ISSN 1870-4069

# Corpus-based Automatic Text Expansion

Balaji Vasan Srinivasan[1], Rishiraj Saha Roy[2], Harsh Jhamtani[3],
Natwar Modani[1], Niyati Chhaya[1]

[1] Adobe Research Big Data Experience Lab,
Bangalore,
India

[2] Max Planck Institute for Informatics,
Saarland Informatics Campus,
Germany

[3] Carnegie Mellon University,
Language Technology Institute,
USA

{balsrini, nmodani, nchhaya}@adobe.com,
rishiraj@mpi-inf.mpg.de, jharsh@cs.cmu.edu

**Abstract.** The task of algorithmically expanding a textual content based on an existing corpus can aid in efficient authoring and is feasible if the desired additional materials are already present in the corpus. We propose an algorithm that automatically expands a piece of text, by identifying paragraphs from the repository as candidates for augmentation to the original content. The proposed method involves: extracting the keywords, searching the corpus, selecting and ranking relevant textual units while maintaining diversity in the overall information in the expanded content, and finally concatenating the selected text units. We propose metrics to evaluate the expanded content for diversity and relevance, and compare them against manual annotations. Results indicate viability of the proposed approach.

**Keywords:** Corpus, text expansion, automated text.

## 1 Introduction

While automated text summarization has been thoroughly researched over the last decade, the reverse task of "expanding" a piece of text has not been explored widely. Our work in this paper is motivated by the use case of automatically "resizing" textual content according to its delivery channel which can be assisted via algorithmic text expansion.

While channels like social media require content limited to a few characters, channels like websites, blogs or emails require an elaborate version of the same content. Content authors in an organization are under severe time pressure to deliver such modified versions along with numerous other stylistic personalizations of the same

piece of content. We believe that automating text expansion can be an important step towards accelerating the authoring workflow for textual content.

In this work, we propose algorithms that take a piece of textual content and expand it to a desired size by adding required content from a repository. The input is a short snippet composed by the author. A search query is constructed using the representative terms in the snippet, and is used to fetch relevant content from the corpus. We propose two algorithms to choose desired content from the retrieved list to produce the final expansion, and evaluate their performance on a real-world data set.

The rest of the paper is organized as follows. We differentiate existing work in this space from our current problem in Section 2. We introduce the proposed algorithm in Section 3 and evaluate its performance based on human annotations in Section 4. We propose metrics to evaluate the expanded content and correlate them with human annotations. Finally, we evaluate the expansion algorithm on a larger dataset to show the viability of the proposed algorithm in Section 5. Section 6 concludes the paper.

## 2 Related Work

Mihalcea et al. [5] identify key concepts in a document and link them to corresponding Wikipedia pages. While this helps to identify relevant Wikipedia areas for the document, this will not help in expanding the seed content from these input sources.

Li et al. [4] use a language model and a *topic development curve* to identify the most relevant text snippet in a document corresponding to a query. Snippets consist of phrases or sentences from multiple parts of the document, without sufficient context. While snippets may be useful in elegant presentation of a search engine results page, they are not suitable for text expansion.

Schlaefer et al. [8] aim to enhance question-answering by expanding a 'seed document' using web resources. The most relevant paragraphs (nuggets) are combined to provide the answers. While avoiding lexical redundancy, the authors retain semantic redundancy as it is desirable to enhance question-answering performance. This may however not be ideal for a content author as (s)he would want to avoid any type of content redundancy for human consumption, be it lexical or semantic. In the proposed algorithms, we address this by jointly optimizing for relevance and diversity of the expanded content.

Taneva and Weikum [9] identify relevant text snippets ('gems') by using an integer linear program to maximize relevance of selected words, and prefer the selection of contiguous words. However, such a method can result in only fragments of a sentence being selected, since the linear program is formulated at a word-level, thereby affecting readability. Biases may also be introduced in the expanded content which may not be preferable to an author.

