Dimensionality Reduction Aids Term Co-Occurrence Based Multi-Document Summarization

Ben Hachey, Gabriel Murray & David Reitter
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW
bhachey@inf.ed.ac.uk, gabriel.murray@ed.ac.uk, dreitter@inf.ed.ac.uk

Abstract

A key task in an extraction system for query-oriented multi-document summarisation, necessary for computing relevance and redundancy, is modelling text semantics. In the Embra system, we use a representation derived from the singular value decomposition of a term co-occurrence matrix. We present methods to show the reliability of performance improvements. We find that Embra performs better with dimensionality reduction.

1 Introduction

We present experiments on the task of query-oriented multi-document summarisation as explored in the DUC 2005 and DUC 2006 shared tasks, which aim to model real-world complex question-answering. Input consists of a detailed query[1] and a set of 25 to 50 relevant documents. We implement an extractive approach where pieces of the original texts are selected to form a summary and then smoothing is performed to create a discursively coherent summary text.

The key modelling task in the extraction phase of such a system consists of estimating responsiveness to the query and avoiding redundancy. Both of these are often approached through some textual measure of semantic similarity. In the Embra[2] system, we follow this approach in a sentence extraction framework. However, we model the semantics of a sentence using a very large distributional semantics (i.e. term co-occurrence) space reduced by singular value decomposition. Our hypothesis is that this dimensionality reduction using a large corpus can outperform a simple term co-occurrence model.

A number of papers in the literature look at singular value decomposition and compare it to unreduced term $\times$ document or term co-occurrence matrix representations. These explore varied tasks and obtain mixed results. For example, Pedersen et al. (2005) find that SVD does not improve performance in a name discrimination task while Matveeva et al. (2005) and Rohde et al. (In prep) find that dimensionality reduction with SVD does help on word similarity tasks.

The experiments contained herein investigate the contribution of singular value decomposition on the query-oriented multi-document summarisation task. We compare the singular value decomposition of a term co-occurrence matrix derived from a corpus of approximately 100 million words (DS+SVD) to an unreduced version of the matrix (DS). These representations are described in Section 2. Next, Section 3 contains a discussion of related work using SVD for summarisation and a description of the sentence selection component in the Embra system. The paper goes on to give an overview of the experimental design and results in Section 4. This includes a detailed analysis of the statistical significance of the results.

2 Representing Sentence Semantics

The following three subsections discuss various ways of representing sentence meaning for information extraction purposes. While the first approach relies solely on weighted term frequencies in a vector space, the subsequent methods attempt to use term context information to better represent the meanings of sentences.

[1] On average, queries contain approximately 34 words words and three sentences.

[2] Edinburgh Multi-document Breviloquence Assay

2.1 Terms and Term Weighting (TF.IDF)

The traditional model for measuring semantic similarity in information retrieval and text mining is based on a vector representation of the distribution of terms in documents. Within the vector space model, each term is assigned a weight which signifies the semantic importance of the term. Often, *tf.idf* is used for this weight, which is a scheme that combines the importance of a term within the current document[3] and the distribution of the term across the text collection. The former is often represented by the term frequency and the latter by the inverse document frequency ($idf_i = \frac{N}{df_i}$), where N is the number of documents and df_i is the number of documents containing term t_i.

2.2 Term Co-occurrence (DS)

Another approach eschews the traditional vector space model in favour of the distributional semantics approach. The DS model is based on the intuition that two words are semantically similar if they appear in a similar set of contexts. We can obtain a representation of a document's semantics by averaging the context vectors of the document terms. (See Besançon et al. (1999), where the DS model is contrasted with a term $\times$ document vector space representation.)

2.3 Singular Value Decomposition (DS+SVD)

Our third approach uses dimensionality reduction. Singular value decomposition is a technique for dimensionality reduction that has been used extensively for the analysis of lexical semantics under the name of latent semantic analysis (Landauer et al., 1998). Here, a rectangular (e.g., term $\times$ document) matrix is decomposed into the product of three matrices ($X_{w\times p} = W_{w\times n}S_{n\times n}(P_{p\times n})^T$) with n 'latent semantic' dimensions. W and P represent terms and documents in the new space. And S is a diagonal matrix of singular values in decreasing order.

