Extractive Summarization under Strict Length Constraints

Yashar Mehdad[1†] **Kapil Thadani**[1†] **Dragomir Radev**[2] **Amanda Stent**[1] **Youssef Billawala**[1] **Karolina Buchner**[3]
[1] Yahoo Research [2] University of Michigan [3] Apple
{ymehdad, thadani}@yahoo-inc.com, radev@umich.edu, {stent, billawala}@yahoo-inc.com, kbbuchner@gmail.com

Abstract

In this paper we report a comparison of various techniques for single-document extractive summarization under strict length budgets, which is a common commercial use case (e.g. summarization of news articles by news aggregators). We show that, evaluated using ROUGE, numerous algorithms from the literature fail to beat a simple lead-based baseline for this task. However, a supervised approach with lightweight and efficient features improves over the lead-based baseline. Additional human evaluation demonstrates that the supervised approach also performs competitively with a commercial system that uses more sophisticated features.

Keywords: text summarization, extractive summarization

1. Introduction

Extractive text-to-text summarization has a rich history (e.g. (Carbonell and Goldstein, 1998; Erkan and Radev, 2004)). In this task, sentences or text snippets are extracted from an input document or set of documents to produce a summary. However, in previous extractive summarization shared tasks such as the DUC shared tasks (`http://duc.nist.gov/`) length restrictions on output summaries were quite generous (100 to 400 words, or 400+ characters) as opposed to the very strict length budgets required by current commercial use cases (160 to 300 characters for search result and news article summarization, especially on mobile devices).

In this work, we focus on a supervised method to generate extractive summaries under strict length constraints without sacrificing meaning or grammaticality. We present evaluation results for the single-document news summarization use case, comparing performance on well written articles as well as on a sampling of news articles of random quality.

2. Related Work

Research on single-document, extractive summarization has been conducted since the 1950s (Luhn, 1958). Traditionally, extractive single document summarization has focused on scoring, ranking, and extracting the most "informative" sentences from a document using various supervised (e.g. (Conroy et al., 2004; Daumé and Marcu, 2006; Lin, 1999; Svore et al., 2007)) and unsupervised (e.g. (Erkan and Radev, 2004; Mei et al., 2010; Mihalcea and Radev, 2011)) methods. Innovations fall into two broad categories: (a) finding ways to assess whether a sentence should be included in a summary; and (b) efficient algorithms for exploring the space of possible summaries.

A recent study compared a number of extractive summarization algorithms (Hong et al., 2014). The best performing algorithm performed global optimization over the input sentence set. However, these algorithms were compared using the DUC 2004 task, (a) which is a multi-document summarization task; and (b) for which the reference summaries were abstractive.

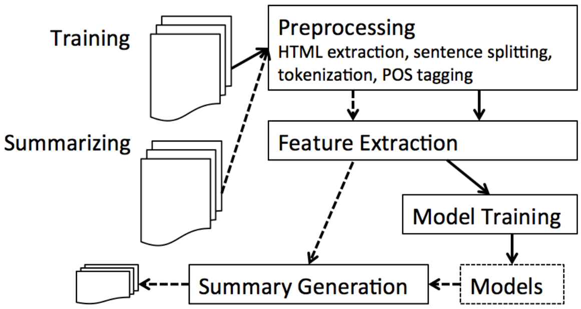

Figure 1: Our summarization system

More related to our framework, snippet extraction is a popular approach for search engines and news aggregators to show some content related to the query and the original document (Li et al., 2008). A simple way to identify snippets is to extract a passage from specific areas of the original page where the important information is found, relying on structural markup to identify such information (Callan, 1994). Although this approach is simple and scalable, document style and structural differences when changing domains or publishers can significantly affect snippet quality.

3. System

Our system takes as input an HTML document. We automatically extract the article text from the HTML, and then automatically preprocess the text to obtain sentences, tokens and part of speech tags. Then, we compute various features over the preprocessed document. Each sentence is scored using a combination of feature values and feature weights, which are learned using a structured perceptron (Collins, 2002). Finally, sentences are extracted in a greedy fashion based on their scores while respecting the length constraint. These steps are illustrated in Figure 2.