There is a significant body of work in the domain of text summarization [7]. Text expansion could be perceived as a summarization task once we have identified the required candidate paragraphs from the repository. However, lack of notions of information coverage (maintaining the distribution of concepts between input and summary) in expansion makes it different from summarization.

## 3 Text Expansion

The primary requirements in expanding a content are that the resulting text should be relevant to what the author is building and also be diverse in the overall information present. We aim to achieve both these requirements with the proposed framework.

The input to our algorithm is a *content snippet* that the author is looking to expand, and the desired *length of the target* expansion (in words). The first step in our algorithm is extracting the top-$k$ keywords in the snippet using the inverse document frequency (IDF) of the words in the corpus, thus capturing the most significant keywords in the snippet with respect to the corpus. A query $q$ is then constructed by concatenating these $k$ keywords.

The choice of $k$ determines the relevance and the amount of content that is available and fetched from the repository. A lower value of $k$ results in the query under-representing the content and fetching articles that may not be very relevant. On the other hand, a higher value of $k$ will result in a very specific query that might not fetch many results from the repository.

We use $q$ to retrieve indexed content from the corpus. The retrieved content is split into paragraphs $\{P_1 \dots P_n\}$; which are the candidates for inclusion in the expanded content. Paragraphs are preferred over sentences because they are more likely to preserve (local) coherence in the final text. We assign every paragraph with a relevance score to the query based on a Lucene index. Often, paragraphs with high relevance scores contain significant information overlap (and hence redundancy). Therefore it is important to choose the relevant paragraphs but still account for the diversity in the overall material. We propose two approaches for this below.

### 3.1 Maximal Marginal Relevance (MMR)-based Ranking

Our first algorithm is inspired from **Maximum Marginal Relevance** (MMR) [1] for selecting the candidate paragraphs. MMR is used for obtaining diversified search result ranking often with multiple optimization objectives, e.g., relevance and diversity. At each iteration of the MMR algorithm, the best item is selected from the set of candidates by minimizing a cost function. For expansion, the cost function is formulated as:

$$\max_{P_i \in R \setminus S} \left[ (\lambda \times score_{rel}(q, P_i)) - ((1-\lambda) \times \max_{P_j \in S} (sim(P_i, P_j))) \right], \qquad (1)$$

where, $R$ is the set of all candidate paragraphs, $S$ is the subset of $R$ that is already selected for the expansion, $R \setminus S$ represents the set of unselected paragraphs so far, $score_{rel}$ is the relevance score of paragraph $P_i$ w.r.t query $q$, $sim(P_i, P_j)$ is the similarity (e.g., cosine similarity) between the vector representations of paragraphs $P_i$ and $P_j$ (reflecting the degree of content overlap), and $\lambda \in [0, 1]$ is a tunable parameter that reflects the trade-off between relevance and redundancy.

At each step, the paragraph that maximizes the above cost function is added to the expanded content $S$ and continued till the length of the expanded content reaches the desired limit, or the list of candidates is exhausted.

### 3.2 Graph-based Ranking

Our second algorithm is based on the graph-based ranking in [6]. We represent each paragraph $P_i$ as a node $v_i \in V$ in a weighted graph $\mathcal{G} = (V, E, W)$. We assign an initial "reward" $r_i^0$ for $P_i$ as the relevance of the paragraph $P_i$ to the query $q$. The cost $c_i$ of the paragraph $P_i$ is taken as the number of words in $P_i$.

An edge $e \in E$ between vertices $v_i$ and $v_j$ exists if there is a non-zero similarity between $P_i$ and $P_j$, and is weighted by a similarity function (again, like cosine similarity) $w_{ij}$ under a vector space representation. The gain $G_{v_i}$ of including a node $v_i$ in the expanded content at iteration $l$ is defined as its current discounted reward plus the weighted sum of the current discounted rewards of all immediate neighbors ($N_i$) of $v_i$ in $\mathcal{G}$, given by:

$$G_{v_i}^l = r_i^{l-1} + \sum_{v_j \in N_i} r_j^{l-1} \times w_{ij}. \quad (2)$$

At step $l$, we add $v_i^*$ with cost $c_i$ less than the remaining budget and maximum gain-to-cost ratio $G_{v_i}/c_i$ to our expansion. The rewards of the neighbor nodes $v_j$ of $v_i^*$ are then updated as $r_j^{l+1} = r_j^l \times (1 - w_{i^*j})$. This avoids inclusion of similar paragraphs thus ensuring diversity. We stop when there are no nodes left with cost lower than the available budget.

