Using Semantic and Context Features for Answer Summary Extraction

Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, Mark Sanderson
RMIT University, Melbourne, Australia
{evi.yulianti, ruey-cheng.chen, falk.scholer, mark.sanderson}@rmit.edu.au

ABSTRACT

We investigate the effectiveness of using semantic and context features for extracting document summaries that are designed to contain answers for non-factoid queries. The summarization methods are compared against state-of-the-art factoid question answering and query-biased summarization techniques. The accuracy of generated answer summaries are evaluated using ROUGE as well as sentence ranking measures, and the relationship between these measures are further analyzed. The results show that semantic and context features give significant improvement to the state-of-the-art techniques.

Keywords

summarization; answer summaries; non-factoid queries

1. INTRODUCTION

The summaries displayed in a search result page, or *snippets*, are classically biased towards the matching texts, so users can better decide which documents are worth examining. More recently, commericial search engines have begun to utilize such summaries to answer the user's question, thereby alleviating the need to read the underlying documents. This approach could potentially lead to good abandonment [6] where people leave the result page without any click interaction, but having gained the information they sought [2]. Web search on mobile devices with low network bandwidth or limited screen size might benefit the most from this research.

A challenge to extract answer summaries is the lexical gap between the query and the sentences containing answers in the document. The answer sentences may share many different vocabularies with the queries, therefore relying only on topical relevance has been shown to be ineffective for finding answers [4]. Here, we investigate semantic and context features [13] to bridge the gap using three learning models. We first examine the effectiveness of a state of-the-art method in factoid question answering. Next, we evaluate our method against a state-of-the-art query-biased summarization technique, using term overlap based measures (i.e. ROUGE) and, finally, sentence ranking measures. The relationship between the measures are also analyzed to better understand their agreement in this task.

ADCS '16, December 05-07, 2016, Caulfield, VIC, Australia

DOI: http://dx.doi.org/10.1145/3015022.3015031

2. RELATED WORK

Current research on document summarization aims to generate relevant summaries that represent the main topic of a document. They could be generic to document content, or biased to some information such as queries [8] or linked social media [15]. Less attention has been paid to extracting single document summaries that are designed to contain answers to the query.

Extracting a short fact as an answer to a factoid question has been the focus of Question Answering (QA) research. There are many questions however that could not be answered by a short fact. For example, the question defined in description queries of TREC Terabyte track – *"What are some of the possible complications and potential dangers of gastric bypass surgery?"* – are best answered with a longer multi sentence summary. Some past work has considered passage extraction from documents to serve as an answer, such as using statistical translation [10], query likelihood passage retrieval [4], and a paid crowdsourcing method [2]. However, none of these efforts have used an automatic summarization approach to extract answers from documents. The research on answer retrieval and answer ranking in Community question-answering (CQA) [12, 11] is also relevant, but without needing to extract or synthesize answers from the underlying collections. In contrast to the prior art, our work focuses on extracting answer summaries directly from retrieved documents.

Our research is closely related to answer sentence retrieval or selection [9, 13]. Severyn and Moschitti [9] used a deep learning method to model sentence-to-sentence similarity on factoid data, while Yang et al. [13] focused on non-factoid questions, utilizing a learning-to-rank technique to tackle the problem of sentence selection. None of these results, however, were evaluated in terms of answer quality. Our work is different from Yang et al. since we aim to generate answer summaries from each retrieved document, instead of displaying a ranked list of sentences retrieved from a set of documents in the collection. This approach, we argue, may produce more readable answers as same-document sentences usually make more coherent summaries.

3. DATASET

We use the WebAP[1] dataset because it contains all the elements that we need to evaluate our methods. The dataset consists of 82 description queries, each with a corresponding top 50 documents, and annotated answer passages to serve as ground truth answers. Keikha et al.[4] built this dataset based on the GOV2 collection and the description queries from the TREC Terabyte track. The authors initially identified the queries that can be answered by just a passage. For each of these queries, they selected relevant documents from the retrieved fifty top ranked documents, and then annotated answer passages from such documents using four relevance levels: perfect, excellent, good, and fair.

Ground truth answers were created by drawing from high-quality passages such as those labeled as perfect or excellent. Passages from lower relevance level might include partial answers or marginally relevant texts, which do not fulfil our standard of answer quality. There are 80 out of 82 queries that have answers of the required quality level. The evaluation was focused only on relevant documents, as irrelevant documents do not contain any answers. Our final dataset consists of 80 queries, 1436 documents, and 3298 ground truth answers. The average number of sentences per document is 255.63 and the average number of sentences in our ground truth answers is 2.67. On average, 93.2% of the sentences in a document are irrelevant; and only 3.4% and 2.7% are part of perfect and excellent answers.

