

Semi-Supervised Learning for Neural Keyphrase Generation

Hai Ye* and Lu Wang

College of Computer and Information Science

Northeastern University

Boston, MA 02115

hye.me@outlook.com luwang@ccs.neu.edu

Abstract

We study the problem of generating keyphrases that summarize the key points for a given document. While sequence-to-sequence (seq2seq) models have achieved remarkable performance on this task (Meng et al., 2017), model training often relies on large amounts of labeled data, which is only applicable to resource-rich domains. In this paper, we propose semi-supervised keyphrase generation methods by leveraging both labeled data and large-scale unlabeled samples for learning. Two strategies are proposed. First, unlabeled documents are first tagged with synthetic keyphrases obtained from unsupervised keyphrase extraction methods or a self-learning algorithm, and then combined with labeled samples for training. Furthermore, we investigate a multi-task learning framework to jointly learn to generate keyphrases as well as the titles of the articles. Experimental results show that our semi-supervised learning-based methods outperform a state-of-the-art model trained with labeled data only.

1 Introduction

Keyphrase extraction concerns the task of selecting a set of phrases from a document that can indicate the main ideas expressed in the input (Turney, 2000; Hasan and Ng, 2014). It is an essential task for document understanding because accurate identification of keyphrases can be beneficial for a wide range of downstreaming natural language processing and information retrieval applications. For instance, keyphrases can be leveraged to improve text summarization systems (Zhang et al., 2004; Liu et al., 2009a; Wang and Cardie, 2013), facilitate sentiment analysis and opinion mining (Wilson et al., 2005; Berend, 2011), and help with document clustering (Hammouda et al., 2005). Though relatively

*Work was done while visiting Northeastern University.

Document:

In this paper, we consider an enthalpy formulation for a two-phase Stefan problem arising from the solidification of aluminum during *casting* process. We solve this free boundary problem in a time varying three-dimensional domain and consider convective heat transfer in the liquid phase. The resulting equations are discretized using a characteristics method in time and a *finite element* method in space, and we propose a numerical algorithm to solve the obtained nonlinear discretized problem. Finally, numerical results are given which are compared with industrial experimental measurements.

Keyphrase:

casting; *thermal*; *conduction*; *convection*; *finite element*

in document not in document

Figure 1: Sample document with labeled keyphrases.

easy to implement, extract-based approaches fail to generate keyphrases that do not appear in the source document, which are frequently produced by human annotators as shown in Figure 1. Recently, Meng et al. (2017) propose to use a sequence-to-sequence model (Sutskever et al., 2014) with copying mechanism for *keyphrase generation*, which is able to produce phrases that are not in the input documents.

While seq2seq model demonstrates good performance on keyphrase generation (Meng et al., 2017), it heavily relies on massive amounts of labeled data for model training, which is often unavailable for new domains. To overcome this drawback, in this work, we investigate *semi-supervised learning for keyphrase generation*, by leveraging abundant unlabeled documents along with limited labeled data. Intuitively, additional documents, though unlabeled, can provide useful knowledge on universal linguistic features and discourse structure, such as context information for keyphrases and that keyphrases are likely to be noun phrases or main verbs. Furthermore, learning with unlabeled data can also mitigate the overfitting problem, which is often caused by small size of labeled training data, and thus improve model generalizability and enhance keyphrase generation performance on unseen data.

Concretely, two major approaches are proposed for leveraging unlabeled data. For the first method,

unlabeled documents are first tagged with synthetic keyphrases, then mixed with labeled data for model pre-training. Synthetic keyphrases are acquired through existing unsupervised keyphrase *extraction* methods (e.g., TF-IDF or TextRank (Mihalcea and Tarau, 2004)) or a self-learning algorithm. The pre-trained model will further be fine-tuned on the labeled data only. For the second approach, we propose a multi-task learning (MTL) framework¹ by jointly learning the main task of keyphrase generation based on labeled samples, and an auxiliary task of title generation (Rush et al., 2015) on unlabeled documents. Here one encoder is shared among the two tasks. Importantly, we test our proposed methods on a seq2seq framework, however, we believe they can be easily adapted to other encoder-decoder-based systems.

