Sentence Centrality Revisited for Unsupervised Summarization

Hao Zheng and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
Hao.Zheng@ed.ac.uk mlap@inf.ed.ac.uk

Abstract

Single document summarization has enjoyed renewed interest in recent years thanks to the popularity of neural network models and the availability of large-scale datasets. In this paper we develop an unsupervised approach arguing that it is unrealistic to expect large-scale and high-quality training data to be available or created for different types of summaries, domains, or languages. We revisit a popular graph-based ranking algorithm and modify how node (aka sentence) centrality is computed in two ways: (a) we employ BERT, a state-of-the-art neural representation learning model to better capture sentential meaning and (b) we build graphs with directed edges arguing that the contribution of any two nodes to their respective centrality is influenced by their relative position in a document. Experimental results on three news summarization datasets representative of different languages and writing styles show that our approach outperforms strong baselines by a wide margin.[1]

1 Introduction

Single-document summarization is the task of generating a shorter version of a document while retaining its most important content (Nenkova et al., 2011). Modern neural network-based approaches (Nallapati et al., 2016; Paulus et al., 2018; Nallapati et al., 2017; Cheng and Lapata, 2016; See et al., 2017; Narayan et al., 2018b; Gehrmann et al., 2018) have achieved promising results thanks to the availability of large-scale datasets containing hundreds of thousands of document-summary pairs (Sandhaus, 2008; Hermann et al., 2015b; Grusky et al., 2018). Nevertheless, it is unrealistic to expect that large-scale and high-quality training data will be available or created for different summarization styles (e.g., highlights vs. single-sentence summaries), domains (e.g., user- vs. professionally-written articles), and languages.

It therefore comes as no surprise that unsupervised approaches have been the subject of much previous research (Marcu, 1997; Radev et al., 2000; Lin and Hovy, 2002; Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Wan, 2008; Wan and Yang, 2008; Hirao et al., 2013; Parveen et al., 2015; Yin and Pei, 2015; Li et al., 2017). A very popular algorithm for extractive single-document summarization is TextRank (Mihalcea and Tarau, 2004); it represents document sentences as nodes in a graph with *undirected* edges whose weights are computed based on sentence similarity. In order to decide which sentence to include in the summary, a node's *centrality* is often measured using graph-based ranking algorithms such as PageRank (Brin and Page, 1998).

In this paper, we argue that the centrality measure can be improved in two important respects. Firstly, to better capture sentential meaning and compute sentence similarity, we employ BERT (Devlin et al., 2018), a neural representation learning model which has obtained state-of-the-art results on various natural language processing tasks including textual inference, question answering, and sentiment analysis. Secondly, we advocate that edges should be *directed*, since the contribution induced by two nodes' connection to their respective centrality can be in many cases unequal. For example, the two sentences below are semantically related:

(1) Half of hospitals are letting patients jump NHS queues for cataract surgery if they pay for it themselves, an investigation has revealed.

(2) Clara Eaglen, from the royal national in-

[1] Our code is available at https://github.com/mswellhao/PacSum.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6236–6247
Florence, Italy, July 28 - August 2, 2019.

stitute of blind people, said: "It's shameful that people are being asked to consider funding their own treatment when they are entitled to it for free, and in a timely manner on the NHS."

Sentence (1) describes a news event while sentence (2) comments on it. Sentence (2) would not make much sense on its own, without the support of the preceding sentence, whose content is more central. Similarity as an undirected measure, cannot distinguish this fundamental intuition which is also grounded in theories of discourse structure (Mann and Thompson, 1988) postulating that discourse units are characterized in terms of their text importance: *nuclei* denote central segments, whereas *satellites* denote peripheral ones.

We propose a simple, yet effective approach for measuring directed centrality for single-document summarization, based on the assumption that the contribution of any two nodes' connection to their respective centrality is influenced by their *relative* position. Position information has been frequently used in summarization, especially in the news domain, either as a baseline that creates a summary by selecting the first n sentences of the document (Nenkova, 2005) or as a feature in learning-based systems (Lin and Hovy, 1997; Schilder and Kondadadi, 2008; Ouyang et al., 2010). We transform undirected edges between sentences into directed ones by differentially weighting them according to their *orientation*. Given a pair of sentences in the same document, one is looking forward (to the sentences following it), and the other is looking backward (to the sentences preceding it). For some types of documents (e.g., news articles) one might further expect sentences occurring early on to be more central and therefore backward-looking edges to have larger weights.

We evaluate the proposed approach on three single-document news summarization datasets representative of different languages, writing conventions (e.g., important information is concentrated in the beginning of the document or distributed more evenly throughout) and summary styles (e.g., verbose or more telegraphic). We experimentally show that position-augmented centrality significantly outperforms strong baselines (including TextRank; Mihalcea and Tarau 2004) across the board. In addition, our best system achieves performance comparable to supervised systems trained on hundreds of thousands of examples (Narayan et al., 2018b; See et al., 2017). We present an alternative to more data-hungry models, which we argue should be used as a standard comparison when assessing the merits of more sophisticated supervised approaches over and above the baseline of extracting the leading sentences (which our model outperforms).

