Modeling Content Importance for Summarization with Pre-trained Language Models

Liqiang Xiao[1], **Lu Wang**[2], **Hao He**[1,3]*, **Yaohui Jin**[1,3]*
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Computer Science and Engineering, University of Michigan
[3]State Key Lab of Advanced Optical Communication System and Network,
Shanghai Jiao Tong University
xiaoliqiang@sjtu.edu.cn, wangluxy@umich.edu
{hehao, jinyh}@sjtu.edu.cn

Abstract

Modeling *content importance* is an essential yet challenging task for summarization. Previous work is mostly based on statistical methods that estimate word-level salience, which does not consider semantics and larger context when quantifying importance. It is thus hard for these methods to generalize to semantic units of longer text spans. In this work, we apply *information theory* on top of *pre-trained language models* and define the concept of importance from the perspective of *information amount*. It considers both the semantics and context when evaluating the importance of each semantic unit. With the help of pre-trained language models, it can easily generalize to different kinds of semantic units (n-grams or sentences). Experiments on CNN/Daily Mail and New York Times datasets demonstrate that our method can better model the importance of content than prior work based on F1 and ROUGE scores.

1 Introduction and Related Work

Text summarization aims to compress long document(s) into a concise summary while maintaining the salient information. It often consists of two critical subtasks, *important information identification* and *natural language generation* (for abstractive summarization). With the advancements of large pre-trained language models (PreTLMs) (Devlin et al., 2019; Yang et al., 2019), state-of-the-art results are achieved on both natural language understanding and generation. However, it is still unclear how well these large models can estimate "content importance" for a given document.

Previous studies for modeling importance are either *empirical-based*, which implicitly encode importance during document summarization, or *theory-based*, which often lacks support by empirical experiments (Peyrard, 2019). Benefiting from the large-scale summarization datasets (Nallapati et al., 2016; Narayan et al., 2018), data-driven approaches (Nallapati et al., 2017; Paulus et al., 2018; Zhang et al., 2019) have made significant progress. Yet most of them conduct the information selection implicitly while generating the summaries. It lacks theory support and is hard to be applied to low-resource domains. In another line of work, structure features (Zheng and Lapata, 2019), such as centrality, position, and title, are employed as proxies for importance. However, the information captured by these features can vary in texts of different genres.

To overcome this problem, theory-based methods (Louis, 2014; Peyrard, 2019; Lin et al., 2006) aim to formalize the concept of importance, and develop general-purpose systems by modeling the background knowledge of readers. This is based on the intuition that humans are good at identifying important content by using their own interpretation of the world knowledge. Theoretical models usually rely on *information theory* (IT) (Shannon, 1948). Louis (2014) uses Dirichlet distribution to represent the background knowledge and employs Bayesian surprise to find novel information. Peyrard (2019) instead models the importance with entropy, assuming the important words should be frequent in the given document but rare in the background.

However, statistical method is only a rough evaluation for informativity, which largely ignores the effect of semantic and context. In fact, the information amount of units is not only determined by frequency, but also by its semantic meaning, context, as well as reader's background knowledge. In addition, bag-of-words approaches are difficult to generalize beyond unigrams due to the sparsity of n-grams when n is large.

In this paper, we propose a novel and general-purpose approach to model content importance for

[*] Corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3606–3611,
November 16–20, 2020. ©2020 Association for Computational Linguistics

summarization. We employ *information theory* on top of *pre-trained language models*, which are expected to better capture the information amount of semantic units by leveraging their meanings and context. We argue that *important content contains information that cannot be directly inferred from context and background knowledge*. Large pre-trained language models are suitable for our study since they are trained from large-scaled datasets consisting of diverse documents and thus containing a wide range of knowledge.

We conduct experiments on popular summarization benchmarks of CNN/Daily Mail and New York Times corpora, where we show that our proposed method can outperform prior importance estimation models. We further demonstrate that our method can be adapted to model semantic units of different scales (n-grams and sentences).

2 Methodology

In this section, we first estimate the amount of information by using information theory with pre-trained language models (§2.1 and §2.2), where we consider both the context and semantic meaning of a given text unit. We then propose a formal definition of importance for text summarization from a perspective of information amount (§2.3).

2.1 Information Theory

Information theory (IT), as invented by Shannon (1948), has been used on *words* to quantify their "informativity". Concretely, IT uses the frequency of semantic units x_i to approximate the probability $P(x_i)$ and uses negative logarithm of frequency as the measurement for information, which is called *self-info*[1]:

$$I(x_i) = -\log_2 P(x_i) \tag{1}$$

It approximates the information amount of a unit (e.g. word) in a given corpus.

