unsupervised and two kinds of supervised training, with three different training sets. In effect, there are far too many experiments to report the results in a single paper. Instead we describe and summarize the results using graphs.

The first experiment is to identify which MSR performed best in an unsupervised setting; this is summed up in Figure 3. We only present results for nouns. The results for verbs and adjectives are quite similar. Our basic findings are that most measures show a noticeable improvement over the baseline, with the exception of LL, where there is no improvement. χ^2 also perform poorly. Without exception, PMI is the superior measure, performing best at all recall points.

The second experiment is to identify the best measure of association for a supervised MSR. Once again our findings are the same for all POSs, all training data and with supervision at both the context and relation level; see Figure 4. PMI is clearly superior, though – unlike the supervised case – most other measures of association are worse than the unweighted baseline. Dice is frequently very close to the baseline, while χ^2 and LL are almost never superior at any recall point.

Having established the best measure of association, we now look to which kind of training – context or relation level – actually yields the best results. For nouns and verbs we find consistently that training at the context level is superior; see Figure 5. The Figure shows that, at most recall points, training at the context level outperforms training at the relation level. This is consistently true across all three sources of training data. For adjectives we find quite different results; see Figure 6. In this case there is a very small difference between training at the relation and context level, but more often than not training at the relation level is superior. One possible reasons for this is that the

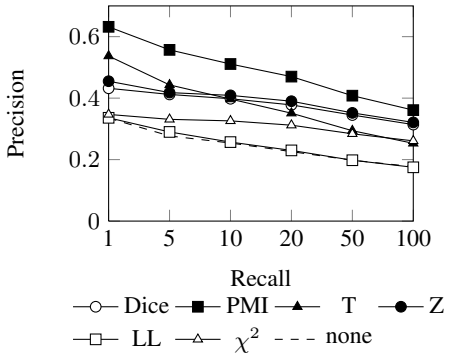

Fig. 3. Scores for nouns, unsupervised

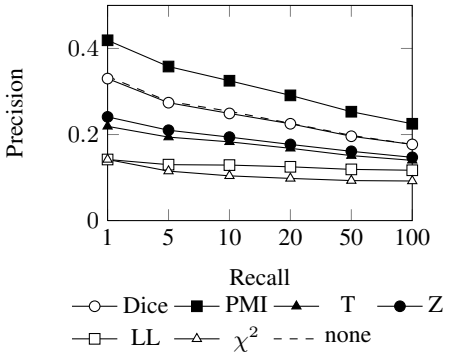

Fig. 4. Scores for nouns, supervised by context with *Roget's* 1911

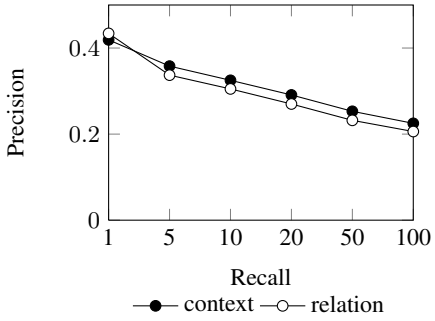

Fig. 5. Context and relation scores for nouns, trained with *Roget's* 1911

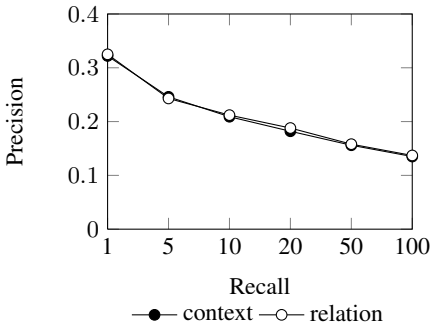

Fig. 6. Context and relation scores for adjectives, trained with *Roget's* 1911

adjective matrix is smaller than the noun and verb matrices, making it more difficult to find large groups of related or unrelated words in a given context.

4.2 Testing Our Measure of Semantic Relatedness

We have now identified the parameters for our supervised and unsupervised systems. Both the supervised and unsupervised MSRs use PMI weighting for all three POSs, while the supervised MSRs use learning at the context level for nouns and verbs and learn at the relation level for adjectives. In the tuning phase we did not attempt to identify which source of training data worked best, nor did we experiment with the combined method. This section examines both of these. The unweighted matrix makes up a lower baseline, while the unsupervised PMI re-weighted matrix makes up the high baseline. We compare the three supervised systems and three combined systems against these baselines on all three POSs; see Table 2.

The findings in Table 2 show that all supervised methods, while consistently outperforming the unweighted baselines, do not outperform the higher baseline of unsu-

Table 2. Evaluation of the various MSRs with statistically significant improvements over the Unsupervised-PMI baseline in bold.