Taking the product $W_{w\times k}S_{k\times k}(P_{p\times k})^T$ over the first k columns gives the best least square approximation of the original matrix X by a matrix of rank k, i.e. a reduction of the original matrix to k dimensions. Similarity between documents can then be computed in the space obtained by taking the rank k product of S and P.

[3]The local importance of a term can also be computed over other textual units, e.g. sentence in extractive summarisation or the context of an entity pair in relation discovery.

This decomposition abstracts away from terms and can be used to model a semantic similarity that is more linguistic in nature. Furthermore, it has been successfully used to model human intuitions about meaning. For example, Landauer et al. (1998) show that latent semantic analysis correlates well with human judgements of word similarity and Foltz (1998) shows that it is a good estimator for textual coherence.

It is hoped that these latter two techniques (dimensionality reduction and the DS model) will provide for a more robust representation of term contexts and therefore better representation of sentence meaning, enabling us to achieve more reliable sentence similarity measurements for extractive summarisation.

3 SVD in Summarisation

This section describes ways in which SVD has been used for summarisation and details the implementation in the Embra system.

3.1 Related Work

In seminal work by Gong and Liu (2001), the authors proposed that the rows of P^T may be regarded as defining topics, with the columns representing sentences from the document. In their SVD method, summarisation proceeds by choosing, for each row in P^T, the sentence with the highest value. This process continues until the desired summary length is reached.

Steinberger and Ježek (2004) have offered two criticisms of the Gong and Liu approach. Firstly, the method described above ties the dimensionality reduction to the desired summary length. Secondly, a sentence may score highly but never "win" in any dimension, and thus will not be extracted despite being a good candidate. Their solution is to assign each sentence an SVD-based score using:

$$Sc_i^{SVD} = \sqrt{\sum_{i=1}^{n} v(i,k)^2 * \sigma(k)^2}\,,$$

where $v(i,k)$ is the kth element of the ith sentence vector and $\sigma(k)$ is the corresponding singular value.

Murray et al. (2005a) address the same concerns but retain the Gong and Liu framework. Rather than extracting the best sentence for each topic, the n best sentences are extracted, with n determined by the corresponding singular values from

matrix S. Thus, dimensionality reduction is no longer tied to summary length and more than one sentence per topic can be chosen.

A similar approach in DUC 2005 using term co-occurrence models and SVD was presented by Jagarlamudi et al. (2005). Their system performs SVD over a term $\times$ sentence matrix and combines a relevance measurement based on this representation with relevance based on a term co-occurrence model by a weighted linear combination.

3.2 Sentence Selection in Embra

The Embra system developed for DUC 2005 attempts to derive more robust representations of sentences by building a large semantic space using SVD on a very large corpus. While researchers have used such large semantic spaces to aid in automatically judging the coherence of documents (Foltz et al., 1998; Barzilay and Lapata, 2005), to our knowledge this is a novel technique in summarisation.

Using a concatenation of Aquaint and DUC 2005 data (100+ million words), we utilised the Infomap tool[4] to build a semantic model based on singular value decomposition (SVD). The decomposition and projection of the matrix to a lower-dimensionality space results in a semantic model based on underlying term relations. In the current experiments, we set dimension of the reduced representation to 100. This is a reduction of 90% from the full dimensionality of 1000 content-bearing terms in the original DS matrix. This was found to perform better than 25, 50, 250 and 500 during parameter optimisation. A given sentence is represented as a vector which is the average of its constituent word vectors. This sentence representation is then fed into an MMR-style algorithm.

MMR (Maximal Marginal Relevance) is a common approach for determining relevance and redundancy in multi-document summarisation, in which candidate sentences are represented as weighted term-frequency vectors which can thus be compared to query vectors to gauge similarity and already-extracted sentence vectors to gauge redundancy, via the cosine of the vector pairs (Carbonell and Goldstein, 1998). While this has proved successful to a degree, the sentences are represented merely according to weighted term frequency in the document, and so two similar sentences stand a chance of not being considered similar if they do not share the same terms.