However, traditional IT suffers from the sparsity problem of longer n-grams and also ignores semantics and context. Advanced compression algorithms in IT (Hirschberg and Lelewer, 1992) attempt to model the context to better estimate the information amount. But due to the sparsity, they can only count up to third-order statistics. Statistical methods are nearly impossible to reliably calculate the probability of x_i conditioned on its context, e.g., $P(x_i|\cdots, x_{i-1}, x_{i+1}, \cdots)$, as the number of combinations of the context can be explosive.

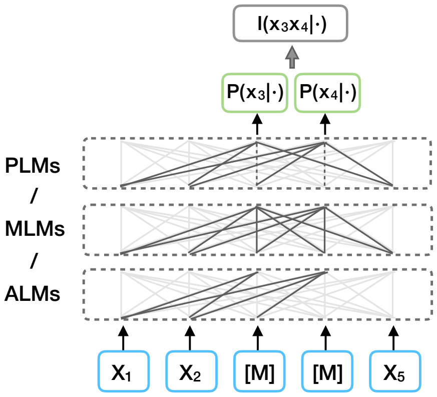

Figure 1: Information amount evaluation with language models. Here we take a subsequence x_3x_4 as example. [M] denotes mask and PLMs/MLMs/ALMs are three different options for language models. $I(x_3x_4|\cdot) = -\log[P(x_3|\cdot)P(x_4|\cdot)]$, where conditions for different models are omitted for brevity.

2.2 Using Language Models in Information Theory

With the development of deep learning, neural language models can efficiently predict the probability of a specified unit, such as a word or a phrase, given its context, which makes it feasible to calculate high-order approximation for the information amount.

We thus propose to use neural language models to replace the statistical models for estimating the information amount of a given semantic unit. Language models can be categorized as follows, and we present information estimation method for each as shown in Fig. 1.

Auto-regressive Language Model (ALM) (Bengio et al., 2000) is the most commonly used probabilistic model to depict the distribution of language, which is usually referred as unidirection LM (UniLM). Given a sequence of tokens $x_{0:T} = [x_0, x_1, \cdots, x_T]$, UniLMs use leftward content to estimate the conditional probability for each token: $P(x_t|x_{<t}) = g_{\text{UniLM}}(x_{<t})$, where g_{UniLM} denotes a neural network for language model and $x_{<t}$ represents the sequence from x_0 to x_{t-1}. Then the joint probability of a subsequence is factorized as:

$$P(x_{m:n}|x_{<m}) = \prod_{t=m}^{n} P(x_t|x_{<t}) \tag{2}$$

After applying Eq. (1) to both sides of Eq. (2), we can obtain the information amount of the subse-

[1] The unit of information is "bit", with base of 2. In the rest of this paper, we omit base 2 for brevity.

quence conditioned on its context as:

$$I(x_{m:n}|x_{<m}) = \sum_{t=m}^{n} I(x_t|x_{<t}) \tag{3}$$

Masked Language Model (MLM) is proposed by Taylor (1953) and combined with pre-training by Devlin et al. (2019) to encode bidirectional context. MLM masks a certain number of tokens from the input sequence, then predicts these tokens based on the unmasked ones. The conditional probability of a masked token x_t can be estimated as: $P(x_t|x_{\neq t}) = g_{\text{MLM}}(x_{\neq t})$, where $\neq t$ indicates that the t-th token is masked. Information amount of a given subsequence of the input is calculated as:

$$I(x_{m:n}|x_{\notin[m:n]}) = \sum_{t=m}^{n} I(x_t|x_{\notin[m,n]}) \tag{4}$$

Since MLMs encode both leftward and rightward context, intuitively, it can better estimate the information of current tokens than UniLMs.

Permutation Language Model (PLM) is proposed by (Yang et al., 2019) to combine ALMs and MLMs, by considering the dependency between the masked tokens as well as overcoming the problem caused by discrepancy of pre-training and fine-tuning in MLMs. It models the dependency of the tokens by maximizing the expected likelihood of all possible permutations of factorization orders. The probability prediction can be formalized as: $P(z_t|z_{<t}) = g_{\text{PLM}}(z_{<t})$ where z denotes a possible permutation sequence of input. Information of a subsequence is estimated as:

$$I(z_{m:n}|z_{<m}) = \sum_{t=m}^{n} I(z_t|z_{<t}) \tag{5}$$

2.3 Modeling Importance with Pre-trained Language Model

We argue that important content should be hard to be predicted based on background knowledge only; it should be also difficult to be inferred from the context. Moreover, detecting important content is to find the most informative part from the input. As described in (Shann, 1989), the information amount is a quantification of the uncertainty we have for the semantic units. But the degree of uncertainty is relative to reader's *background knowledge*. The less knowledge the reader has, the more uncertainty the source shows.