4. METHODS

We describe the means of summarizing and evaluating.

4.1 Summarization Method

We use a learning-to-rank approach to sort sentences in a document and then take top ranked sentences as an answer summary. The summary length is set to the nearest integer length of the ground truth answers (i.e. 2.67: three sentences). A combination of query-biased [8], semantic, and context features [13] were used to identify sentences that contain answers from each document.

The features were extracted for each sentence in the document, after stopping using INQUERY list and performing Krovetz stemming. The full list of features is given in Table 1. The first group (MK) is derived from the work of Metzler and Kanungo [8] on sentence extraction for query-biased summarization. The MK features cover basic lexical and synonym matching techniques such as Term Overlap, Synonym Overlap, and Language Model Score. The features were a common baseline in query-biased summarization experiments [1]. The second group (Semantic) are built on top of three semantic representations of texts: Explicit Semantic Analysis (ESA), word vector representation by using Word2Vec[2], and entities linked using TAGME [13]. Each of these features represents a different class of approach for estimating query-sentence semantic similarity, and they were shown to complement the MK features in previous work [13]. The third group (Context) focus on characterizing neighboring sentences. Yang et al. [13] proposed using meta-features to catch nearby answer-bearing signals. Two meta-features are used to combine existing MK and semantic features in the sentences immediately preceding or following the current sentence.

In summary, each sentence in the document has 27 features in total, consisting of six MK, three Sem, and eighteen Con features. Three learning models are employed in this work, following previous work [13]: CA (Coordinate Ascent), MART (Multiple Additive Regression Trees), and LambdaMART. When learning and testing the model, sentence annotations in the dataset are mapped into numerical grades of relevance: None=0, Fair=1, Good=2, Excellent=3, and Perfect=4.

4.2 Evaluation Method

The accuracy of answer summaries is evaluated using a term overlap based measure, ROUGE [7], that has been commonly used in previous work on summarization. We report the score of ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-SU4 (overlap of word pairs with maximum skip-length of 4 plus unigrams). The ROUGE scores are computed by comparing the generated summaries with the available ground truth answers. We use the maximum value of the generated summary and each of the ground truth answers, following the recommendation of Keikha et al. [3] on evaluating answer passages using ROUGE measures. In addition to ROUGE, we also evaluate the accuracy of summaries based on the effectiveness of sentence rankers, as performed in previous work [8, 5]. Here, NDCG@k and P@k are adopted, where k is set to three following the number of sentences in our summaries. As the number of documents in each query is different, we calculate macro-averaged scores for each of our measures.

5. EXPERIMENTS

The features in our summary are extracted by using the SummaryRank[3] package, and learning-to-rank models are built using the RankLib[4] tool. The hyperparameters for each learning algorithm are set up the same way as in Yang et al. [13]. To generate high-quality answer summaries, we focus on NDCG@3 in the training rather than other more recall-oriented variants of NDCG and ERR. A ten-fold cross validation is conducted for each combination of feature set and learning algorithm.

5.1 Factoid QA Method

To examine the effectiveness of factoid QA methods on extracting non-factoid answer summaries, we first make a comparison with a recent deep-learning-based approach by Severyn and Moschitti [9], which was shown on the TREC QA dataset to outperform several common baselines including the Convolutional Neural Network (CNN) approach [14]. Severyn and Moschitti also make use of CNNs but do so in a renovated architecture that incorporates rich ("external") features such as term overlap and term counts into the decision process. Our replicated result faithfully follows the best setting reported in the original paper, with the width of convolutional filter set to 5, the number of feature maps set to 100, and batch size set to 50 examples [9]. The limitation of CNNs forced us to truncate sentences longer than some predefined threshold, which in our experiment was set to 60 words to minimize the negative impact. 8084 sentences

[1]https://ciir.cs.umass.edu/downloads/WebAP/
[2]https://code.google.com/p/word2vec/
[3]http://rmit-ir.github.io/SummaryRank
[4]http://www.lemurproject.org/ranklib.php

Table 1: List of features

MK	**Exact Match**	Binary value indicating the query being of substring in the sentence
	Term Overlap	Fraction of query terms that occur in the sentence
	Synonym Overlap	Fraction of query terms as well as their synonyms that occur in the sentence
	Language Model Score	Log-likelihood of the query generated from the sentence [8]
	Sentence Length	Number of terms in the sentence
	Sentence Location	Relative location of the sentence within the document
Sem	**ESA**	Cosine similarity between the query and the sentence ESA vectors
	Word2Vec	Average pairwise cosine similarity between any query and sentence word vectors
	TAGME	Jaccard coefficient between the query and the sentence entity sets
Con	$\mathbf{X_{before}}$	Feature X of the sentence immediately before this sentence
	$\mathbf{X_{after}}$	Feature X of the sentence immediately after this sentence

Table 2: Summary accuracy based on ROUGE scores and precision-oriented metrics.