Taken together, our results indicate that directed centrality improves the selection of salient content substantially. Interestingly, its significance for unsupervised summarization has gone largely unnoticed in the research community. For example, `gensim` (Barrios et al., 2016), a widely used open-source implementation of TextRank only supports building undirected graphs, even though follow-on work (Mihalcea, 2004) experiments with position-based directed graphs similar to ours. Moreover, our approach highlights the effectiveness of pretrained embeddings for the summarization task, and their promise for the development of unsupervised methods in the future. We are not aware of any previous neural-based approaches to unsupervised single-document summarization, although some effort has gone into developing unsupervised models for multi-document summarization using reconstruction objectives (Li et al., 2017; Ma et al., 2016; Chu and Liu, 2018).

2 Centrality-based Summarization

2.1 Undirected Text Graph

A prominent class of approaches in unsupervised summarization uses graph-based ranking algorithms to determine a sentence's salience for inclusion in the summary (Mihalcea and Tarau, 2004; Erkan and Radev, 2004). A document (or a cluster of documents) is represented as a graph, in which nodes correspond to sentences and edges between sentences are weighted by their similarity. A node's centrality can be measured by simply computing its degree or running a ranking algorithm such as PageRank (Brin and Page, 1998).

For single-document summarization, let D denote a document consisting of a sequence of sentences $\{s_1, s_2, ..., s_n\}$, and e_{ij} the similarity score for each pair (s_i, s_j). The degree centrality for sentence s_i can be defined as:

$$\textbf{centrality}(s_i) = \sum_{j \in \{1,..,i-1,i+1,..,n\}} e_{ij} \qquad (1)$$

After obtaining the centrality score for each sentence, sentences are sorted in reverse order and the

top ranked ones are included in the summary.

TextRank (Mihalcea and Tarau, 2004) adopts PageRank (Brin and Page, 1998) to compute node centrality recursively based on a Markov chain model. Whereas degree centrality only takes local connectivity into account, PageRank assigns relative scores to all nodes in the graph based on the recursive principle that connections to nodes having a high score contribute more to the score of the node in question. Compared to degree centrality, PageRank can in theory be better since the global graph structure is considered. However, we only observed marginal differences in our experiments (see Sections 4 and 5 for details).

2.2 Directed Text Graph

The idea that textual units vary in terms of their importance or salience, has found support in various theories of discourse structure including Rhetorical Structure Theory (RST; Mann and Thompson 1988). RST is a compositional model of discourse structure, in which elementary discourse units are combined into progressively larger discourse units, ultimately covering the entire document. Discourse units are linked to each other by rhetorical relations (e.g., *Contrast*, *Elaboration*) and are further characterized in terms of their text importance: *nuclei* denote central segments, whereas *satellites* denote peripheral ones. The notion of nuclearity has been leveraged extensively in document summarization (Marcu, 1997, 1998; Hirao et al., 2013) and in our case provides motivation for taking directionality into account when measuring centrality.

We could determine nuclearity with the help of a discourse parser (Li et al. 2016; Feng and Hirst 2014; Joty et al. 2013; Liu and Lapata 2017, inter alia) but problematically such parsers rely on the availability of annotated corpora as well as a wider range of standard NLP tools which might not exist for different domains, languages, or text genres. We instead approximate nuclearity by relative position in the hope that sentences occurring earlier in a document should be more central. Given any two sentences s_i, s_j $(i < j)$ taken from the same document D, we formalize this simple intuition by transforming the undirected edge weighted by the similarity score e_{ij} between s_i and s_j into two directed ones differentially weighted by $\lambda_1 e_{ij}$ and $\lambda_2 e_{ij}$. Then, we can refine the centrality score of s_i based on the directed graph as follows:

$$\textbf{centrality}(s_i) = \lambda_1 \sum_{j<i} e_{ij} + \lambda_2 \sum_{j>i} e_{ij} \qquad (2)$$

where λ_1, λ_2 are different weights for forward- and backward-looking directed edges. Note that when λ_1 and λ_1 are equal to 1, Equation (2) becomes degree centrality. The weights can be tuned experimentally on a validation set consisting of a small number of documents and corresponding summaries, or set manually to reflect prior knowledge about how information flows in a document. During tuning experiments, we set $\lambda_1 + \lambda_2 = 1$ to control the number of free hyper-parameters. Interestingly, we find that the optimal λ_1 tends to be negative, implying that similarity with previous content actually hurts centrality. This observation contrasts with existing graph-based summarization approaches (Mihalcea and Tarau, 2004; Mihalcea, 2004) where nodes typically have either no edge or edges with positive weights. Although it is possible to use some extensions of PageRank (Kerchove and Dooren, 2008) to take negative edges into account, we leave this to future work and only consider the definition of centrality from Equation (6) in this paper.