## 4 Experimental Evaluation

For our initial evaluation, we used a repository of 215 articles (indexed via Apache Lucene) from a proprietary forum including articles around key product features and troubleshooting instructions. We extracted 30 short text fragments and applied the proposed approaches to expand the original snippets using the repository. The input snippets had 33.9 words on an average, ranging from 4 to 86 words. The expansions were run with a target length of 500 words, with $k = 10$. The two methods generated a total of $30 \times 2 = 60$ expansions.

### 4.1 Human Evaluation

We obtained scores from 30 human annotators, each annotating 4 of the generated expansions, evaluating the *relevance* of the expanded content to the seed content and its content *diversity*, on a scale of $0 - 7$. We collected 120 annotations, each of the 60 expansions being rated twice while ensuring that the same annotator does not annotate the output from both algorithms. Fig. 1 plots the fraction of times ($y$) an expansion received a score of at least $x$ computed based on the cumulative distribution of the scores from the kernel density estimates.

The two approaches are comparable on relevance, as the same keyword extraction and search process applies for both of them. Diversity was observed to be better for MMR, possibly because it directly optimizes for low content-level overlap in its objective function via the choice of $\lambda$.

### 4.2 Automated Evaluation

While our results with human annotations are encouraging, evaluating the expansion performance on larger datasets requires a metric-based objective estimate of relevance and diversity that correlates well with human annotations. While metrics like KL-divergence [7] are widely used for measuring summarization quality, they cannot be applied as they are for evaluating expansion, because of the differences in the underlying tasks. We therefore propose two metrics to capture the degrees of relevance and diversity in the expanded content.

To measure the relevance of the expanded content to the input, we compute the similarity between each paragraph in the expanded content and the input. The average of maximum similarity of every paragraph in the expansion against all input text units can be computed as the relevance, but this will yield higher relevance even when the expanded text units match with very few input text units. On the other hand, taking an average will lead to a reduced relevance score unless it is relevant to *all input* text units. To address these issues, we use a *decayed weighted average* for computing the relevance:

$$\mathrm{rel}(c_{\mathrm{inp}}^{1\ldots N}, c_{\mathrm{exp}}^{1\ldots M}) = \frac{1}{M}\sum_{i=1}^{M}\frac{\sum_{k=1}^{TopK(c_{\mathrm{inp}},c_{\mathrm{exp}}^{i})}\gamma^{k}\,\mathrm{sim}(c_{\mathrm{exp}}^{i}, c_{\mathrm{inp}}^{k})}{\sum_{k=1}^{K}\gamma^{k}}, \tag{3}$$

where $c_{\mathrm{inp}}^{1\ldots N}$ are the text units in the input and $c_{\mathrm{exp}}^{1\ldots M}$ are the text units in the expanded content. $TopK(c_{\mathrm{inp}}, c_{\mathrm{exp}}^{i})$ returns the top-$K$ text units in the input content similar to $c_{\mathrm{exp}}^{i}$ (in the decreasing order of their similarly as computed by $sim(c_{\mathrm{exp}}^{i}, c_{\mathrm{inp}}^{k})$). The parameter $\gamma$, $(0 \leq \gamma \leq 1)$, penalizes the addition of a text unit that is similar to only a small set of the input text units via a decayed-weighted-average. The $sim()$ could be a standard similarity function between text units. We use cosine similarity here.

The Pearson Correlation Coefficient between the scores from Eq. 3 against the human evaluation is $0.7130$ indicating a strong correlation. We also compute the Wilcoxon-Mann-Whitney (WMW) statistic (extended from learning-to-rank problems [2]) between all the scores from the same annotator and the corresponding scores given by Eq. 3 was observed to be $0.9008$. The WMW statistic measures the probability that any pair of expanded content samples is ordered correctly based on Eq. 3 against the annotator's ranking.