3.1. Features

We implemented various features drawn from the summarization literature that capture aspects of salience, diversity, coverage, content and readability. Table 1 presents features from each category implemented in our framework.

[†] Equal contribution.

Category	Examples
Position	sentence position, paragraph position, in-paragraph position
Length	word length, character length, summary length
Content	similarity to query / headline / title/ document / summary
Lexical cues	containing url / special pattern / quote / question / capitalization / number / money
Syntactic cues	containing pronouns/ nouns/ certain part of speech

Table 1: Feature categories implemented in our summarization framework

Position features indicate the inverse position of each sentence and paragraph in the document, and of each sentence in its containing paragraph. **Length** features indicate the length of each sentence, and of the summary so far, in words and characters. **Content** features indicate similarity of each sentence to the input query, headline or title as well as to the document and to the summary so far, and are computed as cosine similarity over term frequency vectors. **Lexical cue** features are binary features indicating whether a sentence contains various lexical items such as URLs, quotations, or numbers. Finally, **Syntactic cue** features are binary and normalized count features indicating whether a sentence contains a syntactic phenomenon such as a pronoun, and if so, how many. Although we experimented with all these features, the best performing model for sentence scoring contained only the very efficient and easy to compute position, length and content features; the lexical and syntactic cue features were used in a preprocessor that prunes sentences that should not be in a summary (e.g. a sentence that contains a URL).

3.2. Supervised Learning

An input document D is represented as a set $\{x_1, \ldots, x_n\}$ of n sentences. An extractive summary S is composed of sentences from this document, i.e., $S \subseteq D$. For any document, we seek to recover the highest-scoring summary $\widehat{S}$ which can fits within a pre-determined budget b.

$$\widehat{S} = \underset{S \subseteq D}{\arg\max}\ score(S, D) \quad \text{s.t. } cost(S) < b \tag{1}$$

The tractability of this inference formulation depends on the factorization of the function $score$ over the summary. In this work, we forego exact solutions to (1) in order to accommodate richer scoring functions that can model phenomena such as diversity (Carbonell and Goldstein, 1998) and coherence (Barzilay and Lapata, 2005; Christensen et al., 2013). Summaries are scored using a linear model

$$score(S, D) = \mathbf{w}^{\top}\mathbf{\Phi}(S, D) \tag{2}$$

where $\mathbf{\Phi}$ is a feature map for summaries and $\mathbf{w}$ is a vector of learned parameters.

In order to train the parameters $\mathbf{w}$, we assume the existence of a training dataset $\mathcal{D}$ comprised of instances $\langle D, S^* \rangle$ where S^* represents a reference summary for document D. The parameters are estimated using the structured perceptron (Collins, 2002) which minimizes a 0/1 loss over $\mathcal{D}$ and incorporates parameter averaging for generalization. The basic learning procedure is described in Algorithm 1. When inference is inexact and carried out via search—as in the case of our framework—convergence and performance can be improved using *violation-fixing* weight updates (Huang and Feyong, 2012). In addition to greedy search, we also experimented with beams of various sizes to reduce search errors but did not observe performance improvements[1].

Algorithm 1 Structured perceptron (Collins, 2002)

```
Input: training dataset $\mathcal{D}$, feature map $\mathbf{\Phi}$, learning rate $\eta$
Output: vector of learned parameters $\mathbf{w}$
$\mathbf{w}_0 \leftarrow \mathbb{0}^{|\mathbf{\Phi}|}$
$k \leftarrow 0$
while not converged do
    for instance $\langle D, S^* \rangle \in \mathcal{D}$ do
        $\widehat{S} \leftarrow \arg\max_S\ \mathbf{w}_k^{\top}\mathbf{\Phi}(D, S)$
        if $\widehat{S} \neq S^*$ then
            $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \eta\left(\mathbf{\Phi}(D, S^*) - \mathbf{\Phi}(D, \widehat{S})\right)$
            $k \leftarrow k + 1$
return average weights $\frac{1}{k}\sum_j \mathbf{w}_j$
```

4. Experiments - Systems Compared

4.1. Baselines and Other Methods

We compare the extractive summaries generated by our framework to three simple baselines (*Lead-based* and two variations of *Greedy*) as well as to an array of standard summarization methods from the literature.