3 Sentence Similarity Computation

The key question now is how to compute the similarity between two sentences. There are many variations of the similarity function of TextRank (Barrios et al., 2016) based on symbolic sentence representations such as tf-idf. We instead employ a state-of-the-art neural representation learning model. We use BERT (Devlin et al., 2018) as our sentence encoder and fine-tune it based on a type of sentence-level distributional hypothesis (Harris, 1954; Polajnar et al., 2015) which we explain below. Fine-tuned BERT representations are subsequently used to compute the similarity between sentences in a document.

3.1 BERT as Sentence Encoder

We use BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018) to map sentences into deep continuous representations. BERT adopts a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017) and uses two unsupervised prediction tasks, i.e., masked language modeling and next sentence prediction, to pre-train the encoder.

The language modeling task aims to predict masked tokens by jointly conditioning on both left and right context, which allows pre-trained representations to fuse both contexts in contrast to conventional uni-directional language models. Sentence prediction aims to model the relationship between two sentences. It is a binary classification task, essentially predicting whether the second sentence in a sentence pair is indeed the next sentence. Pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference. We use BERT to encode sentences for unsupervised summarization.

3.2 Sentence-level Distributional Hypothesis

To fine-tune the BERT encoder, we exploit a type of sentence-level distributional hypothesis (Harris, 1954; Polajnar et al., 2015) as a means to define a training objective. In contrast to skip-thought vectors (Kiros et al., 2015) which are learned by reconstructing the surrounding sentences of an encoded sentence, we borrow the idea of negative sampling from word representation learning (Mikolov et al., 2013). Specifically, for a sentence s_i in document D, we take its previous sentence s_{i-1} and its following sentence s_{i+1} to be positive examples, and consider any other sentence in the corpus to be a negative example. The training objective for s_i is defined as:

$$\log \sigma({v'_{s_{i-1}}}^{\top} v_{s_i}) + \log \sigma({v'_{s_{i+1}}}^{\top} v_{s_i}) + \mathbb{E}_{s \sim P(s)} \left[\log \sigma(-{v'_s}^{\top} v_s) \right] \quad (3)$$

where v_s and v'_s are two different representations of sentence s via two differently parameterized BERT encoders; σ is the sigmoid function; and $P(s)$ is a uniform distribution defined over the sentence space.

The objective in Equation (3) aims to distinguish context sentences from other sentences in the corpus, and the encoder is pushed to capture the meaning of the intended sentence in order to achieve that. We sample five negative samples for each positive example to approximate the expectation. Note, that this approach is much more computationally efficient, compared to reconstructing surrounding sentences (Kiros et al., 2015).

Dataset	# docs	avg. document		avg. summary	
		words	sen.	words	sen.
CNN+DM	11,490	641.9	28.0	54.6	3.9
NYT	4,375	1,290.5	50.7	79.8	3.5
TTNews	2,000	1,037.1	21.8	44.8	1.1

Table 1: Statistics on NYT, CNN/Daily Mail, and TTNews datasets (test set). We compute the average document and summary length in terms of number of words and sentences, respectively.

3.3 Similarity Matrix

Once we obtain representations $\{v_1, v_2, ..., v_n\}$ for sentences $\{s_1, s_2, \ldots, s_n\}$ in document D, we employ pair-wise dot product to compute an unnormalized similarity matrix $\bar{\mathrm{E}}$:

$$\bar{\mathrm{E}}_{ij} = {v_i}^{\top} v_j \quad (4)$$

We could also use cosine similarity, but we empirically found that the dot product performs better.

The final normalized similarity matrix E is defined based on $\bar{\mathrm{E}}$:

$$\tilde{\mathrm{E}}_{ij} = \bar{\mathrm{E}}_{ij} - \left[\min \bar{\mathrm{E}} + \beta(\max \bar{\mathrm{E}} - \min \bar{\mathrm{E}}) \right] \quad (5)$$

$$\mathrm{E}_{ij} = \begin{cases} \tilde{\mathrm{E}}_{ij} & \text{if } \tilde{\mathrm{E}}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Equation (5) aims to remove the effect of absolute values by emphasizing the relative contribution of different similarity scores. This is particularly important for the adopted sentence representations which in some cases might assign very high values to all possible sentence pairs. Hyper-parameter β ($\beta \in [0, 1]$) controls the threshold below which the similarity score is set to 0.

4 Experimental Setup

In this section we present our experimental setup for evaluating our unsupervised summarization approach which we call PACSUM as a shorthand for **P**osition-**A**ugmented **C**entrality based **Sum**marization.

4.1 Datasets

We performed experiments on three recently released single-document summarization datasets representing different languages, document information distribution, and summary styles. Table 1 presents statistics on these datasets (test set); example summaries are shown in Table 5.