To compute the diversity within the expanded content, we used a decayed-weighted average of similarity within the text units of the expanded content (similar to the relevance computation). A measure similar to Eq. 3 would give the information overlap (redundancy) within the expanded content, and the diversity is measured by modifying it as:

$$\mathrm{div}(c_{\mathrm{exp}}^{1\ldots M}) = 1 - \frac{1}{M}\sum_{i=1}^{M}\frac{\sum_{k=1}^{TopK(c_{\mathrm{exp}},c_{\mathrm{exp}}^{i})}\gamma^{k}\,\mathrm{sim}(c_{\mathrm{exp}}^{i}, c_{\mathrm{exp}}^{k})}{\sum_{k=1}^{K}\gamma^{k}}. \tag{4}$$

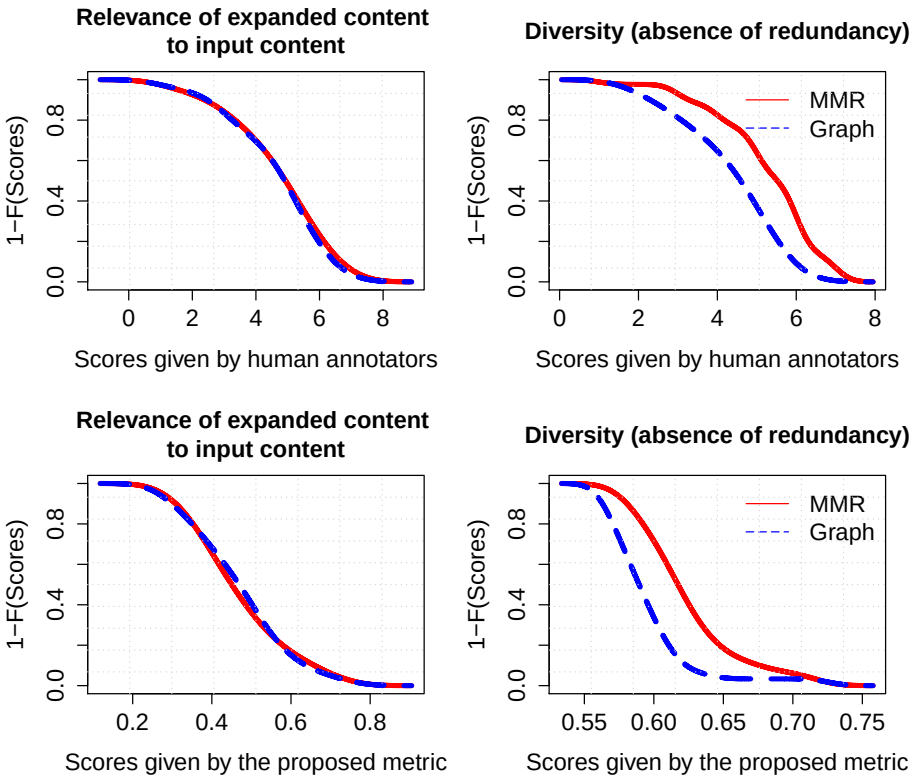


**Fig. 1.** Evaluation of the two proposed expansion approaches based on annotations of 60 different expansions, from 30 input fragments and two approaches. Each expanded content was annotated by 2 independent annotators. Every point *(x,y)* on the plotted graph indicates the percentage *y* of expanded content that was rated at least *x* by the user. *F* is the cumulative distribution computed using kernel density estimates of the score distribution.

Note that the arguments to the $TopK$ function are different in Eqs. 3 and 4. Eq. 4 captures the similarity within the expanded content and uses it for the diversity computation.

The Pearson Correlation Coefficient between Eq. 4 and the human evaluation is $0.3730$, indicating a moderate correlation. The WMW statistic was $0.7795$ indicating a good agreement with the human annotators.