[4] `http://infomap.stanford.edu/`

```
for each sentence in document:
    for each word in sentence:
        get word vector from semantic model
    average word vectors to form sentence vector
    sim1 = cossim(sentence vector, query vector)
    sim2 = highest(cossim(sentence vector, all extracted vectors))
    score = λ*sim1 - (1-λ)*sim2
extract sentence with highest score
repeat until desired length
```

Figure 1: Sentence extraction algorithm

Our implementation of MMR (Figure 1) uses λ annealing following (Murray et al., 2005a). λ decreases as the summary length increases, thereby emphasising relevance at the outset but increasingly prioritising redundancy removal as the process continues.

4 Experiment

The experimental setup uses the DUC 2005 data (Dang, 2005) and the Rouge evaluation metric to explore the hypothesis that query-oriented multi-document summarisation using a term co-occurrence representation can be improved using SVD. We frame the research question as follows:

> Does SVD dimensionality reduction lead to an increase in Rouge score compared to the DS representation?

4.1 Materials

The DUC 2005 task[5] was motivated by Amigo et al.'s (2004) suggestion of evaluations that model real-world complex question answering. The goal is to synthesise a well-organised, fluent answer of no more than 250 words to a complex question from a set of 25 to 50 relevant documents. The data includes a detailed query, a document set, and at least 4 human summaries for each of 50 topics.

The preprocessing was largely based on LT TTT and LT XML tools (Grover et al., 2000; Thompson et al., 1997). First, we perform tokenisation and sentence identification. This is followed by lemmatisation.

At the core of preprocessing is the LT TTT program *fsgmatch*, a general purpose transducer which processes an input stream and adds annotations using rules provided in a hand-written grammar file. We also use the statistical combined part-of-speech (POS) tagger and sentence boundary disambiguation module from LT TTT (Mikheev,

[5] `http://www-nlpir.nist.gov/projects/duc/duc2005/tasks.html`

1997). Using these tools, we produce an XML markup with sentence and word elements. Further linguistic markup is added using the *morpha* lemmatiser (Minnen et al., 2000) and the *C&C* named entity tagger (Curran and Clark, 2003) trained on the data from MUC-7.

4.2 Methods

The different system configurations (DS, DS+SVD, TF.IDF) were evaluated against the human upper bound and a baseline using Rouge-2 and Rouge-SU4. Rouge estimates the coverage of appropriate concepts (Lin and Hovy, 2003) in a summary by comparing it several human-created reference summaries. Rouge-2 does so by computing precision and recall based on macro-averaged bigram overlap. Rouge-SU4 allows bigrams to be composed of non-contiguous words, with as many as four words intervening. We use the same configuration as the official DUC 2005 evaluation,[6] which is based on word stems (rather than full forms) and uses jackknifing ($k-1$ cross-evaluation) so that human gold-standard and automatic system summaries can be compared.

The independent variable in the experiment is the model of sentence semantics used by the sentence selection algorithm. We are primarily interested in the relative performance of the DS and DS+SVD representations. As well as this, we include the DUC 2005 baseline, which is a lead summary created by taking the first 250 words of the most recent document for each topic. We also include a *tf.idf*-weighted term $\times$ sentence representation (TF.IDF) for comparison with a conventional MMR approach.[7] Finally, we include an upper bound calculated using the DUC 2005 human reference summaries. Preprocessing and all other aspects of the sentence selection algorithm remain constant over all systems.

In general, Rouge shows a large variance across data sets (and so does system performance). It is important to test whether obtained nominal differences are due to chance or are actually statistically significant.

To test whether the Rouge metric showed a reliably different performance for the systems, the Friedman rank sum test (Friedman, 1940; Demšar, 2006) can be used. This is a hypothesis test not unlike an ANOVA, however, it is non-parametric, i.e. it does not assume a normal distribution of the measures (i.e. precision, recall and F-score). More importantly, it does not require homogeneity of variances.