Extensive experiments are conducted in scientific paper domain. Results on five different datasets show that all of our semi-supervised learning-based models can uniformly significantly outperform a state-of-the-art model (Meng et al., 2017) as well as several competitive unsupervised and supervised keyphrase extraction algorithms based on F_1 and recall scores. We further carry out a cross-domain study on generating keyphrases for news articles, where our models yield better F_1 than a model trained on labeled data only. Finally, we also show that training with unlabeled samples can further produce performance gain even when a large amount of labeled data is available.

2 Related Work

Keyphrase Extraction and Generation. Early work mostly focuses on the keyphrase extraction task, and a two-step strategy is typically designed. Specifically, a large pool of candidate phrases are first extracted according to pre-defined syntactic templates (Mihalcea and Tarau, 2004; Wan and Xiao, 2008; Liu et al., 2009b, 2011) or their estimated importance scores (Hulth, 2003). In the second step, re-ranking is applied to yield the final keyphrases, based on supervised learning (Frank et al., 1999; Witten et al., 1999; Hulth, 2003; Lopez and Romary, 2010; Kim and Kan, 2009), unsupervised graph algorithms (Mihalcea and Tarau, 2004; Wan and Xiao, 2008; Bougouin et al., 2013), or topic modelings (Liu et al., 2009c, 2010). Keyphrase generation is stud-

¹We use “semi-supervised learning” as a broad term to refer to the two methods proposed in this paper.

ied in more recent work. For instance, Liu et al. (2011) propose to use statistic machine translation model to learn word-alignments between documents and keyphrases, enabling the model to generate keyphrases which do not appear in the input. Meng et al. (2017) train seq2seq-based generation models (Sutskever et al., 2014) on large-scale labeled corpora, which may not be applicable to a new domain with minimal human labels.

Neural Semi-supervised Learning. As mentioned above, though significant success has been achieved by seq2seq model in many NLP tasks (Luong et al., 2015; See et al., 2017; Dong and Lapata, 2016; Wang and Ling, 2016; Ye et al., 2018), they often rely on large amounts of labeled data, which are expensive to get. In order to mitigate the problem, semi-supervised learning has been investigated to incorporate unlabeled data for modeling training (Dai and Le, 2015; Ramachandran et al., 2017). For example, neural machine translation community has studied the usage of source-side or target-side monolingual data to improve translation quality (Gülçehre et al., 2015), where generating synthetic data (Sennrich et al., 2016; Zhang and Zong, 2016), multi-task learning (Zhang and Zong, 2016), and autoencoder-based methods (Cheng et al., 2016) are shown to be effective. Multi-task learning is also examined for sequence labeling tasks (Rei, 2017; Liu et al., 2018). In our paper, we study semi-supervised learning for keyphrase generation based on seq2seq models, which has not been explored before. Besides, we focus on leveraging source-side unlabeled articles to enhance performance with synthetic keyphrase construction or multi-task learning.

3 Neural Keyphrase Generation Model

In this section, we describe the neural keyphrase generation model built on a sequence-to-sequence model (Sutskever et al., 2014) as illustrated in Figure 2. We denote the input source document as a sequence $\mathbf{x} = x_1 \cdots x_{|\mathbf{x}|}$ and its corresponding keyphrase set as $a = \{a_i\}_{i=1}^{|a|}$, with a_i as one keyphrase.