The CNN/DailyMail dataset (Hermann et al., 2015a) contains news articles and associated highlights, i.e., a few bullet points giving a brief

overview of the article. We followed the standard splits for training, validation, and testing used by supervised systems (90,266/1,220/1,093 CNN documents and 196,961/12,148/10,397 DailyMail documents). We did not anonymize entities.

The LEAD-3 baseline (selecting the first three sentences in each document as the summary) is extremely difficult to beat on CNN/DailyMail (Narayan et al., 2018b,a), which implies that salient information is mostly concentrated in the beginning of a document. NYT writers follow less prescriptive guidelines[2], and as a result salient information is distributed more evenly in the course of an article (Durrett et al., 2016). We therefore view the NYT annotated corpus (Sandhaus, 2008) as complementary to CNN/DailyMail in terms of evaluating the model's ability of finding salient information. We adopted the training, validation and test splits (589,284/32,736/32,739) widely used for evaluating abstractive summarization systems. However, as noted in Durrett et al. (2016), some summaries are extremely short and formulaic (especially those for obituaries and editorials), and thus not suitable for evaluating extractive summarization systems. Following Durrett et al. (2016), we eliminate documents with summaries shorter than 50 words. As a result, the NYT test set contains longer and more elaborate summary sentences than CNN/Daily Mail (see Table 1).

Finally, to showcase the applicability of our approach across languages, we also evaluated our model on TTNews (Hua et al., 2017), a Chinese news summarization corpus, created for the shared summarization task at NLPCC 2017. The corpus contains a large set of news articles and corresponding human-written summaries which were displayed on the Toutiao app (a mobile news app). Because of the limited display space on the mobile phone screen, the summaries are very concise and typically contain just one sentence. There are 50,000 news articles with summaries and 50,000 news articles without summaries in the training set, and 2,000 news articles in test set.

4.2 Implementation Details

For each dataset, we used the documents in the training set to fine-tune the BERT model; hyperparameters $(\lambda_1, \lambda_2, \beta)$ were tuned on a validation set consisting of 1,000 examples with gold summaries, and model performance was evaluated on the test set.

We used the publicly released BERT model[3] (Devlin et al., 2018) to initialize our sentence encoder. English and Chinese versions of BERT were respectively used for the English and Chinese corpora. As mentioned in Section 3.2, we fine-tune BERT using negative sampling; we randomly sample five negative examples for every positive one to create a training instance. Each mini-batch included 20 such instances, namely 120 examples. We used Adam (Kingma and Ba, 2014) as our optimizer with initial learning rate set to 4e-6.

5 Results

5.1 Automatic Evaluation

We evaluated summarization quality automatically using ROUGE F1 (Lin and Hovy, 2003). We report unigram and bigram overlap (ROUGE-1 and ROUGE-2) as a means of assessing informativeness and the longest common subsequence (ROUGE-L) as a means of assessing fluency.

NYT and CNN/Daily Mail Table 2 summarizes our results on the NYT and CNN/Daily Mail corpora (examples of system output can be found in the Appendix). We forced all extractive approaches to select three summary sentences for fair comparison. The first block in the table includes two state-of-the-art supervised models. REFRESH (Narayan et al., 2018b) is an extractive summarization system trained by globally optimizing the ROUGE metric with reinforcement learning. POINTER-GENERATOR (See et al., 2017) is an abstractive summarization system which can copy words from the source text while retaining the ability to produce novel words. As an upper bound, we also present results with an extractive oracle system. We used a greedy algorithm similar to Nallapati et al. (2017) to generate an oracle summary for each document. The algorithm explores different combinations of sentences and generates an oracle consisting of multiple sentences which maximize the ROUGE score against the gold summary.

The second block in Table 2 presents the results of the LEAD-3 baseline (which simply creates a summary by selecting the first three sentences in a document) as well as various instantiations of

[2] https://archive.nytimes.com/www.nytimes.com/learning/issues_in_depth/10WritingSkillsIdeas.html

[3] https://github.com/google-research/bert

Method	NYT			CNN+DM		
	R-1	R-2	R-L	R-1	R-2	R-L
ORACLE	61.9	41.7	58.3	54.7	30.4	50.8
REFRESH[4] (Narayan et al., 2018b)	41.3	22.0	37.8	41.3	18.4	37.5
POINTER-GENERATOR (See et al., 2017)	42.7	22.1	38.0	39.5	17.3	36.4
LEAD-3	35.5	17.2	32.0	40.5	17.7	36.7
DEGREE (tf-idf)	33.2	13.1	29.0	33.0	11.7	29.5
TEXTRANK (tf-idf)	33.2	13.1	29.0	33.2	11.8	29.6
TEXTRANK (skip-thought vectors)	30.1	9.6	26.1	31.4	10.2	28.2
TEXTRANK (BERT)	29.7	9.0	25.3	30.8	9.6	27.4
PACSUM (tf-idf)	40.4	20.6	36.4	39.2	16.3	35.3
PACSUM (skip-thought vectors)	38.3	18.8	34.5	38.6	16.1	34.9
PACSUM (BERT)	41.4	21.7	37.5	40.7	17.8	36.9

Table 2: Test set results on the NYT and CNNDailyMail datasets using ROUGE F1 (R-1 and R-2 are shorthands for unigram and bigram overlap, R-L is the longest common subsequence).