Lead-based The baseline, lead-based, algorithm takes the first sentences in the document that fit within the budget.

Greedy The greedy algorithm is exactly the one described in (McDonald, 2007): sentences are ordered by a sentence scoring function, and then added one by one to the summary until the budget is full. The sentence scoring function may be entirely precomputable (incorporating assessments of relevance of sentence to the document, independence of the sentence, etc.) or may also incorporate aspects that are computed on the fly (e.g. redundancy of sentence with respect to the summary).

In addition to the standard greedy search algorithm, we implemented a version in which the first sentence in the document must be included in the summary (*S1 + Greedy*).

MEAD (Radev et al., 2004) provides a sentence classifier. Sentence features include similarity to centroid, length, and position. The default reranker orders the sentences and iteratively decides whether to add each sentence to the summary, based on its similarity to previously included sentences. At each step, if the quota of words or sentences has not been filled, and the sentence is not too

[1] This phenomenon has also been reported in prior work (McDonald, 2007).

similar to any sentence already in the summary, the sentence in question is added to the summary. After the allocated length has been reached, the reranker increases the scores of the selected sentences and discounts the disqualified (by similarity) or unselected sentences.

In addition to standard MEAD, we implemented a variation in which the first sentence in the document is required to be in the summary (*S1 + MEAD*).

Divrank (Mei et al., 2010) is based on a reinforced random walk over a lexical similarity graph. This model automatically balances the prestige and the diversity of the top ranked vertices in a principled way.

Lexrank (Erkan and Radev, 2004) is based on eigenvector similarity over a lexical similarity graph. The nodes of the graph correspond to input sentences and the edges to weighted cosine similarity. The most central sentences are selected for the summary.

MaxCov (Takamura and Okumura, 2009) applies approximation algorithms for the budgeted maximum coverage problem (Khuller et al., 1999) to document summarization. It assumes the existence of a vocabulary in which each word is associated with some positive profit (word score in the summary). Given a collection of subsets of this vocabulary (sentences), each associated with some cost (number of characters), the budgeted maximum coverage problem identifies a summary whose total cost remains within the budget and whose union maximizes the summary score.

Personalized PageRank (PPR) (Agirre and Soroa, 2009) is a variation of the pagerank algorithm where the ranking process is performed depending on the relationships among all sentences in the document.

Knapsack We use the algorithm described in (McDonald, 2007), except that the scoring function may be varied, as described above in the synopsis of the greedy algorithm.

Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) is a variant of greedy summarization which encourages the sentences in the summary to be diverse and non-redundant. As new sentences are considered for the summary, the greedy objective balances the relevance of any new sentence with its similarity to each sentence already in the summary.

4.2. Data

We collected 172 highly popular articles from the Yahoo home page: (a) Yahoo-produced content; (b) content from Yahoo partners; and (c) news articles of any type, covering a wide range of journalistic quality (“random”). We obtained gold-standard extractive 300-character reference summaries for each document from one to four (usually two) professional annotators. This makes our data and reference summaries much more similar to the commercial use case than other publicly available data sets such as DUC.

For training the model for our supervised framework, we used a subset of the New York Times Annotated Corpus containing 1800 articles with extractive summaries of length less than 300 characters. The model contains “lightweight” features that require minimal preprocessing, including positional features (sentence position in the document and the paragraph), relevance features (relevance of the sentence to the document, computed using cosine similarity over term frequency vectors) and a binary feature—computed on the fly—to identify when a candidate sentence is adjacent to a sentence already present in the summary. The scoring function we used for the greedy and knapsack algorithms uses the same features.

