Long-Span Language Models for Query-focused Unsupervised Extractive Text Summarization

Mittul Singh[1], Arunav Mishra[2], Youssef Oualil[1], Klaus Berberich[2], and Dietrich Klakow[1]

[1] Spoken Language Systems (LSV)
[2] Max Planck Institute for Informatics
Saarland Informatics Campus, Saarbrücken, Germany
msingh@lsv.uni-saarland.de,amishra@mpi-inf.mpg.de

Abstract. Effective unsupervised query-focused extractive summarization systems use query-specific features along with short-range language models (LMs) in sentence ranking and selection summarization subtasks. We hypothesize that applying long-span n-gram-based and neural LMs that better capture larger context can help improve these subtasks. Hence, we outline the first attempt to apply long-span models to a query-focused summarization task in an unsupervised setting. We also propose the ***A****cross* ***S****entence* ***B****oundary* LSTM-based LMs, ***ASB****LSTM* and *bi****ASB****LSTM*, that is geared towards the query-focused summarization subtasks. Intrinsic and extrinsic experiments on a real word corpus with 100 Wikipedia event descriptions as queries show that using the long-span models applied in an integer linear programming (ILP) formulation of MMR criterion are the most effective against several state-of-the-art baseline methods from the literature.

1 Introduction & Background

Extractive text summarization system has been traditionally considered as an effective tool to address *information overloading* by facilitating efficient consumption of information that spans across multiple documents. Broadly, an extractive multi-document summarization task can be setup as *supervised*: having example summaries to design and train systems, or *unsupervised*: having access only to a large text corpus.

In the supervised setting, recent approaches [1, 2] that leverage deep neural network-based models (e.g., attention-based encoder-decoder LSTM [1]) require large number of training examples. Moreover, often the interpretability is poor thereby limiting their usages as black boxes. This issue makes it hard to gather insights into the methods which limit the scope of improvement. However, with enough examples, neural network-based models with end-to-end training have recently shown impressive results.

On the other hand, recent unsupervised approaches have utilized query-specific information with unigram Language Models (LMs) [3], and architectures such as Restricted Boltzmann Machines [4, 5]) for effective performance. Though state-of-the-art in extractive summarization, similar techniques that rely on document and corpus statistics have earlier been outperformed by unsupervised long-span neural LMs on similar sentence ranking tasks [6, 7] such as in question answering [8].

A typical unsupervised system that operates on a large corpus implements two stages [3–5]: **1)** *sentence ranking* to generate a candidate set of sentences from the entire corpus; and **2)** *sentence selection* from the candidate set to compose a (length budgeted) summary. In the sentence selection stage, traditional unsupervised extractive multi-document summarization systems address a *global inference problem* [9, 10] that aims to generate a length-budgeted summary with the candidate sentences that maximize the overall *relevance* while avoiding informational *redundancy*. In the past, several Integer Linear Programming (ILP)-based approaches [9, 11] based on the popular Maximal Marginal Relevance (MMR) [12] criterion have been shown to achieve high-quality results. Such an objective comes with explicit relevance and redundancy functions that can leverage LMs [3]. However, these ILP-based approaches work for summarizing a small number of pre-selected documents and do not scale to larger number input documents (e.g., entire corpus). Thus, a preliminary sentence ranking step is required to generate a smaller set of (top-k ranked) candidate sentences.

In this paper, we aim to study the effectiveness of *long-span-based neural sentence LMs* [13] for an unsupervised query-focused extractive multi-document summarization task. In a long-span LM, word probabilities are estimated by considering long-range dependencies within a large local context (e.g. few surrounding sentences, passages, or entire source document) in contrast to short-context models that use word independence (e.g., count-based LMs) and Markovian restrictions that use the previous word (e.g., n-gram LMs). Specifically, we address the problem recently proposed by [3] inputs short (single sentence) Wikipedia event descriptions as queries, and outputs a focused extractive summary from a longitudinal collection. In this task, we focus on developing long-span LMs that are robust to variable sentence lengths; and effective for computing relevance to query and inter-sentence redundancies for global inference.

We make the following key contributions in this paper: **1)** To the best of our knowledge, we are the first to incorporate across sentence-based [8] and LSTM-based long-span LMs for an unsupervised query-focused extractive summarization task. For sentence selection, we extend the ILP-based approach to incorporate the proposed LM. For application of long-span LM to this ILP, sentence *relevance* is computed by scoring the candidate sentences conditioned on the query whereas, for *inter-sentence redundancy*, we propose comparing query words given the candidate sentences (Section 2) to allow comparison of arbitrary length sentences. **2)** We present two *Across Sentence Boundary* LSTM-based LMs: *ASBLSTM* and *biASBLSTM* (Section 2.1), that build over an LSTM-based LM for our task. **3)** Intrinsic and extrinsic experiments are performed (Section 3) to evaluate the effectiveness of the LMs with a test query set containing 100 Wikipedia event descriptions released by [3] on the English Gigaword corpus [14].