TEXTRANK (Mihalcea and Tarau, 2004). Specifically, we experimented with three sentence representations to compute sentence similarity. The first one is based on tf-idf where the value of the corresponding dimension in the vector representation is the number of occurrences of the word in the sentence times the idf (inverse document frequency) of the word. Following `gensim`, We preprocessed sentences by removing function words and stemming words. The second one is based on the skip-thought model (Kiros et al., 2015) which exploits a type of sentence-level distributional hypothesis to train an encoder-decoder model trying to reconstruct the surrounding sentences of an encoded sentence. We used the publicly released skip-thought model[5] to obtain vector representations for our task. The third one is based on BERT (Devlin et al., 2018) fine-tuned with the method proposed in this paper. Finally, to determine whether the performance of PageRank and degree centrality varies in practice, we also include a graph-based summarizer with DEGREE centrality and tf-idf representations.

The third block in Table 2 reports results with three variants of our model, PACSUM. These include sentence representations based on tf-idf, skip-thought vectors, and BERT. Recall that PACSUM uses directed degree centrality to decide which sentence to include in the summary. On both NYT and CNN/Daily Mail datasets, PACSUM (with BERT representations) achieves the highest ROUGE F1 score, compared to other unsupervised approaches. This gain is more pronounced on NYT where the gap between our best system and LEAD-3 is approximately 6 absolute ROUGE-1 F1 points. Interestingly, despite limited access to only 1,000 examples for hyperparameter tuning, our best system is comparable to supervised systems trained on hundreds of thousands of examples (see rows REFRESH and POINTER-GENERATOR in the table).

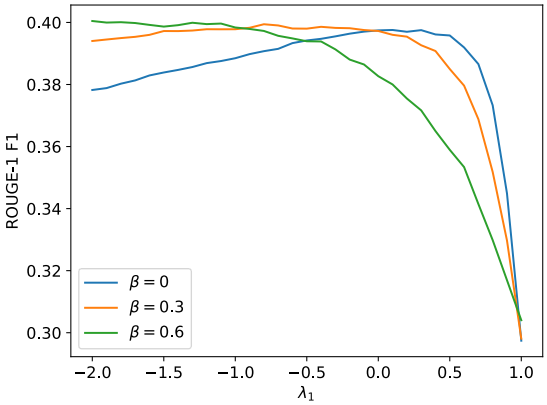

Figure 1: PACSUM's performance against different values of λ_1 on the NYT validation set with with $\lambda_2 = 1$. Optimal hyper-parameters $(\lambda_1, \lambda_2, \beta)$ are $(-2, 1, 0.6)$.

As can be seen in Table 2, DEGREE (tf-idf) is very close to TEXTRANK (tf-idf). Due to space limitations, we only show comparisons between DEGREE and TEXTRANK with tf-idf, however, we observed similar trends across sentence representations. These results indicate that considering global structure does not make a difference when selecting salient sentences for NYT and CNN/Daily Mail, possibly due to the fact

[4]The ROUGE scores here on CNN/Daily Mail are higher than those reported in the original paper, because we extract 3 sentences in Daily Mail rather than 4.

[5]https://github.com/ryankiros/skip-thoughts

Method	TTNews		
	R-1	R-2	R-L
ORACLE	45.6	31.4	41.7
POINTER-GENERATOR	42.7	27.5	36.2
LEAD	30.8	18.4	24.9
TEXTRANK (tf-idf)	25.6	13.1	19.7
PACSUM (BERT)	32.8	18.9	26.1

Table 3: Results on Chinese TTNews corpus using ROUGE F1 (R-1 and R-2 are shorthands for unigram and bigram overlap, R-L is the longest common subsequence).

Method	NYT	CNN+DM	TTNews
ORACLE	49.0*	53.9*	60.0*
REFRESH	42.5	34.2	—
LEAD	34.7*	26.0*	50.0*
PACSUM	44.4	31.1	56.0

Table 4: Results of QA-based evaluation on NYT, CNN/Daily Mail, and TTNews. We compute a system's final score as the average of all question scores. Systems statistically significant from PACSUM are denoted with an asterisk * (using a one-way ANOVA with posthoc Tukey HSD tests; $p < 0.01$).

that news articles in these datasets are relatively short (see Table 1). The results in Table 2 further show that PACSUM substantially outperforms TEXTRANK across sentence representations, directly confirming our assumption that position information is beneficial for determining sentence centrality in news single-document summarization. In Figure 1 we further show how PACSUM's performance (ROUGE-1 F1) on the NYT validation set varies as λ_1 ranges from -2 to 1 ($\lambda_2 = 1$ and $\beta = 0, 0.3, 0.6$). The plot highlights that differentially weighting a connection's contribution (via relative position) has a huge impact on performance (ROUGE ranges from 0.30 to 0.40). In addition, the optimal λ_1 is negative, suggesting that similarity with the previous content actually hurts centrality in this case.