POS	Measure	Top 1	Top 5	Top 10	Top 20	Top 50	Top 100
Nouns	Unweighted	0.376	0.296	0.262	0.239	0.207	0.186
	Unsupervised-PMI	0.645	0.579	0.537	0.490	0.423	0.374
	context-1911	0.440	0.363	0.330	0.303	0.262	0.233
	context-1987	0.456	0.376	0.334	0.296	0.252	0.223
	context-WN	0.466	0.370	0.333	0.291	0.252	0.224
	Combined-1911	0.659	**0.588**	**0.548**	**0.501**	**0.431**	**0.382**
	Combined-1987	0.651	0.584	**0.549**	**0.501**	**0.430**	**0.381**
	Combined-WN	0.654	0.586	0.541	**0.495**	**0.430**	**0.380**
Verbs	Unweighted	0.398	0.331	0.318	0.299	0.276	0.256
	Unsupervised-PMI	0.582	0.526	0.487	0.444	0.396	0.357
	context-1911	0.468	0.394	0.368	0.334	0.303	0.283
	context-1987	0.480	0.418	0.382	0.356	0.318	0.299
	context-WN	0.482	0.426	0.393	0.365	0.324	0.303
	Combined-1911	0.605	0.533	**0.500**	**0.455**	**0.401**	**0.362**
	Combined-1987	0.588	**0.537**	**0.499**	**0.453**	0.399	**0.360**
	Combined-WN	0.587	0.531	**0.495**	**0.451**	0.395	0.356
Adjectives	Unweighted	0.317	0.259	0.224	0.205	0.163	0.139
	Unsupervised-PMI	0.600	0.480	0.431	0.368	0.295	0.247
	relation-1911	0.358	0.273	0.243	0.212	0.175	0.148
	relation-1987	0.357	0.277	0.250	0.217	0.179	0.153
	relation-WN	0.353	0.278	0.242	0.213	0.175	0.148
	Combined-1911	0.602	0.484	0.431	0.368	0.296	0.247
	Combined-1987	0.603	0.483	0.431	0.367	0.296	0.247
	Combined-WN	0.595	0.483	0.430	0.368	0.296	0.247

pervised PMI. The combined systems fare much better. By harnessing elements of both supervised and unsupervised matrix re-weighting, often we can find a statistically significant improvement (the bold results in Table 2) over the high baseline of unsupervised PMI for both nouns and verbs. For adjectives, we we do not find a significant improvement using the combined MSRs. We hypothesize that this is due to the smaller matrix size. Perhaps a larger amount of data is needed before supervision can offer a meaningful benefit.

In terms of sources of training data, it would appear that the 1911 and 1987 versions of *Roget's* performed comparably. The combined system trained with the 1911 *Roget's Thesaurus* shows a significant improvement on 9 out of 12 recall points for nouns and verbs. The 1987 version significantly improves on 8 out of 12 recall points. *WordNet* 3.0 still can improve the MSRs at a statistically significant level, although only 5 times. This may not seem surprising, because we evaluate our MSRs on *Roget's Thesaurus*, so those trained using data from *Roget's Thesaurus* could have an edge. That said, it is not completely clear that identifying words in the same *Roget's* Head should benefit more from training with *Roget's* Semicolon Groups than training with *WordNet* synsets.

5 Conclusion and Discussion

We have expanded on the methods in [6] to show how our MSRs can be implemented with a variety of measures of association and applied to different parts-of-speech. We have also noted that the supervised matrix weighting can be applied in two ways: learning at the relation level and learning at the context level. Finally we explore the use of *WordNet* as training data for identifying words in the same *Roget's* Head. We have found PMI to be the strongest measure of association for all of our measures. Learning at the context level worked best for nouns and verbs, but learning at the relation level was best for adjectives. Training data from *Roget's Thesaurus* proved superior to *WordNet*'s data on our task, though *WordNet* still improved over our high baseline. Ultimately, our combined MSR has a statistically significant improvement for nouns and verbs, though for adjectives the differences are too small to determine significance.

Acknowledgments

Partially funded by the Natural Sciences and Engineering Research Council of Canada.

References

1. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research **37** (Mar 2010) 141–188
2. Patwardhan, S.: Incorporating dictionary and corpus information into a vector measure of semantic relatedness. Master's thesis, University of Minnesota, Duluth (August 2003)
3. Weeds, J., Weir, D.: Co-occurrence retrieval: A flexible framework for lexical distributional similarity. Comput. Linguist. **31**(4) (2005) 439–475
4. Mohammad, S., Hirst, G.: Distributional measures of concept-distance: A task-oriented evaluation. In Jurafsky, D., Gaussier, É., eds.: EMNLP, ACL (2006) 35–43
5. Hagiwara, M., Ogawa, Y., Toyama, K.: Supervised synonym acquisition using distributional features and syntactic patterns. Journal of Natural Language Processing. **16** (2009 2005) 59–83
6. Kennedy, A., Szpakowicz, S.: A supervised method of feature weighting for measuring semantic relatedness. In: Proceedings of Canadian AI 2011, Ottawa, Ontario, Canada, Springer (2011) 222–233
7. Yih, W.t.: Learning term-weighting functions for similarity measures. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. EMNLP '09, Morristown, NJ, USA, Association for Computational Linguistics (2009) 793–802
8. Hajishirzi, H., Yih, W.t., Kolcz, A.: Adaptive near-duplicate detection via similarity learning. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '10, New York, NY, USA, ACM (2010) 419–426
9. Broda, B., Piasecki, M.: Supermatrix: a general took for lexical semantic knowledge acquisition. Technical report, Institute of Applied Informatics, Wroclaw University of Technology, Poland (2008)
10. Evert, S.: The statistics of word cooccurrences: word pairs and collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart (2004)
11. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1998) 768–774