2 Approach

We next describe the two summarization stages and the proposed language model.

Sentence Ranking: As the first stage, the primary goal is to generate a candidate set CS of sentences to reduce the search space during the summarization process. For this purpose, we employ two steps: **1)** For a given event description as a query q, we retrieve a set of top-k documents using query likelihood retrieval framework as pseudo-relevance feedback [15]. **2)** Then, sentences within the retrieved documents are ranked according to the generative probability $P(s|q)$ where s is a candidate sentence. Finally, we consider the top-100 sentences as the CS.

Sentence Selection: In this second stage, we solve an ILP to select sentences from the generated CS with the criterion [11]: **Max** $\sum_i \lambda rel_i \xi_i - (1-\lambda)\sum_{j \neq i} red_{ij} M_{ij}$, where rel_i is the relevance score of a sentence; and red_{ij} represents the redundancy between two candidate sentences. ξ_i and M_{ij} are indicator variables and λ controls the importance of relevance and redundancy. We refer to [11] for full details on the ILP.

In our setup, rel_i computes the likelihood $P(s|q)$ of generating a sentence s from a given query q. The redundancy red_{ij} between two sentences is computed as the Jensen-Shanon Divergence (JSD) between sentence LMs. In the ILP objective, while computing red_{ij} between two sentences, we find that the long-span LMs, if estimated naïvely, suffer from lack of query-relevant context (e.g., terms that are semantically related to the query terms) which leads to poor estimate of redundancy w.r.t the query. Thus, we propose to compute JSD between query LMs conditioned on the candidate sentences to estimate their redundancy. For the given query q; two candidates s_i and s_j; and the respective LMs $P(q|s_i)$ and $P(q|s_j)$, $red_{ij} = -JSD(P(q|s_i)||P(q|s_j))$. Intuitively, calculating redundancy in such a manner compares the predictive nature of different query terms given the sentences of arbitrary lengths.

2.1 ASBLSTM Language Model

LSTM-based methods have shown impressive improvements in language modelling tasks (e.g., Machine Translation and Speech Recognition) in comparison to standard count-based methods [15]. The main advantage of LSTMs is a single state that encodes the global linguistic context and controls its longevity using a forget gate. However, the individual hidden state in LSTM tends to lose the long-term context [13] which also becomes essential for query-based sentence ranking tasks.

To reduce this loss, [8] have incorporated across sentence information explicitly. These LMs learn to trigger words across sentences instead of just the within-sentence triggers. Intuitively, in such a triggering scheme a sentence is less divergent (or more relevant) to an adjacent sentence (query) if the words in the sentence predict words in the adjacent sentence with a higher probability. However, in a standard LSTM architecture, the recurrent state focusses more on the within-sentence words as triggers while losing the information around the sentence. Thus, implying that sentences with more within-sentence triggers are heavily boosted while not considering the impact of across sentence triggers, which is more relevant for a query-specific setup. We address this issue by introducing an extra memory state (as shown in Figure 1) into the architecture that stores the LSTM state of the previous sentence s_{-1} and uses this state while scoring the current sentence s, hence, calculating $P(s|s_{-1})$. We refer to this LM as the Across Sentence Boundary-based LSTM or *ASBLSTM*.

Here, we assume that each sentence is represented by the hidden layer state achieved at the end of the sentence. Hence, the previous sentence information is contained in the hidden state ($H_{s_{-1}}$) observed at the end of the previous sentence. Using this hidden

Language Models	*Uni*	Dirichlet smoothed unigram LM proposed by [3].
	ASB	An across boundary n-gram sentence LM proposed by [8]. For training, we look at a window of one previous sentence in the source document.
	LSTM	A stateful *LSTM*-based language model estimated with projection and hidden layers containing 200 nodes, and a vocabulary-sized output layer for sentence selection. The stateful LSTM variant initializes the first hidden state with the last hidden state of the previous sentence.
	biASB, *biLSTM*, *biASBLSTM*	The **bi-** suffix is added to denote that the above described LMs and *ABSLSTM* LM introduced in Section 2.1 are trained using both left and right context. Training using bi-directional context alleviates the inherent boosting of sentences that *follow* query words within a document.
Schemes	*Random*	This approach selects sentences form CS at uniform random.
	MEAD	Uses centroid based *MEAD D*, lead-based *MEAD LB* [16] that come with the open source framework.
	Rdh	Uses Dirichlet smoothed unigram LMs for sentence selection with corpus background model [3].
	Gil	Maximizes query-salient *words* with an ILP [10] using an advanced unigram query LM [3].