We thus employ pre-trained language models, which contain a wide range of knowledge, to represent background knowledge. If a semantic unit is frequently mentioned in the training corpus, it will get *high probability* during inference and thus *low information amount*. We further propose a notion of **importance** as the information amount conditional on the background knowledge:

$$\text{Imp}(x_i|X - x_i, K) = -\log P_{LM_K}(x_i|X - x_i) \tag{6}$$

where $X - x_i$ means the context excluding[2] the unit x_i from input X and K denotes the knowledge encoded in the pre-trained model. In practice, when calculating the importance of a semantic unit, we first exclude all its occurrences from the input document, and let the PreTLMs predict the probability of each occurrence, based on which the information amount is calculated. As the same unit may appear at multiple positions in the input, summation is used as the final value of information amount.

Based on our notion of importance, a summarization model is to maximize the overall importance of a subset $\mathbf{x}$ of the input X, with a length constraint, such as $\sum_{x_i \in \mathbf{x}} |x_i| < l_{max}$:

$$\arg\max_{\mathbf{x} \subset X} \text{Imp}(\mathbf{x}) = \sum_{x_i \in \mathbf{x}} \text{Imp}(x_i|X - x_i, K) \tag{7}$$

3 Experimental Setups

Semantic Units and Tasks. Our theory can be generalized for evaluating the importance for any scale of semantic units. To verify the effectiveness of our theory, we instantiate the semantic unit with three common forms: **unigram**, **bigram** and **sentence**. In this way, our method can also be regarded as a general unsupervised information extraction system, serving as a keyphrase extraction or sentence-level extractive summarization model. As our method exploits the existed PreTLMs and needs no additional training, it has the potential of benefiting the low-resource languages and domains.

In *unigram* scenario, we simply instantiate semantic unit x_i as a token w_t and calculate its importance with $\text{Imp}(w_t) = -\log P(w_t|w_{\neq t}, K)$. For evaluation, top-$k$ important ones are selected and F_1 score is calculated by comparing against the reference, where the value of k is set by grid search.

Importance of *bigrams*, e.g., $x_i = w_t w_{t+1}$, can be represented as a joint probability of two tokens: $\text{Imp}(w_t w_{t+1}) = -\log P(w_t w_{t+1}|w_{t \notin [t,t+1]}, K)$. Same as unigrams, F_1 score is computed to evaluate the accuracy.

By extending the formula of bigram importance to longer sequences, we get importance definition

[2] MLMs hide targets by replacing them with special tokens, PLMs use attention masks, and ALMs only see leftward context.

	CNN/DM					NYT				
	UNI.	BI.	SENTENCE			UNI.	BI.	SENTENCE		
Models	**F_1**	**F_1**	**R-1**	**R-2**	**R-L**	**F_1**	**F_1**	**R-1**	**R-2**	**R-L**
TF·IDF	17.08	12.25	-	-	-	22.65	11.65	-	-	-
STM	38.78	16.76	-	-	-	**34.10**	16.49	-	-	-
BayesianSR	37.72	23.04	27.50	8.19	25.19	23.95	19.47	25.12	8.89	22.54
LexRank	12.04	11.43	33.96	11.79	30.17	18.70	13.97	27.32	11.93	23.75
TextRank	-	-	33.20	11.80	29.60	-	-	**33.20**	13.10	29.00
TextRank+BERT	-	-	30.80	9.60	27.40	-	-	29.70	9.00	25.30
SumBasic	-	-	31.72	9.60	28.58	-	-	23.16	7.18	20.06
Imp + GPT-2 (ALM)	34.73	26.02	35.06	12.41	32.62	27.96	15.96	26.69	9.22	24.13
Imp + BERT (MLM)	39.93	**29.39**	**37.53**	**14.71**	**34.71**	31.86	**20.07**	32.26	**14.48**	**29.28**
Imp + DistillBERT (MLM)	38.59	28.29	34.25	11.75	31.68	32.84	19.75	29.16	12.23	26.53
Imp + XLNet (PLM)	33.90	25.44	37.04	13.50	34.01	30.01	18.89	29.24	11.82	26.40

Table 1: Results of importance modeling. UNI./BI. denote unigram and bigram. R-1/R-2/R-L are ROUGE-1, ROUGE-2 and ROUGE-L respectively. Best results per metric are in bold. Among our models (bottom), Imp yields significantly higher scores on all metrics except when using unigrams as semantic unit and with sentences (based on R-1) on NYT (Welch's t-test, p<0.05).

for a *sentence* as: $\mathrm{Imp}(s_i) = I(s_i|w_{\notin s_i}, K) = -\log P(s_i|w_{\notin s_i}, K)$. For evaluation, we select a subset of sentences with Eq. (7) and calculate the ROUGE scores (Lin, 2004) against reference summary. The length constraints for CNN/DM and NYT are set to 105 and 95 tokens respectively.