Keyphrase Sequence Formulation. Different from the setup by Meng et al. (2017), where input article is paired with each keyphrase a_i to consist a training sample, we concatenate the keyphrases in a into a keyphrase sequence $\mathbf{y} =$

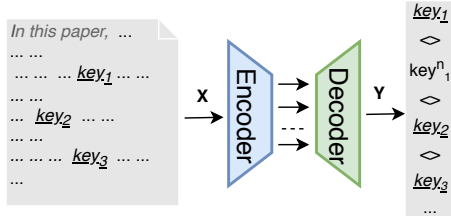


Figure 2: Neural keyphrase generation model built on sequence-to-sequence framework. Input is the document, and output is the keyphrase sequence consisting of phrases present (key_i) or absent (key_i^n) in the input.

$a_1 \diamond a_2 \diamond \dots \diamond a_{|a|}$, where \diamond is a segmenter to separate the keyphrases². With this setup, the seq2seq model is capable to generate all possible keyphrases in one sequence as well as capture the contextual information between the keyphrases from the same sequence.

Seq2Seq Attentional Model. With source document \mathbf{x} and its keyphrase sequence \mathbf{y} , an encoder encodes \mathbf{x} into context vectors, from which a decoder then generates \mathbf{y} . We set the encoder as one-layer bi-directional LSTM model and the decoder as another one-layer LSTM model (Hochreiter and Schmidhuber, 1997). The probability of generating target sequence $p(\mathbf{y}|\mathbf{x})$ is formulated as:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{|\mathbf{y}|} p(y_t|\mathbf{y}_{<t}, \mathbf{x}) \quad (1)$$

where $\mathbf{y}_{<t} = \overrightarrow{y_1} \dots \overrightarrow{y_{t-1}}$.

Let $\mathbf{h}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$ denote the hidden state vector in the encoder at time step t , which is the concatenation of forward hidden vector $\overrightarrow{\mathbf{h}}_t$ and backward hidden vector $\overleftarrow{\mathbf{h}}_t$. Specifically, $\overrightarrow{\mathbf{h}}_t = \mathbf{f}_{\text{LSTM}_e}(x_t, \overrightarrow{\mathbf{h}}_{t-1})$ and $\overleftarrow{\mathbf{h}}_t = \mathbf{f}_{\text{LSTM}_e}(x_t, \overleftarrow{\mathbf{h}}_{t+1})$, where $\mathbf{f}_{\text{LSTM}_e}$ is an LSTM unit in encoder.

Decoder hidden state is calculated as $\mathbf{s}_t = \mathbf{f}_{\text{LSTM}_d}(y_{t-1}, \mathbf{s}_{t-1})$, where $\mathbf{f}_{\text{LSTM}_d}$ is an LSTM unit in decoder. We apply global attention mechanism (Luong et al., 2015) to calculate the context vector:

$$\mathbf{c}_t = \sum_{i=1}^{|\mathbf{x}|} \alpha_{t,i} \mathbf{h}_i$$

$$\alpha_{t,i} = \frac{\exp(\mathbf{W}_{\text{att}}[\mathbf{s}_t; \mathbf{h}_i])}{\sum_{k=1}^{|\mathbf{x}|} \exp(\mathbf{W}_{\text{att}}[\mathbf{s}_t; \mathbf{h}_k])} \quad (2)$$

²We concatenate keyphrases following the original keyphrase order in the corpora, and we set \diamond as “;” in our implementation. The effect of keyphrase ordering will be studied in the future work.

Algorithm 1 Keyphrase Ranking

Input: Generated top R keyphrase sequences $\mathcal{S} = [y_1, \dots, y_R]$ ranked with generation possibility from high to low with beam search

Output: Ranked keyphrase set \mathcal{A} with importance from high to low

function KEY-RANK(\mathcal{S})

```

 $\mathcal{A} \leftarrow \text{list}()$  ▷ set  $\mathcal{A}$  as a empty list.
 $\mathcal{Q} \leftarrow \text{dict}()$  ▷ to skip keyword that is already in  $\mathcal{A}$ .
for  $y_i \in \mathcal{S}$  do
   $y_i \leftarrow y_i.\text{split}(\text{“}\diamond\text{”})$ ; ▷ split  $y_i$  by “ $\diamond$ ”
  for  $a \in y_i$  do
    ▷ if  $a$  has not been merged in  $\mathcal{A}$ , then keep it.
    if not  $\mathcal{Q}.\text{has\_key}(a)$  then
       $\mathcal{A}.\text{append}(a)$ 
       $\mathcal{Q}.\text{update}(\text{dict}(\{a : \text{“”}\}))$ 

```

where $\alpha_{t,i}$ is the attention weight; \mathbf{W}_{att} contains learnable parameters. In this paper, we omit the bias variables to save space.