To (partially) rank the systems against each other, we used a cascade of Wilcoxon signed ranks tests. These tests are again non-parametric (as they rank the differences between the system results for the datasets). As discussed by Demšar (2006), we used Holm's procedure for multiple tests to correct our error estimates (p).

p	Metric	hypothesis
0.000262	Rouge-2	base<TF.IDF ***
0.021640	Rouge-2	base<DS *
0.000508	Rouge-2	base<DS+SVD ***
0.014845	Rouge-2	DS<TF.IDF *
0.507702	Rouge-2	TF.IDF<DS+SVD
0.047016	Rouge-2	DS<DS+SVD *
0.000080	Rouge-SU4	base<TF.IDF ***
0.006803	Rouge-SU4	base<DS **
0.000006	Rouge-SU4	base<DS+SVD ***
0.012815	Rouge-SU4	DS<TF.IDF *
0.320083	Rouge-SU4	TF.IDF<DS+SVD
0.001053	Rouge-SU4	DS<DS+SVD **

Table 1: Holm-corrected Wilcoxon hypothesis test results.

4.3 Results

Friedman tests for each Rouge metric (with F-score, precision and recall included as observations, with the dataset as group) showed a reliable effect of the system configuration ($\chi^2_{F,SU4} = 106.6$, $\chi^2_{P,SU4} = 96.1$, $\chi^2_{R,SU4} = 105.5$, all $p < 0.00001$).

Post-hoc analysis (Wilcoxon) showed (see Table 1) that all three systems performed reliably better than the baseline. TF.IDF performed better than simple DS in Rouge-2 and Rouge-SU4. DS+SVD performed better than DS ($p_2 < 0.05$, $p_{SU4} < 0.005$). There is no evidence to support a claim that DS+SVD performed differently from TF.IDF.

However, when we specifically compared the performance of TF.IDF and DS+SVD with the Rouge-SU4 F score for only the specific (as opposed to general) summaries, we found that DS+SVD scored reliably, but only slightly better

[6] i.e. `ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 d`

[7] Specifically, we use $tf_{i,j} * log(\frac{N}{df_i})$ for term weighting where $tf_{i,j}$ is the number of times term i occurs in sentence j, N is the number of sentences, and df_i is the number of sentences containing term i.

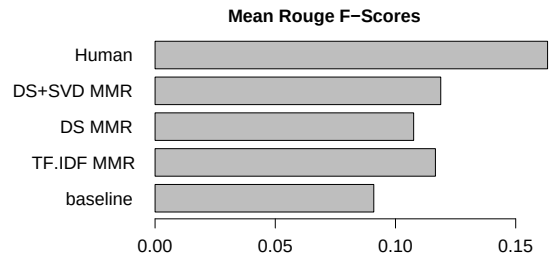

Figure 2: Mean system performance over 50 datasets (F-scores). Precision and Recall look qualitatively similar.

(Wilcoxon, $p<0.05$). This result is unadjusted, and post-hoc comparisons with other scores or for the general summaries did not show reliable differences.

Having established the reliable performance improvement of DS+SVD over DS, it it important to take the effect size into consideration (with enough data, small effects may be statistically significant, but practically unimportant). Figure 2 illustrates that the gain in mean performance is substantial. If the mean Rouge-SU4 score for human performance is seen as upper bound, the DS+SVD system showed a 25.4 percent reduction in error compared to the DS system.[8]

A similar analysis for precision and recall gives qualitatively comparable results.

5 Discussion and Future Work

The positive message from the experimental results is that SVD dimensionality reduction improves performance over a term co-occurrence model for computing relevance and redundancy in a MMR framework. We note that we cannot conclude that the DS or DS+SVD systems outperform a conventional *tf.idf*-weighted term $\times$ sentence representation on this task. However, results from Jagarlamudi et al. (2005) suggest that the DS and term $\times$ sentence representations may be complementary in which case we would expect a further improvement through an ensemble technique.

Previous results comparing SVD with unreduced representations show mixed results. For example, Pedersen et al. (2005) experiment with term co-occurrence representations with and without SVD on a name discrimination task and find that the unreduced representation tends to perform better. Rohde et al. (In prep), on the other hand, find that a reduced matrix does perform better on word pair similarity and multiple-choice vocabulary tests. One crucial factor here may be the size of the corpus. SVD may not offer any reliable 'latent semantic' advantage when the corpus is small, in which case the efficiency gain from dimensionality reduction is less of a motivation anyway.

We plan to address the question of corpus size in future work by comparing DS and DS+SVD derived from corpora of varying size. We hypothesise that the larger the corpus used to compile the term co-occurrence information, the larger the potential contribution from dimensionality reduction. This will be explored by running the experiment described in this paper a number of times using corpora of different sizes (e.g. 0.5m, 1m, 10m and 100m words).