Summarizer	**R-1**	**R-2**	**R-4**
Greedy	0.51	0.34	0.29
S1 + Greedy	0.58	0.45	0.40
Lead-based	0.66	0.59	0.55
PPR (α=0.5, simceil=0.15)	0.27	0.1	0.06
MaxCov	0.43	0.25	0.20
Divrank	0.46	0.33	0.28
Lexrank	0.50	0.38	0.34
MEAD	0.51	0.39	0.35
S1 + MEAD	0.51	0.40	0.36
Knapsack	0.57	0.43	0.38
MMR (λ=0.8)	0.61	0.49	0.44
Ours	**0.71**	**0.61**	**0.56**

Table 2: Results of automated evaluation under ROUGE

4.3. Automatic Evaluation

For evaluation, we produced 300-character extractive summaries from the input documents by passing them through the systems described earlier.

We evaluate the performance of all systems against manually creative extractive reference summaries. We use ROUGE (Lin, 2004), a well-established automatic evaluation metric based on lexical overlap which has been widely used in the scientific community and has been shown to correlate well with human evaluations. We follow the standards suggested by Owczarzak et al. (2012). As is standard, we report ROUGE R-1, R-2 and R-4.

Table 2 shows evaluation results. The lead-based baseline outperforms all the methods from the literature that we included. However, our framework outperforms this baseline. Analyzing the reference summaries, we observed that they are primarily lead-based unless one of the first sentences in the input document is a result of errors in article extraction from HTML (e.g. a byline is extracted as part of the article, or a photo caption is included as part of the article) or a repetition of the article title.

5. Experiments - Side by Side Editorial Evaluation

We also conducted a manual, side-by-side evaluation comparing the summaries produced by our system to those produced by a commercial summarizer.

5.1. Dataset

We collected a sample set of popular articles from the Yahoo home page covering the period 04/20–04/24 2015. These articles fall into three categories:

Popular: 120 most viewed documents from 04/20–04/25.

Random Partner: a subsample of 139 random scraped ar-

Title	Randy Travis Surprises, Moves Spectators with First Awards Show Appearance
Our system	Traviss moment on camera was brief and slightly shaky, but showcased the fact that he is on the way to full recovery. Travis underwent an acrimonious divorce with his first wife, Lib Hatcher, in 2010, which appeared to set off a chain of troublesome events to follow.
Commercial system	When country star Lee Brice took the stage at the 50th annual Academy of Country Music Awards to announce the Song of the Year prize, spectators were delighted by his unexpected stanza of Randy Travis's 1987 hit "Forever And Ever, Amen." They were even more excited, however, when...
Title	Source: Bruce Jenner Began Transgender Journey in 1980s, Then Stopped After Meeting Kris Jenner
Our system	A source with direct knowledge told ET that Jenner talked about transitioning 30 years ago, and that by the mid-1980s Jenner had started hormone therapy, electrolysis and had plastic surgery to make his features look more feminine. Bruce Jenner's Emotional New Promo: 'How Does My Story End?'
Commercial system	ET has exclusive new information about when Bruce Jenner began his gender transition. A source with direct knowledge told ET that Jenner talked about transitioning 30 years ago, and that by the mid-1980s Jenner had started hormone therapy, electrolysis and had plastic surgery to make his features...
Title	Which passports are the best (and worst) to have?
Our system	Arton, an advisory firm that helps people take part in citizenship and residency programs for investors, has created the Passport Index, which ranks travel documents based on how many countries their holders can visit without having to obtain a visa in advance.
Commercial system	How "powerful" is your passport? This is the question that the folks at Arton Capital have set out to answer. The more nations you can access merely by showing up and getting a landing visa at the airport (or even entering visa-free), the more powerful your passport is.

Table 6: Sample summaries from our system and a commercial system

Summary	#	%	Equally bad	Equally good
Ours much better	14	12	5 (4%)	26 (22%)
Ours better	65	54		
No preference	31	26		
Other better	8	7		
Other much better	2	2		

Table 3: Results on *Popular* articles

Summary	#	%	Equally bad	Equally good
Ours much better	18	15	11 (8%)	24 (17%)
Ours better	41	24		
No preference	36	30		
Other better	26	22		
Other much better	18	15		

Table 4: Results on *Random Partner* articles

Summary	#	%	Equally bad	Equally good
Ours much better	2	1	43 (31%)	16 (12%)
Ours better	22	16		
No preference	62	45		
Other better	50	36		
Other much better	3	2		

Table 5: Results on *Random Other* articles

ticles from Yahoo and its publishing partners from the same time period.