Table 1: Baseline schemes and language models for sentence ranking and selection.

state, the output layer (O_t) is defined as the combination of within-sentence context captured by the present hidden state H_t and the across-sentence boundary context captured by the hidden state $H_{s_{-1}}$ as, $O_t = g_{softmax}(WH_t + W_s H_{s_{-1}})$. At inference, we compute $P(s|q)$ and $P(q|s)$ for sentence ranking and selection stages. For the query-specific long-span LMs, the query is used as the previous sentence of a candidate s for computing $P(s|q)$, whereas their order is reversed while computing $P(s|q)$.

3 Experimental Evaluation

Data: Test query set contains 100 random event descriptions that happened between 1987 and 2007 with Wikipedia articles central to the events as the gold standard summaries. This test set was publicly released by [3]. Our target is English Gigaword corpus [14] with about 9 million news articles published between 1994 to 2008 taken from four different news sources. We evaluate the 250 worded system-generated summaries against the gold standard Wikipedia articles using standard Rouge-2 and -SU4 measures. We make our data publicly available[3].

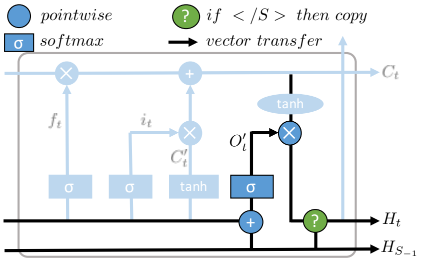
The disparate quantity of text between the gold standard Wikipedia articles and the system generated summaries result in low Rouge scores. Thus, we perform one-tailed student's t-test over the Rouge-SU4 scores. Significant improvements at levels 0.05 and 0.01 are indicated by △ and ▲ while decrements by ▽ and ▼.

Methods: Table 1 describes the different baseline methods under comparison.

Implementation: Using long-span LMs for this summarization system, we restrict the LM vocabulary to 80000 most frequent words. This vocabulary is then appended with words included in the test queries. Rest of the words are replaced by an out-of-vocabulary symbol. This modification allows for constraining the parametric size of LMs leading to a faster processing of the data. Our LMs are trained on the pseudo-relevant documents for all the queries, allowing the LMs to learn the triggering information but staying agnostic to explicit query information. Only at the time of inference, query text is used in LMs to help score candidate sentences.

[3] `http://resources.mpi-inf.mpg.de/d5/asblstmSumm`

Fig. 1: ASBLSTM block with additional $H_{s_{-1}}$ state. Last hidden state of the previous sentence, $H_{s_{-1}}$, is updated for a sentence separator symbol (</S>)



LM used for CS	Rouge-2 F1	Rouge-SU4 F1
Uni	0.0406	0.0542 (▽)
ASB	**0.0415**	**0.0549** (-)
LSTM	0.0377	0.0516 (▲)
biLSTM	0.0383	0.0509 (▲)
ASBLSTM	0.0395	0.0527 (▲)
biASBLSTM	0.0390	0.0516 (▲)

Table 2: Oracle summary results with significance test against the *ASB*.

LM used for CS	LM used for selection	Rouge-2 F1	Rouge-SU4 F1
	Random	0.0293	0.0388 (▼/▼)
	MEAD LB	0.0287	0.0400 (▼/▼)
	MEAD D	0.0305	0.0422 (▼/▼)
	Rdh	0.0364	0.0510 (-/△)
	Gil	0.0381	0.0531 (△/-)
Uni	*ASB*	0.0382	0.0548 (▲/△)
	biASB	0.0383	0.0549 (▲/△)
	LSTM	0.0388	**0.0553** (▲/△)
	biLSTM	0.0387	0.0550 (▲/△)
	ASBLSTM	0.0387	0.0547 (▲/△)
	biASBLSTM	**0.0389**	0.0552 (▲/▲)
ASB	*ASB*	0.0371	0.0534 (-/-)
	biASB	0.0376	0.0537 (△/△)
	LSTM	0.0393	0.0550 (▲/▲)
	biLSTM	0.0392	**0.0552** (▲/▲)
	ASBLSTM	0.0396	0.0547 (▲/▲)
	biASBLSTM	**0.0399**	**0.0552** (▲/▲)
ASBLSTM	*ASB*	0.0378	0.0536 (△/△)
	biASB	0.0377	0.0539 (△/△)
	LSTM	0.0371	0.0527 (-/-)
	biLSTM	0.0369	0.0524 (-/-)
	ASBLSTM	0.0368	0.0523 (-/-)
	biASBLSTM	**0.0383**	**0.0538** (▲/△)

Table 3: Sentence selection results with significance tests against (*Rdh*/ *Gil*).