Unlike official DUC evaluations, which rely on human judgements of readability and informativeness, our experiments rely solely on Rouge n-gram evaluation metrics. It has been shown in DUC 2005 and in work by Murray et al. (2005b; 2006) that Rouge does not always correlate well with human evaluations, though there is more stability when examining the correlations of macro-averaged scores. Rouge suffers from a lack of power to discriminate between systems whose performance is judged to differ by human annotators.

Thus, it is likely that future human evaluations would be more informative. Another way that the evaluation issue might be addressed is by using an annotated sentence extraction corpus. This could proceed by comparing gold standard alignments between abstract and full document sentences with predicted alignments using correlation analysis.

6 Conclusions

We have presented experiments with query-oriented multi-document summarisation. The experiments explore the question of whether SVD dimensionality reduction offers any improvement over a term co-occurrence representation for sentence semantics for measuring relevance and redundancy. While the experiments show that our system does not outperform a term $\times$ sentence *tf.idf* system, we have shown that the SVD reduced representation of a term co-occurrence space built from a large corpora performs better than the unreduced representation. This contra-

[8] Pairwise effect size estimates over datasets aren't sensible. Averaging of differences between pairs was affected by outliers, presumably caused by Rouge's error distribution.

dicts related work where SVD did not provide an improvement over unreduced representations on the name discrimination task (Pedersen et al., 2005). However, it is compatible with other work where SVD has been shown to help on the task of estimating human notions of word similarity (Matveeva et al., 2005; Rohde et al., In prep). A detailed analysis using the Friedman test and a cascade of Wilcoxon signed ranks tests suggest that our results are statistically valid despite the unreliability of the Rouge evaluation metric due to its low variance across systems.

Acknowledgements

This work was supported in part by Scottish Enterprise Edinburgh-Stanford Link grant R36410 and, as part of the EASIE project, grant R37588. It was also supported in part by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811, publication).

We would like to thank James Clarke for detailed comments and discussion. We would also like to thank the anonymous reviewers for their comments.

References

Enrique Amigo, Julio Gonzalo, Victor Peinado, Anselmo Penas, and Felisa Verdejo. 2004. An empirical study of information synthesis tasks. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, USA.

Romaric Besançon, Martin Rajman, and Jean-Cédric Chappelier. 1999. Textual similarities based on a distributional approach. In *Proceedings of the 10th International Workshop on Database And Expert Systems Applications*, Firenze, Italy.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 2003 Conference on Computational Natural Language Learning*, Edmonton, Canada.

Hoa T. Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, Vancouver, B.C., Canada.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, Jan.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25.

Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11:86–92.

Yihon Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, USA.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT—a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Ben Hachey and Claire Grover. 2004. A rhetorical status classifier for legal text summarisation. In *Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*, Barcelona, Spain.

Jagadeesh Jagarlamudi, Prasad Pingali, and Vasudeva Varma. 2005. A relevance-based language modeling approach to DUC 2005. In *Proceedings of the Document Understanding Conference*, Vancouver, B.C., Canada.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting*, Edmonton, Alberta, Canada.

Irina Matveeva, Gina-Anne Levow, Ayman Farahat, and Christiaan Royer. 2005. Term represetation with generalized latent semantic analysis. In *Proceedings of the 2005 Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Andrei Mikheev. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3).

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005a. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005b. Evaluating automatic summaries of meeting recordings. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, USA.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting*, New York City, NY, USA.

Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. 2005. Name discrimination by clustering similar contexts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

Douglas L. T. Rohde, Laur M. Gonnerman, and David C. Plaut. In prep. An improved method for deriving word meaning from lexical co-occurrence. `http://dlt4.mit.edu/~dr/COALS/Coals.pdf` (1 May 2006).

Josef Steinberger and Karel Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the 5th International Conference on Information Systems Implementation and Modelling*, Ostrava, Czech Republic.

Henry Thompson, Richard Tobin, David McKelvie, and Chris Brew. 1997. LT XML: Software API and toolkit for XML processing. *http://www.ltg.ed.ac.uk/software/*.