System	R-1	R-2	R-SU4	NDCG@3	P@3
Severyn and Moschitti [9]	0.550	0.318	0.343	0.196	0.164
MK	**0.599**‡	**0.365**†	**0.389**†	**0.229**	**0.183**

(1.1% of all sentences) were affected by text truncation and only 94 of them (0.5% of all relevant sentences) were relevant, which appeared to be a reasonable trade-off.

Training a CNN model on the full WebAP data would take about two weeks which is too costly. We work around this problem by reducing each query to the top 1,000 sentences with the highest term overlap scores. The reduced data has 78,124 sentences in total, still larger than the factoid TREC QA dataset. The concentration of relevant sentences in the data is nevertheless two times higher, which might give the CNN method a slight advantage in this comparison.

The experimental results are given in Table 2, where the significance tests are directed against MK that is considered as a state-of-the-art query-biased summarization technique (†: $p < 0.05$ and ‡: $p < 0.01$). Here we use MART algorithm to learn the MK features, as suggested in the original paper [8]. Our results show that the CNN-based model is inferior to the MK approach in terms of ROUGE, NDCG@3, and P@3, suggesting that the query-biased summarization approach is more effective than the more sophisticated neural network model on this task.

5.2 Effect of Semantic and Context Features

In the second experiment, we examine the effectiveness of semantic and context features in extracting answer summaries from documents. A breakdown over different combinations of features and learning algorithms is given in Table 3, with the effectiveness of each combination measured in ROUGE scores and precision-oriented metrics. Best results are printed in boldface.

In general, we found that the results on NDCG@3 and P@3 are in line with ROUGE scores. The results show that in most cases adding semantic features leads to significant improvements over MK. LambdaMART obtains the biggest gain in overall effectiveness, boasting 9.0%, 20.3%, and 18.3% increases respectively in terms of ROUGE-1, -2, and -SU4 scores. On NDCG@3 and P@3, it also achieves 21.2% and 26.3% increases. Adding context features into MK+Sem boosts the overall effectiveness even further. The improvements are observed across all models, but using MART and LambdaMART, the difference appears to be more pronounced Combining MK+Sem+Con and LambdaMART gives the best result in our test, which amounts to a 12.8% increase in ROUGE-1, 31.6% in ROUGE-2, 28.4% in ROUGE-SU4, 47.2% in NDCG@3, and 49.7% in P@3.

5.3 Ablation Analysis

To understand the importance of each feature in extracting answer summaries, we perform an ablation analysis by removing one feature at a time from the set of 27 features. We choose to apply LambdaMART as it is shown the most accurate according to our results in section 5.2. Table 4 displays the top five features with the largest decrease in ROUGE-2 scores induced by feature ablations. The decreases are computed against the score of using a complete set of features. The accuracy of the model significantly degrades when removing the ESA feature, which is around two times lower than the decrease caused by removing the second top feature. The two most important features belong to semantic categories which supports our finding above regarding the effectiveness of semantic features. The next two important features are associated with context from the sentence after, that appear to be more critical than a query biased feature: LM score. The decreases in ROUGE-2 induced by these context features are however not significant.

5.4 Correlation between Measures

The key difference between the two types of measures adopted in this work is that ROUGE evaluates a summary as a single text unit to be compared with the ground truth answers, while the sentence ranking measures work at the sentence level. A summary that contains sentences each with a perfect annotation score but they in combination are not a ground truth answer, will obtain perfect score in the latter measure, but more likely lower result in the former one. Therefore, to better understand how well ROUGE and sentence ranking measures correlate each other, we compute a Pearson correlation between them. For this analysis, we use all results generated using different combinations of features and learning algorithms described in section 5.2. In total there are 12,924 (i.e. 1436 x 3 x 3) sets of scores, each consisting of ROUGE-1, -2, -SU4, P@3, and NDCG@3. All correlations are found to be statistically significant with $p < 0.01$ (see Table 5). All ROUGE scores are found to have moderate correlation with NDCG@3 and P@3, and the

Table 3: Summary accuracy across different feature sets and learning models. Significant differences with respect to MK are marked as †/‡ and with respect to MK+Sem as $^{*}/^{**}$ (for $p<0.05$ and $p<0.01$).

