TCDSCSS: Dimensionality Reduction to Evaluate Texts of Varying Lengths - an IR Approach

Arun Jayapal
Dept of Computer Science
Trinity College Dublin
jayapala@cs.tcd.ie

Martin Emms
Dept of Computer Science
Trinity College Dublin
martin.emms@cs.tcd.ie

John D.Kelleher
School of Computing
Dublin Institute of Technology
john.d.kelleher@dit.ie

Abstract

This paper provides system description of the cross-level semantic similarity task for the SEMEVAL-2014 workshop. Cross-level semantic similarity measures the degree of relatedness between texts of varying lengths such as *Paragraph to Sentence* and *Sentence to Phrase*. *Latent Semantic Analysis* was used to evaluate the cross-level semantic relatedness between the texts to achieve above baseline scores, tested on the training and test datasets. We also tried using a bag-of-vectors approach to evaluate the semantic relatedness. This bag-of-vectors approach however did not produced encouraging results.

1 Introduction

Semantic relatedness between texts have been dealt with in multiple situations earlier. But it is not usual to measure the semantic relatedness of texts of varying lengths such as *Paragraph to Sentence* (P2S) and *Sentence to Phrase* (S2P). This task will be useful in natural language processing applications such as *paraphrasing* and *summarization*. The working principle of information retrieval system is the motivation for this task, where the queries are not of equal lengths compared to the documents in the index. We attempted two ways to measure the semantic similarity for P2S and S2P in a scale of 0 to 4, 4 meaning both texts are similar and 0 being dissimilar. The first one is Latent Semantic Analysis (LSA) and second, a bag-of-vecors (BV) approach. An example of target similarity ratings for comparison type *S2P* is provided in table 1.

Sentence: *Schumacher was undoubtedly one of the very greatest racing drivers there has ever been, a man who was routinely, on every lap, able to dance on a limit accessible to almost no-one else.*

Score	Phrase
4	the unparalleled greatness of Schumachers driving abilities
3	driving abilities
2	formula one racing
1	north-south highway
0	orthodontic insurance

Table 1: An Example - *Sentence to Phrase* similarity ratings for each scale

2 Data

The task organizers provided training data, which included 500 pairs of *P2S*, *S2P*, *Phrase to Word* (P2W) and their similarity scores. The training data for *P2S* and *S2P* included text from different genres such as Newswire, Travel, Metaphoric and Reviews. In the training data for *P2S*, newswire text constituted 36% of the data, while reviews constituted 10% of the data and rest of the three genres shared 54% of the data.

Considering the different genres provided in the training data, a chunk of data provided for NIST TAC's Knowledge Base Population was used for building a term-by-document matrix on which to base the LSA method. The data included newswire text and web-text, where the web-text included data mostly from blogs. We used 2343 documents from the NIST dataset[1], which were available in eXtended Markup Language format.

Further to the NIST dataset, all the paragraphs in the training data[2] of *paragraph to sentence* were added to the dataset. To add these paragraphs to the dataset, we converted each paragraph into a

[1]Distributed by LDC (Linguistic Data Consortium)
[2]provided by the SEMEVAL task-3 organizers

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 619–623,
Dublin, Ireland, August 23-24, 2014.

new document and the documents were added to the corpus. The unique number of words identified in the corpus were approximately 40000.

3 System description

We tried two different approaches for evaluating the P2S and S2P. Latent Semantic Analysis (LSA) using SVD worked better than the Bag-of-Vectors (BV) approach. The description of both the approaches are discussed in this section.

3.1 Latent Semantic Analysis

LSA has been used for information retrieval allowing retrieval via vectors over latent, arguably conceptual, dimensions, rather than over surface word dimensions (Deerwester et al., 1990). It was thought this would be of advantage for comparison of texts of varying length.

3.1.1 Representation

The data corpus was converted into a mxn term-by-document matrix, A, where the counts ($c_{m,n}$) of all terms (w_m) in the corpus are represented in rows and the respective documents (d_n) in columns:

$$A = \begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix} \end{array}$$

The document indexing rules such as text tokenization, case standardization, stop words removal, token stemming, and special characters and punctuations removal were followed to get the matrix A.

Singular Value Decomposition (SVD) decomposes the matrix into U, Σ and V matrices (ie., $A = U\Sigma V^T$) such that U and V are orthonormal matrices and Σ is a diagonal matrix with singular values. Retaining just the first k columns of U and V, gives an approximation of A

$$A \approx A_k = U_k \Sigma_k V_k^T \quad (1)$$

According to LSA, the columns of U_k are thought of as representing latent, *semantic* dimensions, and an arbitrary m-dimensional vector $\vec{v}$ can be projected onto this semantic space by taking the dot-product with each column of U_k; we will call the result $\overrightarrow{v_{sem}}$.