The probability to predict y_t in the decoder at time step t is factorized as:

$$p_{\text{vocab}}(y_t|\mathbf{y}_{<t}, \mathbf{c}_t) = \mathbf{f}_{\text{softmax}}(\mathbf{W}_{d_1} \cdot \tanh(\mathbf{W}_{d_2}[\mathbf{s}_t; \mathbf{c}_t])) \quad (3)$$

where $\mathbf{f}_{\text{softmax}}$ is the softmax function and \mathbf{W}_{d_1} and \mathbf{W}_{d_2} are learnable parameters.

Pointer-generator Network. Similar to Meng et al. (2017), we utilize copying mechanism via pointer-generator network (See et al., 2017) to allow the decoder to directly copy words from input document, thus mitigating out-of-vocabulary (OOV) problem. At time step t , the generation probability p_{gen} is calculated as:

$$p_{\text{gen}} = \mathbf{f}_{\text{sigmoid}}(\mathbf{W}_c \mathbf{c}_t + \mathbf{W}_s \mathbf{s}_t + \mathbf{W}_y y_t) \quad (4)$$

where $\mathbf{f}_{\text{sigmoid}}$ is a sigmoid function; \mathbf{W}_c , \mathbf{W}_s and \mathbf{W}_y are learnable parameters. p_{gen} plays a role of switcher to choose to generate a word from a fixed vocabulary with probability p_{vocab} or directly copy a word from the source document with the attention distribution α_t . With a combination of a fixed vocabulary and the extended source document vocabulary, the probability to predict y_t is:

$$p(y_t) = p_{\text{gen}} p_{\text{vocab}}(y_t|\mathbf{y}_{<t}, \mathbf{c}_t) + (1 - p_{\text{gen}}) \sum_{i:y_i=y_t} \alpha_{t,i} \quad (5)$$

where if y_t does not appear in the fixed vocabulary, then the first term will be zero; and if y_t is outside source document, the second term will be zero.