Fig. 1 plots the cumulative distribution of relevance and diversity (Eqs. 3 and 4) similar to Fig. 1. Similarity of the two distributions further established a strong correlation between proposed metrics and the human annotations. Note that a relevance score reaches a maximum of $0.75$ with a median around $0.5$. The cumulative distribution of the diversity is also very similar to that from the annotations, with some deviations as indicated by a lower correlation coefficient. We note that the diversity score reaches a maximum value of $0.7$ with a median of around $0.6$.

**Table 1.** A sample input from the Australian legal dataset and the expanded content from the two proposed algorithms (with a desired size of 500).

| **Input Content Snippet** | **Expanded with the MMR-based approach in Sec. 3.1** | **Expanded with the Graph-based approach in Sec. 3.2** |
|---|---|---|
| Damages claimed from respondents for breach of guarantee of profit shortfall. First respondent had ostensible authority to bind second respondent to oral variation. Profit shortfall amount for 1998 contracts evidence agency. | However it became clear during cross-examination of Mr Forbes and Mr Brauer that the sales which the respondents claimed should have been credited to the 1998 year actually took place in 1997, and were properly accounted as 1997 sales, as claimed by the applicants.<br>In summary, the respondents claimed that these documents were critical to properly investigating: It was not in contention between the parties that the source financial documents were missing and unavailable.<br>Did Forbes Australia experience a profit shortfall in the financial year ending 31 December 1998? | Did Forbes Australia experience a profit shortfall in the financial year ending 31 December 1998?<br>The material is relevant to both the applicants' claims concerning the 1998 profit shortfall and the respondents' defence.<br>2. a claim for $1,691,284 which is alleged to be the profit shortfall in respect of the 1999 calendar year.<br>It also follows that the thirty-eighth and thirty-ninth respondents should recover judgment for breach of duty.<br>That shortfall was claimed in the amount of $71,663.65. |

## 5 Experiments on a Public Dataset

Finally, we evaluate our proposed approaches on the Australian Legal Case Reports dataset[1], a collection of 3890 legal cases from the Federal Court of Australia (FCA) from 2006 to 2009. The dataset includes a gold standard summary for every case in the form of 'catchphrases' and 'key sentences' [3]. The legal articles were 6406-word long on an average, while the summaries were 65-word long on an average.

We used the entire set of cases as the content repository and used the gold standard summaries as input to our expansion algorithms. Note that our objective is not to reconstruct the original content from the summary, but rather to test the quality of expansion across several pieces of expanded content.

Table 1 shows a sample input and the corresponding output expanded by the two proposed algorithms. Fig. 2 plots the relevance and diversity of the expanded content for various output sizes based on Eqs. 3 and 4. Fig. 2 also shows the similarity of the expanded content to the original content based on Eq. 3 across the two approaches.

The medians of the relevance and diversity across all the runs are approximately 0.5 and 0.65 respectively, similar to the distribution obtained for the annotated dataset. This indicates a similar quality in the expanded content for the two datasets that we have experimented with. Fig. 2 indicates that the MMR-based algorithm performs marginally better that the graph-based one in terms of relevance and diversity for small expansion sizes.

[1] http://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports, Accessed 16 March 2017.