In the experiments reported later, the m-dimensional vector $\vec{v}$ is sometimes a vector of word counts, and sometimes a thresholded or 'boolean' version, mapping all non-zero numbers to 1.

3.1.2 Similarity Calculation

To evaluate the similarity of a paragraph, p, and a sentence, s, first these are represented as vectors of word counts, $\vec{p}$ and $\vec{s}$, then these are projected in the latent semantic space, to give $\overrightarrow{p_{sem}}$ and $\overrightarrow{s_{sem}}$, and then between these the cosine similarity metric is calculated:

$$cos(\overrightarrow{p_{sem}} . \overrightarrow{s_{sem}}) = \frac{\overrightarrow{p_{sem}} . \overrightarrow{s_{sem}}}{|\overrightarrow{p_{sem}}| . |\overrightarrow{s_{sem}}|} \quad (2)$$

The cosine similarity metric provides a similarity value in the range of 0 to 1, so to match the target range of 0 to 4, the cosine values were multiplied by 4. Exactly the same procedure is used for the sentence to phrase comparison.

Further, the number of retained dimensions of U_k was varied, giving different dimensionalities of the LSA space. The results of testing at the reduced dimensions are discussed in 4.1

3.2 Bag-of-Vectors

Another method we experimented on could be termed a 'bag-of-vectors' (BV) approach: each word in an item to be compared is replaced by a vector representing its co-occurrence behavior and the obtained bags of vectors enter into the comparison process.

3.2.1 Representation

For the BV approach, the same data sources as was used for the LSA approach is turned into a $m \times m$ term-by-term co-occurrence matrix C:

$$C = \begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \cdots & w_m \end{array} \\ \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,m} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,m} \end{pmatrix} \end{array}$$

The same preprocessing steps as for the LSA approach applied (text tokenization, case standardization, stop words removal, special characters and punctuations removal). Via C, if one has a bag-of-words representing a paragraph, sentence or phrase, one can replace it by a bag-of-vectors, replacing each word w_i by the corresponding row of C – we will call these rows word-vectors.

3.2.2 Similarity Calculation

For calculating P2S similarity, the procedure is as follows. The paragraph and sentence are tokenized, and stop-words were removed and are represented as two vectors $\vec{p}$ and $\vec{s}$.

For each word p_i from $\vec{p}$, its word vector from C is found, and this is compared to the word vector for each word s_i in $\vec{s}$, via the cosine measure. The highest similarity score for each word p_i in $\vec{p}$ is stored in a vector $\overrightarrow{S_p}$ shown in (3). The overall semantic similarity score between paragraph and sentence is then the mean value of the vector $\overrightarrow{S_p} \times 4$ – see (4).

$$S_p = \begin{bmatrix} S_{p_1} & S_{p_2} & \cdots & S_{p_i} \end{bmatrix} \quad (3)$$

$$S_{sim} = \frac{\sum_{i=1}^{n} S_{p_i}}{n} \times 4 \quad (4)$$

Exactly corresponding steps are carried out for the *S2P* similarity. Although experiments were carried out this particular BV approach, the results were not encouraging. Details of the experiments carried out are explained in 4.2.

4 Experiments

Different experiments were carried out using LSA and BV systems described in sections 3.1 and 3.2 on the dataset described in section 2. Pearson correlation and Spearman's rank correlation were the metrics used to evaluate the performance of the systems. Pearson correlation provides the degree of similarity between the system's score for each pair and the gold standard's score for the said pair while Spearman's rank correlation provides the degree of similarity between the rankings of the pairs according to similarity.

4.1 LSA

The LSA model was used to evaluate the semantic similarity between *P2S* and *S2P*.

4.1.1 Paragraph to Sentence

An initial word-document matrix A was built by extracting tokens just based on spaces, stop words removed and tokens sorted in alphabetical order. As described in 3.1.1, via the SVD of A, a matrix U_k is obtained which can be used to project an m dimensional vector into a k dimensional one. In one setting the paragraph and sentence vectors which are projected into the LSA space have unique word counts for their dimensions. In another setting before projection, these vectors are thresholded into 'boolean' versions, with 1 for every non-zero count.