Datasets. We evaluate our method on the test set of two popular summarization datasets: CNN/Daily Mail (abbreviated as CNN/DM) (Nallapati et al., 2017) and New York Times (Sandhaus, 2008). Following See et al. (2017)[3], we use the non-anonymized version that does not replace the name entities, which is most commonly used in recent work. We preprocess them as described in (Paulus et al., 2018). For unigram experiments, we remove all the stop words and punctuation in the reference summaries and treat the notional words as the predicting targets. For bigram, we first collect all the bigrams in source document and then discard the ones containing stop words or punctuation. The rest bigrams are employed as the predicting targets.

Comparisons. We compare our method with two types of models: (1) the methods that estimate importance for n-grams. We consider TF·IDF, a numerical statistic to reflect how important a term is to a document, and STM (Peyrard, 2019), a simple theoretic model for content importance based on statistical information theory. (2) unsupervised models for extractive summarization. We adopt centrality-based models LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004) and TextRank+BERT (Zheng and Lapata, 2019), a frequency-based model SumBasic (Ani Nenkova, 2005), and BayesianSR (Louis, 2014) which scores words or sentences with Bayesian surprise.

4 Results

We conduct extensive experiments with pre-trained models[4] in all three types of language models, including ALM: GPT-2 (Radford et al., 2019); MLMs: BERT (Devlin et al., 2019), and DistillBert (Sanh et al., 2019); PLMs: XLNet (Yang et al., 2019).

As shown in Table 1, our method Imp consistently outperform prior models. Among comparisons, we can see that theory-based methods, STM and BayesianSR, achieve better results. This is because they have statistics estimated for background distribution, which helps filter out common words. The significant advantage of our method verifies our hypothesis that pre-trained language models better characterize the background knowledge, which in turn more precisely calculate the importance of each semantic unit. Moreover, our methods have a more significant improvement on bigram-level prediction than unigram-level. This is due to the fact that Imp-based models overcome the sparsity issue, where they can evaluate the importance of a phrase by considering its semantic meaning and context.

[3] https://github.com/JafferWilson/Process-Data-of-CNN-DailyMail

[4] We use the implementations and parameters from huggingface.co/transformers/index.html

Surprisingly, our method can also generalize to sentence-level semantic units and serve as an unsupervised extract-based model for summarization. Our models achieve significantly higher ROUGE scores than previous work by average 2.02. This observation inspires a potential future direction for sentence-level importance modeling based on background knowledge as well as context information.

We also compare the performance of PreTLMs in different categories. MLMs, including BERT and DISTILLBERT, have the best overall performance, since they are able to encode bidirectional context. PLM, i.e. XLNet, is slightly inferior to MLMs because the probabilities of the words are related to the order of their permutation, which may hurt importance estimation by our method.

5 Future Work

In the future work, we would like to fine-tune the current language models on the target of $\max P(x_i|X - x_i)$ to better align with the interpretation of information theory. Currently, the PreTLMs mostly mask the text randomly, which still differ from our current method's objective.

Background knowledge also deserves further investigation. The background knowledge of our methods comes from the pre-training process of language models, suggesting that the information distribution largely depends on the training data. Meanwhile, most PreTLMs are trained with Wikipedia or books, which may affect the determination of content importance from text with different styles. So domain-specific knowledge, such as genres or topics, can be included in the future work.

6 Conclusion

We propose to use large pre-trained language models to estimate the information amount of given text units, by filtering out the background knowledge as encoded in the large models. We show that the large pre-trained models can be used as unsupervised methods for content importance estimation, where significant improvement over nontrivial baselines is achieved on both keyphrase extraction and sentence-level extractive summarization tasks.

Acknowledgement

This research is supported in part by National Key Research and Development Program of China under Grant 2018YFC0830400. Lu Wang is supported in part by National Science Foundation through Grant IIS-1813341.

References

Lucy Vanderwende Ani Nenkova. 2005. The impact of frequency on summarization. *Microsoft Research*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 932–938.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Daniel S. Hirschberg and Debra A. Lelewer. 1992. Context modeling for text compression.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*.

Annie Louis. 2014. A bayesian method to incorporate background knowledge during automatic text summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 333–338.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411.

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 Thirty-First AAAI Conference on Artificial Intelligence, February 4-9,* *2017, San Francisco, California, USA*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1059–1073.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Patrick Shann. 1989. The selection of a parsing strategy for an on-line machine translation system in a sublanguage domain. a new practical comparison. In *Proceedings of the First International Workshop on Parsing Technologies*, pages 264–276, Pittsburgh, Pennsylvania, USA. Carnegy Mellon University.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6236–6247.