We also observed that PACSUM improves further when equipped with the BERT encoder. This validates the superiority of BERT-based sentence representations (over tf-idf and skip-thought vectors) in capturing sentence similarity for unsupervised summarization. Interestingly, TEXTRANK performs worse with BERT. We believe that this is caused by the problematic centrality definition, which fails to fully exploit the potential of continuous representations. Overall, PACSUM obtains improvements over baselines on both datasets highlighting the effectiveness of our approach across writing styles (highlights vs. summaries) and narrative conventions. For instance, CNN/Daily Mail articles often follow the inverted pyramid format starting with the most important information while NYT articles are less prescriptive attempting to pull the reader in with an engaging introduction and develop from there to explain a topic.

TTNews Dataset Table 3 presents our results on the TTNews corpus using ROUGE F1 as our evaluation metric. We report results with variants of TEXTRANK (tf-idf) and PACSUM (BERT) which performed best on NYT and CNN/Daily Mail. Since summaries in the TTNews corpus are typically one sentence long (see Table 1), we also limit our extractive systems to selecting a single sentence from the document. The LEAD baseline also extracts the first document sentence, while the ORACLE selects the sentence with maximum ROUGE score against the gold summary in each document. We use the popular POINTER-GENERATOR system of See et al. (2017) as a comparison against supervised methods.

The results in Table 3 show that POINTER-GENERATOR is superior to unsupervised methods, and even comes close to the extractive oracle, which indicates that TTNews summaries are more abstractive compared to the English corpora. Nevertheless, even in this setting which disadvantages extractive methods, PACSUM outperforms LEAD and TEXTRANK showing that our approach is generally portable across different languages and summary styles. Finally, we show some examples of system output for the three datasets in Appendix.

5.2 Human Evaluation

In addition to automatic evaluation using ROUGE, we also evaluated system output by eliciting human judgments. Specifically, we assessed the degree to which our model retains key information from the document following a question-answering (QA) paradigm which has been previously used to evaluate summary quality and document compression (Clarke and Lapata, 2010; Narayan et al., 2018b). We created a set of questions based on the gold summary under the assumption that it highlights the most important document content. We then examined whether partici-

NYT
Gold Summary: Marine Corps says that V-22 Osprey, hybrid aircraft with troubled past, will be sent to Iraq in September, where it will see combat for first time. The Pentagon has placed so many restrictions on how it can be used in combat that plane – which is able to drop troops into battle like helicopter and then speed away like airplane – could have difficulty fulfilling marines longstanding mission for it. limitations on v-22, which cost $80 million apiece, mean it can not evade enemy fire with same maneuvers and sharp turns used by helicopter pilots. **Questions:** • Which aircraft will be sent to Iraq? V-22 Osprey • What are the distinctive features of this type of aircraft? able to drop troops into battle like helicopter and then speed away like airplane • How much does each v-22 cost? $80 million apiece
CNN+DM
Gold Summary: "We're all equal, and we all deserve the same fair trial," says one juror. The months-long murder trial of Aaron Hernandez brought jurors together. Foreperson: "It's been an incredibly emotional toll on all of us." **Questions:** • Who was on trial? Aaron Hernandez • Who said: "It's been an incredibly emotional toll on all of us"? Foreperson
TTNEWS
Gold Summary : 皇马今夏清洗名单曝光，三小将租借外出，科恩特朗、伊利亚拉门迪将被永久送出伯纳乌球场. (Real Madrid's cleaning list was exposed this summer, and the three players will be rented out. Coentrao and Illarramendi will permanently leave the Bernabeu Stadium.) **Question:** 皇马今夏清洗名单中几人将被外租？三 (How many people will be rented out by Real Madrid this summer? three)

Table 5: NYT, CNN/Daily Mail and TTNews with corresponding questions. Words highlighted in red are answers to those questions.

pants were able to answer these questions by reading system summaries alone without access to the article. The more questions a system can answer, the better it is at summarizing the document.

For CNN/Daily Mail, we worked on the same 20 documents and associated 71 questions used in Narayan et al. (2018b). For NYT, we randomly selected 18 documents from the test set and created 59 questions in total. For TTNews, we randomly selected 50 documents from the test set and created 50 questions in total. Example questions (and answers) are shown in Table 5.