Feature Set	Model	R-1	R-2	R-SU4	NDCG@3	P@3
MK	CA	0.613	0.402	0.422	0.266	0.217
MK+Sem		0.633‡	0.429‡	0.446‡	0.289‡	0.231†
MK+Sem+Con		0.644‡	0.435‡	0.456‡	0.294‡	0.240‡
MK	MART	0.599	0.365	0.389	0.229	0.183
MK+Sem		0.619	0.396†	0.417†	0.260†	0.212‡
MK+Sem+Con		0.632‡	0.427‡**	0.447‡**	0.300‡**	0.246‡**
MK	λ-MART	0.586	0.354	0.377	0.231	0.179
MK+Sem		0.639‡	0.426‡	0.446‡	0.280‡	0.226‡
MK+Sem+Con		**0.661**‡**	**0.466**‡**	**0.484**‡**	**0.340**‡**	**0.268**‡**

Table 4: Top 5 features. Significant decreases of ROUGE-2 scores induced by the feature ablations are indicated by †/‡ (for $p<0.05$ and $p<0.01$)

No	Feature	Category	Decrease in R-2
1	ESA	Semantic	0.043‡ (-9.23%)
2	TAGME	Semantic	0.025 (-5.36%)
3	$\text{Length}_{\text{after}}$	Context	0.018 (-3.86%)
4	$\text{SynOverlap}_{\text{after}}$	Context	0.015 (-3.22%)
5	LM	MK	0.014 (-3.00%)

highest correlation is obtained by ROUGE-2.

Table 5: Correlation between Measures

	R-2	R-SU4	N@3	P@3
R-1	0.922‡	0.945‡	0.564‡	0.520‡
R-2	–	0.985‡	0.659‡	0.617‡
R-SU4	–	–	0.644‡	0.599‡
N@3	–	–	–	0.855‡

6. DISCUSSION AND FUTURE WORK

A state-of-the-art factoid QA method using neural network model is shown to be insufficient to generate accurate answer summaries for non factoid queries, affirming that our task is challenging. Results using three different learning models consistently show that using semantic and context features can help to extract better answer summaries from documents. This confirms the effectiveness of these features, which have been used in previous work [13]. It is worth exploring other information, for example: Community Question Answering content, and other techniques that could help in finding answers.

While we found here that there is a moderate correlation between ROUGE and sentence ranking measures, there is a gap to explore whether this information could help to create a better learning-to-rank based technique for our task. Next, it is also important to conduct user studies to evaluate the answer summaries according to users perspective, and analyse their agreement with offline metrics [15, 5].

7. ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (DP140102655) and the Indonesia Endowment Fund for Education (LPDP).

8. REFERENCES

[1] M. Ageev, D. Lagun, and E. Agichtein. Improving Search Result Summaries by Using Searcher Behavior Data. In *Proc. of SIGIR*, pages 13–22, 2013.

[2] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz. Direct Answers for Search Queries in the Long Tail. In *Proc. of SIGCHI*, pages 237–246, 2012.

[3] M. Keikha, J. H. Park, and W. B. Croft. Evaluating Answer Passages Using Summarization Measures. In *Proc. of SIGIR*, pages 963–966, 2014.

[4] M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson. Retrieving Passages and Finding Answers. In *Proc. of ADCS*, pages 81–84, 2014.

[5] L. Leal Bando, F. Scholer, and A. Turpin. Query-biased summary generation assisted by query expansion. *JASIST*, 66(5):961–979, 2015.

[6] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *Proc. of SIGIR*, pages 43–50, 2009.

[7] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. of ACL workshop on Text Summarization Branches Out*, volume 8, 2004.

[8] D. Metzler and T. Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*, pages 40–47, 2008.

[9] A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proc. of SIGIR*, pages 373–382, 2015.

[10] R. Soricut and E. Brill. Automatic Question Answering Using the Web: Beyond the Factoid. *Inf. Retr.*, 9(2):191–206, Mar. 2006.

[11] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to Rank Answers on Large Online QA Collections. In *Proc. of ACL*, pages 719–727, 2008.

[12] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proc. of SIGIR*, pages 475–482, 2008.

[13] L. Yang, Q. Ai, D. Spina, R.-C. Chen, L. Pang, W. B. Croft, J. Guo, and F. Scholer. Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval. In *Proc. of ECIR*, pages 115–128. 2016.

[14] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*, 2014.

[15] E. Yulianti, S. Huspi, and M. Sanderson. Tweet-biased summarization. *JASIST*, 67(6):1289–1300, 2016.