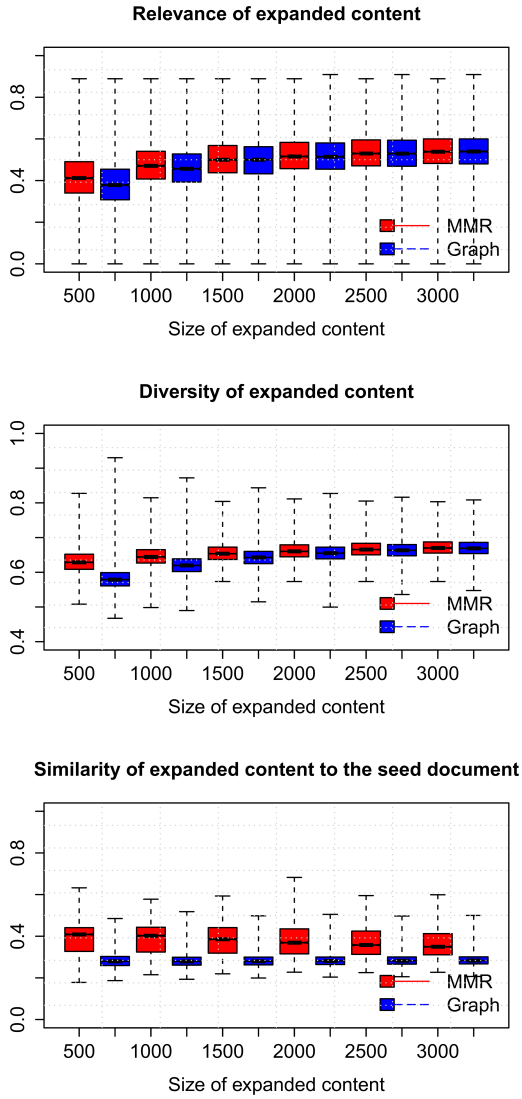


**Fig. 2.** Relevance and diversity of the expanded content across 3886 samples from the Australian legal dataset for various target sizes.

However, for larger expansion, the relevance and diversity of both MMR and graph-based expansions become comparable. The relevance and diversity that the addition of a new paragraph brings to the expansion, perhaps becomes very low beyond a certain output size leading to eventual saturation of these scores, as seen in Fig. 2.

The proposed approach does not aim at reconstructing the original content whose summary was used for the expansion. However a certain degree of similarity to the original content is desirable from an authoring perspective.

We therefore compute the similarity of the expanded content with the original content using Eq. 3 with $K = 1$. The similarity is higher for the MMR-based approach than the graph-based expansion. For both the approaches, the similarity marginally decreases for higher expansion sizes perhaps because of the algorithms' quest for content diversity.

## 6 Conclusions and Future Work

We studied the problem of expanding a piece of text by reusing content from an existing corpus and proposed two alternative approaches within the same framework. We also proposed metrics to evaluate the relevance and diversity of the expanded content which was shown to correlate well with human annotations.

Results show that automated expansion is indeed feasible and is a promising direction of research. Incorporating coherence of the expanded material appears to be the most promising future direction. We believe that automated text expansion will play a key role in smart authoring workflows for several domains in the near future.

## References

1. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335–336 (1998)
2. Fung, G., Rosales, R., Krishnapuram, B.: Learning rankings via convex hull separation. Advances in Neural Information Processing Systems, vol. 18 (2005)
3. Galgani, F., Compton, P., Hoffmann, A.: Combining different summarization techniques for legal text. In: Proceedings of the workshop on innovative hybrid approaches to the processing of textual data, pp. 115–123 (2012)
4. Li, Q., Candan, K. S., Qi, Y.: Extracting relevant snippets fromweb documents through language model based text segmentation. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 287–290 (2007) doi: 10.1109/WI.2007.115
5. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM Conference on Information and Knowledge Management, pp. 233–242 (2007) doi: 10.1145/1321440.1321475
6. Modani, N., Khabiri, E., Srinivasan, H., Caverlee, J.: Creating diverse product review summaries: A graph approach. In: International Conference on Web Information Systems Engineering, vol. 9418, pp. 169–184 (2015), doi: 10.1007/978-3-319-26190-4_12
7. Nenkova, A., McKeown, K.: Automatic summarization. Foundations and Trends in Information Retrieval, vol. 5, no. 2–3, pp. 103–123 (2011) doi: 10.1561/1500000015
8. Schlaefer, N., Chu-Carroll, J., Nyberg, E., Fan, J., Zadrozny, W., Ferrucci, D.: Statistical source expansion for question answering. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 345–354 (2011) doi: 10.1145/2063576.2063632
9. Taneva, B., Weikum, G.: Gem-based entity-knowledge maintenance. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 149–158 (2013) doi: 10.1145/2505515.2505715