Sentence Ranking: Since the relevance judgments for sentence in a candidate set CS generated with an LM is not available; we design an extrinsic experiment in contrast to an information retrieval style evaluation. We leverage the notion that a CS containing more query-informative sentences will lead to generating better summaries. First, we create a CS by reranking using a LM and selecting top-100 sentences. Then this is input to an *Oracle* genetic algorithm proposed by Riedhammer et al. [17] that is aware of gold standard Wikipedia articles to generate the best possible summary.

Table 2, reports the Rouge scores of the oracle summarizer with different candidate sets CS as input. The CS generated using simpler unigram and *n*-gram based ASB outperform those generated with LSTM-based models. A recent work [18] argues that the neural LSTM models are more suited for the semantically oriented task rather than retrieval-based ranking tasks, where *n*-gram-based LMs work better. Our finding conforms with this argument during sentence ranking stage.

Sentence Selection: Table 3 reports the performance of the different schemes using the CS generated with estimated LMs. We find that using *biASBLSTM* LM for sentence selection acorss all the CS proves to be most effective. As expected, the *Random* to be the worst. *MEAD_LB* considers only the documents' lead paragraphs and is also not able to achieve good scores. *MEAD_D* represents centroid based method and performs significantly worse than the ILP-based *Rdh* and *Gil* methods. The best combination is to use an n-gram-based *ASB*-based sentence LM for the ranking, and *biASBLSTM* LM for selection stage. This shows significant improvements over the state-of-the-art coverage-based *Gil* and MMR-style *Rdh* by approximately 5% and 10% in Rouge-2 F1.

In summary, LSTMs have been shown to capture semantic relations from within the text. In summarization, such semantic relations can better model the notion of inter-sentence redundancy [18]. This observation is also reflected in our experiments where we find that adding more context, in fact, improves the quality of short text summaries.

4 Conclusion

In this paper, we proposed applying long-span models to query-focused unsupervised text summarization. We presented the *ASBLSTM* LM for the sentence ranking and selection stages of summarization. In summary, the *ASBLSTM* LM outperform other models in for the sentence selection stage. A scheme that uses an n-gram-based across sentence boundary (ASB) LM for sentence ranking and *biASBLSTM* LM for selection stage of summarization, demonstrated to be the most effective.

References

[1] Cao Z. et al. Attsum: Joint learning of focusing and summarization with neural attention. *arXiv preprint arXiv:1604.00125*, 2016.

[2] Nallapati R. et al. Classify or select: Neural architectures for extractive document summarization. *arXiv preprint arXiv:1611.04244*, 2016.

[3] Mishra et al. Event digest: A holistic view on past events. In: *SIGIR* 2016.

[4] Zhong S. et al. Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications*, 42(21) , 2015.

[5] Yousefi-Azar M. et al. Text summarization using unsupervised deep learning. *Expert Systems with Applications*, 68, 2017.

[6] Cao Z. et a. Ranking with recursive neural networks and its application to multi-document summarization. In: *AAAI* 2015.

[7] Palangi H. et al. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. In: *TASLP* 2016.

[8] Momtazi S. et al. Trained trigger language model for sentence retrieval in qa: Bridging the vocabulary gap. In: *CIKM* 2011.

[9] McDonald R. et al. A study of global inference algorithms in multi-document summarization. In: *ECIR* 2007.

[10] Gillick D. et al. A scalable global model for summarization. In: *ILP-NAACL-HLT* 2009.

[11] Riedhammer K. et al. Long story short – global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10), 2010.

[12] Carbonell J. et al. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: *SIGIR* 1998.

[13] Oualil Y. et al. Long-short range context neural networks for language modeling. *arXiv preprint arXiv:1708.06555*, 2017.

[14] English Gigaword Corpus. `https://catalog.ldc.upenn.edu/ldc2003t05`

[15] Zhai C. X. et al. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1), 2008.

[16] Radev D. R. et al. Mead-a platform for multidocument multilingual text summarization. In: *LREC* 2004.

[17] Riedhammer K. et al. Packing the meeting summarization knapsack. In: *INTERSPEECH* 2008.

[18] Guo J. et al. A deep relevance matching model for ad-hoc retrieval. In: *CIKM* 2016.