Random Other: a subsample of 139 random articles from any article publisher from the same time period.

300 character summaries were automatically produced for all of these articles using both our framework and a commercial system that uses more sophisticated features based on entity mentions in the input document.

For each document, we presented professional annotators with the two summaries produced by these two methods. Order of presentation of the summaries was randomized. We asked the annotators to choose if they "highly prefer" or "prefer" one summary over another, or they had no preference. We also asked them to justify their answer.

5.2. Results

Based on the manual evaluation results (Tables 3, 4 and 5), our framework outperforms the commercial system for the *Popular* and *Random Partner* datasets. However, the same trend was not reflected for the *Random Other* dataset. Annotators complained of unusual character and formatting issues in our output *Random Other* documents; the commercial system has better handling of unicode symbols and other character formatting in its output.

Table 6 shows example outputs. In general, we observed that our summaries are less redundant with article titles (examples 1 and 3); however, our summaries tend to contain extraneous short sentences to fill up the budget (example 2). By contrast, the commercial system quite often simply produces a lead-based summary (examples 1–3). It also frequently exceeds the character budget, leading to truncated sentences in the output (examples 1 and 2). The performance of both summarizers was upper-bounded by errors due to incorrect article extraction from HTML, including incorrect inclusion of photo captions, ads for other articles (example 2), etc.

6. Conclusions

We presented a comparative study on supervised and unsupervised approaches to extractive summarization under strict length considerations. We compared our supervised summarization framework, which implements a collection of light-weight features, with various unsupervised and supervised baselines using automatic evaluation. We also manually evaluated the summaries generated by our system against those produced by a commercial summarizer. The evaluation results demonstrate the effectiveness of our supervised system for extractive summarization under strict length considerations.

7. Bibliographical References

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the EACL*.

Barzilay, R. and Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the ACL*.

Callan, J. (1994). Passage-level evidence in document retrieval. In *Proceedings of SIGIR*.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*.

Christensen, J., Mausam, Soderland, S., and Etzioni, O. (2013). Towards coherent multi-document summarization. In *Proceedings of NAACL-HLT*.

Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Conroy, J. M., Goldstein, J., Schlesinger, J. D., and O'Leary, D. P. (2004). Left-brain/right-brain multi-document summarization. In *Proceedings of the DUC*.

Daumé, III, H. and Marcu, D. (2006). Bayesian query-focused summarization. In *Proceedings of the ACL*.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Hong, K., Conroy, J. M., Favre, B., Kulesza, A., Lin, H., and Nenkova, A. (2014). A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of LREC*.

Huang, L. and Feyong, S. (2012). Structured perceptron with inexact search. In *Proceedings of HLT-NAACL*.

Khuller, S., Moss, A., and Naor, J. S. (1999). The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45.

Li, Q., Candan, K. S., and Yan, Q. (2008). Extracting relevant snippets for web navigation. In *Proceedings of AAAI*.

Lin, C.-Y. (1999). Training a selection function for extraction. In *Proceedings of CIKM*.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165.

McDonald, R. (2007). A study of global inference algorithms in multi-document summarization. In *Proceedings of ECIR*.

Mei, Q., Guo, J., and Radev, D. R. (2010). DivRank: the interplay of prestige and diversity in information networks. In *Proceedings of KDD*.

Mihalcea, R. and Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press, Cambridge; New York.

Owczarzak, K., Conroy, J. M., Dang, H. T., and Nenkova, A. (2012). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*.

Radev, D. R., Jing, H., Styś, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938, December.

Svore, K., Vanderwende, L., and Burges, C. (2007). Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of EMNLP-CONLL*.

Takamura, H. and Okumura, M. (2009). Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the EACL*.