We compared our best system PACSUM (BERT) against REFRESH, LEAD-3, and ORACLE on CNN/Daily Mail and NYT, and against LEAD-3 and ORACLE on TTNews. Note that we did not include TEXTRANK in this evaluation as it performed worse than LEAD-3 in previous experiments (see Tables 2 and 3). Five participants answered questions for each summary. We used the same scoring mechanism from Narayan et al. (2018b), i.e., a correct answer was marked with a score of one, partially correct answers with a score of 0.5, and zero otherwise. The final score for a system is the average of all its question scores. Answers for English examples were elicited using Amazon's Mechanical Turk crowdsourcing platform while answers for Chinese summaries were assessed by in-house native speakers of Chinese. We uploaded the data in batches (one system at a time) on AMT to ensure that the same participant does not evaluate summaries from different systems on the same set of questions.

The results of our QA evaluation are shown in Table 4. ORACLE's performance is below 100, indicating that extracting sentences by maximizing ROUGE fails in many cases to select salient content, capturing surface similarity instead. PACSUM significantly outperforms LEAD but is worse than ORACLE which suggests there is room for further improvement. Interestingly, PACSUM performs on par with REFRESH (the two systems are not significantly different).

6 Conclusions

In this paper, we developed an unsupervised summarization system which has very modest data requirements and is portable across different types of summaries, domains, or languages. We revisited a popular graph-based ranking algorithm and refined how node (aka sentence) centrality is computed. We employed BERT to better capture sentence similarity and built graphs with directed edges arguing that the contribution of any two nodes to their respective centrality is influenced by their relative position in a document. Experimental results on three news summarization datasets demonstrated the superiority of our approach against strong baselines. In the future, we would like to investigate whether some of the ideas introduced in this paper can improve the performance of supervised systems as well as sentence selection in multi-document summarization.

Acknowledgments

The authors gratefully acknowledge the financial support of the European Research Council (Lapata; award number 681760). This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9118. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation therein.

References

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of TextRank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Sergey Brin and Michael Page. 1998. Anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th Conference on World Wide Web*, pages 107–117, Brisbane, Australia.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany.

Eric Chu and Peter J. Liu. 2018. Unsupervised neural multi-document abstractive summarization. *CoRR*, abs/1810.05739.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, USA.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015a. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015b. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Morgan, Kaufmann.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA.

Lifeng Hua, Xiaojun Wan, and Lei Li. 2017. Overview of the nlpcc 2017 shared task: Single document summarization. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 942–947. Springer.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria.

Cristobald de Kerchove and Paul Van Dooren. 2008. The pagetrust algorithm: How to rank web pages when negative links are allowed? In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 346–352. SIAM.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Piji Li, Zihao Wang, Wai Lam, Zhaochun Ren, and Lidong Bing. 2017. Salience estimation via variational auto-encoders for multi-document summarization. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3497–3503, San Francisco, California.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 283–290, Washington, DC, USA.

Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 457–464, Pennsylvania, Philadelphia.

Chin Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 71–78, Edmonton, Canada.

Yang Liu and Mirella Lapata. 2017. Learning contextually informed representations for linear-time discourse parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1300–1309, Copenhagen, Denmark.

Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. 2016. An unsupervised multi-document summarization framework based on neural document model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1514–1523, Osaka, Japan.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain.

Daniel Marcu. 1998. Improving summarization through rhetorical parsing tuning. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 206–215, Montréal, Canada.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 170–173, Barcelona, Spain.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3075–3081, San Francisco, California.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence*, pages 1436–1441, Pittsburgh, Pennsylvania.

Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

You Ouyang, Wenjie Li, Qin Lu, and Renxian Zhang. 2010. A study on position information in document summarization. In *Coling 2010: Posters*, pages 919–927, Beijing, China.

Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. 2015. Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1954, Lisbon, Portugal.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Tamara Polajnar, Laura Rimell, and Stephen Clark. 2015. An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 1–11, Lisbon, Portugal.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL-ANLP 2000 Workshop: Automatic Summarization*, pages 21–30, Seattle, Washington.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Frank Schilder and Ravikumar Kondadadi. 2008. Fastsum: Fast and accurate query-based multi-document summarization. In *Proceedings of ACL-08: HLT, Short Papers*, pages 205–208, Columbus, Ohio.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Xiaojun Wan. 2008. An exploration of document impact on graph-based multi-document summarization. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 755–762, Honolulu, Hawaii.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st Annual International ACL SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, Singapore.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1383–1389, Buenos Aires, Argentina.

A Appendix

A.1 Examples of System Output

Table 6 shows examples of system output. Specifically, we show summaries produced from GOLD, LEAD, TEXTRANK and PACSUM for test documents in NYT, CNN/Daily Mail and TTNews. GOLD is the gold summary associated with each document; LEAD extracts the first document sentences; TextRank (Mihalcea and Tarau, 2004) adopts PageRank (Brin and Page, 1998) to compute node centrality recursively based on a Markov chain model; PACSUM is position augmented centrality based summarization approach introduced in this paper.