Dimensions	100%	90%	50%	30%	10%
Basic word-doc representation	0.499	-	0.494	0.484	0.426
Evaluation-boolean counts	0.548	-	0.533	0.511	0.420
Constrained tokenization	0.368	0.564	0.540	0.516	0.480
Added data	0.461	0.602	0.568	0.517	0.522

Table 2: Pearson scores at different dimensions - *Paragraph to Sentence*

The Pearson scores for these settings are in the first and second rows of table 2. They show the variation with the number of dimensions of the LSA representation (that is the number of columns of U that are kept)[3]. An observation is that the usage of boolean values instead of word counts showed improved results.

Further experiments were conducted, retaining the boolean treatment of the vectors to be projected. In a new setting, further improvements were made to the pre-processing step, creating a new word-document matrix A using constrained tokenization rules, removing unnecessary spaces and tabs, and tokens stemmed[4]. The performance of the similarity calculation is shown as the third row of Table 2: there is a trend of increase in correlation scores with respect to the increase in dimensionality up to a maximum of 0.564, reached at 90% dimension.

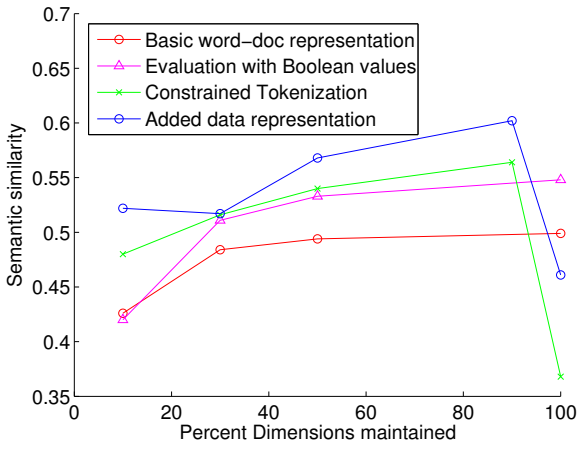

Figure 1: Paragraph to Sentence - Pearson correlation scores for four different experiments at different dimensions[3] (represented in percent) of U_k

Not convinced with the pearson scores, more

[3]Here, the dimension $X\%$ means $k = (X/100) \times N$, where N is the total number of columns in A in the unreduced SVD.

[4]Stemmed using Porter Stemmer module availabe from http://tartarus.org/∼martin/PorterStemmer/

documents were added to the dataset to build a new word-document matrix representation A. The documents included all the paragraphs from the training set. Each paragraph provided in the training set was added to the dataset as a separate document. The experiment was performed maintaining the settings from the previous experiment and the results are shown in the fourth row of table 2. The increase in trend of correlation scores with respect to the increase in dimensionality is followed by the new U produced from A after applying SVD. Figure 2 provides the distribution of similarity scores evaluated at 90% dimension of the model with respect to the gold standard.

Further to compare the performance of different experiments, all the experiment results are plotted in Figure 1. It can be observed that every subsequent model built has shown improvements in performance. The first two experiments shown in the first two rows of table 2 are shown in red and blue lines in the figure. It can be observed that in both the settings, the pearson correlation scores were increasing as the the number of dimensions maintained also increased, whereas in the other two settings, the pearson correlation scores reached their maximum at 90% and came down at 100% dimension, which is unexpected and so is not justified. It is observed from Figure 2 that the scores

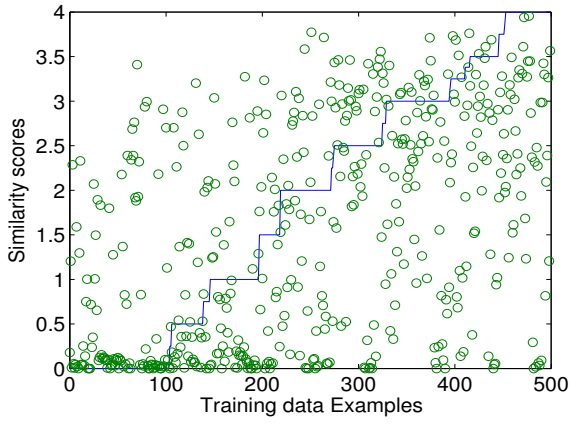

Figure 2: Semantic similarity scores - Gold standard (Line plot) vs System scores (Scatter plot) for examples in training data

of the system in scatter plot are not always clustered around the gold standard scores, plotted as a line. As the gold standard score goes up, the system prediction accuracy has come down. One reason for this pattern can be attributed to the training set which had data mostly data from Newswire and webtext. Therefore, during evaluation all the words from paragraph and/or sentence would not have got a position while getting projected on the latent semantic space, which we believe has pulled down the accuracy.