	NYT	CNN+DM	TTNews
GOLD	Marine Corps says that V-22 Osprey, hybrid aircraft with troubled past, will be sent to Iraq in September, where it will see combat for first time. The Pentagon has placed so many restrictions on how it can be used in combat that plane – which is able to drop troops into battle like helicopter and then speed away like airplane – could have difficulty fulfilling marines longstanding mission for it. Limitations on v-22, which cost $80 million apiece, mean it can not evade enemy fire with same maneuvers and sharp turns used by helicopter pilots.	”We’re all equal, and we all deserve the same fair trial.” says one juror. The months-long murder trial of Aaron Hernandez brought jurors together. Foreperson: ”It’s been an incredibly emotional toll on all of us.”	皇马今夏清洗名单曝光，三小将租借外出，科恩特朗、伊利亚拉门迪将被永久送出伯纳乌球场. (Real Madrid’s cleaning list was exposed this summer, and the three players will be rented out. Coentrao and Illarramendi will permanently leave the Bernabeu Stadium.)
TEXTRANK	The Pentagon has placed so many restrictions on how it can be used in combat that the plane – which is able to drop troops into battle like a helicopter and then speed away from danger like an airplane – could have difficulty fulfilling the marines ’ longstanding mission for it. Because of these problems, Mr. Coyle, the former pentagon weapons tester, predicted the marines will use the v-22 to ferry troops from one relatively safe spot to another, like a flying truck. In December 2000, four more marines, including the program’s most experienced pilot, were killed in a crash caused by a burst hydraulic line and software problems.	A day earlier, Strachan, the jury foreperson, announced the first-degree murder conviction in the 2013 shooting death of Hernandez’s onetime friend Odin Lloyd. Before the trial, at least one juror – Rosalie Oliver – had n’t heard of the 25-year-old defendant who has now gone from a $ 40 million pro-football contract to a term of life without parole in a maximum-security prison. Rosalie Oliver – the juror who had n’t heard of Hernandez before the trial – said that, for her, the first shot was enough.	2个赛季前，皇马花费3500万欧元引进了伊利亚拉门迪，巴斯克人在安切洛蒂手下就知道，他在皇马得不到好机会，现在主教练换成了贝尼特斯，情况也没有变化。(Two seasons ago, Real Madrid spent 35 million euros to introduce Illarramendi. The Basques knew under Ancelotti that he could not get a good chance in Real Madrid. Now the head coach has changed to Benitez. The situation has not changed.)
LEAD	the Marine Corps said yesterday that the V-22 Osprey, a hybrid aircraft with a troubled past, will be sent to Iraq this September, where it will see combat for the first time. But because of a checkered safety record in test flights, the v-22 will be kept on a short leash. The Pentagon has placed so many restrictions on how it can be used in combat that the plane – which is able to drop troops into battle like a helicopter and then speed away from danger like an airplane – could have difficulty fulfilling the marines ’ longstanding mission for it.	(CNN) After deliberating for more than 35 hours over parts of seven days, listening intently to the testimony of more than 130 witnesses and reviewing more than 400 pieces of evidence, the teary-eyed men and women of the jury exchanged embraces. Since late January, their work in the Massachusetts murder trial of former NFL star Aaron Hernandez had consumed their lives. It was nothing like “Law & Order.”	新浪体育显示图片厄德高新赛季可能会被皇马外租，皇马主席弗罗伦蒂诺已经获得了贝尼特斯制定的“清洗黑名单”。(Sina Sports shows that Ödegaard this season may be rented by Real Madrid, Real Madrid President Florentino has obtained the ”cleansing blacklist” developed by Benitez.)
PACSUM	The Marine Corps said yesterday that the V-22 Osprey, a hybrid aircraft with a troubled past, will be sent to Iraq this September, where it will see combat for the first time. The Pentagon has placed so many restrictions on how it can be used in combat that the plane — which is able to drop troops into battle like a helicopter and then speed away from danger like an airplane — could have difficulty fulfilling the Marines’ longstanding mission for it. The limitations on the V-22, which cost $80 million apiece, mean it cannot evade enemy fire with the same maneuvers and sharp turns used by helicopter pilots.	(CNN) After deliberating for more than 35 hours over parts of seven days, listening intently to the testimony of more than 130 witnesses and reviewing more than 400 pieces of evidence, the teary-eyed men and women of the jury exchanged embraces. Since late January, their work in the Massachusetts murder trial of former NFL star Aaron Hernandez had consumed their lives. ”It ’s been an incredibly emotional toll on all of us.” Lesa Strachan told CNN ’s Anderson Cooper Thursday in the first nationally televised interview with members of the jury.	厄德高、卢卡斯-席尔瓦和阿森西奥将被租借外出，而科恩特朗和伊利亚拉门迪，则将被永久送出伯纳乌球场。(Ödegaard, Lucas Silva and Asencio will be rented out, while Coentrao and Illarramendi will permanently leave the Bernabeu Stadium.)

Table 6: Example gold summaries and system output for NYT, CNN/Daily Mail and TTNews documents.