Dimensions	100%	90%	70%	50%	30%	10%
Basic word-doc representation	0.493	-	-	0.435	0.423	0.366
Evaluation boolean counts	0.472	-	-	0.449	0.430	0.363
Constrained tokenization	0.498	0.494	0.517	0.485	0.470	0.434
Added data	0.493	0.504	0.498	0.498	0.488	0.460

Table 3: Pearson scores at different dimensions[3] - *Sentence to Phrase*

4.1.2 Sentence to Phrase

The experiments carried out for *P2S* provided in 4.1.1 were conducted for *S2P* examples as well. The pearson scores produced by different experiments at different dimensions are provided in table 3. This table shows that the latest word-document representation made with added documents, did not have any impact on the correlation scores, while the earlier word-document representation provided in 3^{rd} row, which used the original dataset preprocessed with constrained tokenization rules, removing unnecessary spaces and tabs, and tokens stemmed, provided better correlation score at 70% dimension. Further the comparison of different experiments carried out at different settings are plotted in Figure 3.

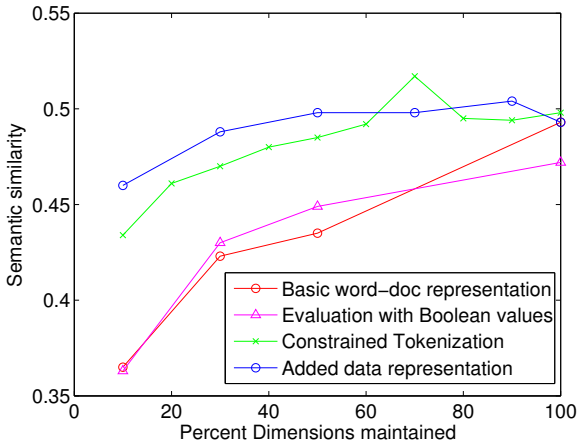

Figure 3: Sentence to Phrase - Pearson correlation scores for four different experiments at different dimensions[3] (represented in percentage) of U_k

4.2 Bag of Vectors

BV was tested in two different settings. The first representation was created with bi-gram co-occurance count as mentioned in section 3.2.1 and experiments were carried out as mentioned in section 3.2.2. This produced negative Pearson correlation scores for *P2S* and *S2P*. Then we created another representation by getting co-occurance count in a window of 6 words in a sentence, on evaluation produced correlation scores of 0.094 for *P2S* and 0.145 for *S2P*. As BV showed strong negative results, we did not continue using the method for evaluating the test data. But we strongly believe that the BV approach can produce better results if we could compare the sentence to the paragraph rather than the paragraph to the sentence as mentioned in section 3.2.2. During similarity calculation, when comparing sentence to the paragraph, for each word in the sentence, we look for the best semantic match from the paragraph, which would increase the mean value by reducing the number of divisions representing the number of words in the sentence. In the current setting, it is believed that while computing the similarity for the paragraph to sentence, the words in the paragraph (longer text) will consider a few words in the sentence to be similar multiple times. This could not be right when we compare the texts of varying lengths.

5 Conclusion and Discussion

On manual verification, it was identified that the dataset used to build the representation did not have documents related to the genres Metaphoric, CQA and Travel. The original dataset mostly had documents from Newswire text and blogs which included reviews as well. Further, it can be identified from tables 2 and 3, the word-document representation with added documents from the training set improved Pearson scores. This allowed to assume that the dataset did not have completely relevant set of documents to evaluate the training set which included data from different genres. For evaluation of the model on test data, we submitted two runs and best of them reported Pearson score of 0.607 and 0.552 on *P2S* and *S2P* respectively. In the future work, we should be able to experiment with more relevant data to build the model using LSI and also use statistically strong unsupervised classifier pLSI (Hofmann T, 2001) for the same task. Further to this, as discussed in 4.2 we would be able to experiment with the BV approach by comparing the sentence to the paragraph, which we believe will yield promising results to compare the texts of varying lengths.

References

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman 1990. *Indexing by latent semantic analysis* Journal of the American society for information science, 41(6):391–401

Thomas Hofmann 2001. *Unsupervised Learning by Probabilistic Latent Semantic Analysis* Journal Machine Learning, Volume 42 Issue 1-2, January-February 2001 Pages 177 - 196