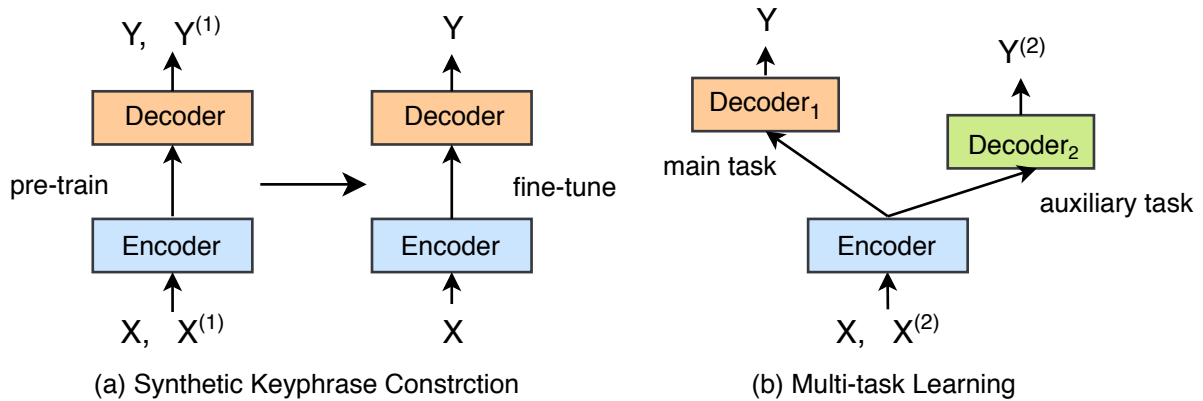


Figure 3: Our two semi-supervised learning methods which are based on (a) synthetic keyphrase construction, and (b) multi-task learning. (X, Y) represents labeled sample. $X^{(1)}$ and $X^{(2)}$ denotes unlabeled documents. $Y^{(1)}$ refers to synthetic keyphrases of $X^{(1)}$ and $Y^{(2)}$ means the title of $X^{(2)}$. For method of synthetic keyphrase construction, model will be pre-trained on the mixture of gold-standards and synthetic data, then fine-tuned on labeled data. For multi-task learning, model parameters of main task and auxiliary task will be jointly updated. Encoder parameters of the two tasks are shared.

Supervised Learning. With a labeled dataset $\mathcal{D}_p = \{\mathbf{x}_{(i)}, \mathbf{y}_{(i)}\}_{i=1}^N$, the loss function of seq2seq model is as follows:

$$\mathcal{L}(\theta) = - \sum_{i=1}^N \log p(\mathbf{y}_{(i)} | \mathbf{x}_{(i)}; \theta) \quad (6)$$

where θ contains all model parameters.

Keyphrase Inference and Ranking Strategy. Beam search is utilized for decoding, and the top R keyphrase sequences are leveraged for producing the final keyphrases. Here we use a beam size of 50, and R as 50. We propose a ranking strategy to collect the final set of keyphrases. Concretely, in sequence we collect unique keyphrases from the top ranked beams to lower ranked beams, and keyphrases in the same sequence are ordered as in the generation process. Intuitively, higher ranked sequences are likely of better quality. As for keyphrases in the same sequence, we find that more salient keyphrases are usually generated first. The ranking method is presented in Algorithm 1.

4 Semi-Supervised Learning for Keyphrase Generation

As illustrated in Figure 3, two methods are proposed to leverage abundant unlabeled data. The first is to provide synthetic keyphrases using unsupervised keyphrase extraction methods or self-learning algorithm, then mixed with labeled data

for model training, which is described in Section 4.1. Furthermore, we introduce multi-task learning that jointly generates keyphrases and the title of the document (Section 4.2). We denote the large-scale unlabeled documents as $\mathcal{D}_u = \{\mathbf{x}'_{(i)}\}_{i=1}^M$ and labeled data as $\mathcal{D}_p = \{\mathbf{x}_{(i)}, \mathbf{y}_{(i)}\}_{i=1}^N$, where $M \gg N$.

4.1 Synthetic Keyphrase Construction

The first proposed technique is to construct synthetic labeled data by assigning keyphrases for unlabeled documents, and then mix the synthetic data with human labeled data for modeling training. Intuitively, adding training samples with synthetic keyphrases has two potentially benefits: (1) the encoder is exposed to more documents in training, and (2) the decoder also benefit from additional information from identifying contextual information for keyphrases. We propose and compare two methods to extract synthetic keyphrases.

Unsupervised Learning Methods. Unsupervised learning methods on keyphrase extraction have been long studied in previous work (Mihalcea and Tarau, 2004; Wan and Xiao, 2008). Here we select two effective and widely used methods to select keyphrases on unlabeled dataset \mathcal{D}_u , which are TF-IDF and TextRank (Mihalcea and Tarau, 2004). We combine the two methods into a hybrid approach, in which we first adopt the two methods to separately select top K keyphrases from the document, we then take the union with duplicate removal. To construct the keyphrase se-

Algorithm 2 Training Procedures for Synthetic Keyphrase Construction

Input: $\mathcal{D}_p, \mathcal{D}_s, \theta$ **Output:** θ **function** PRE-TRAIN($\mathcal{D}_p, \mathcal{D}_s, \theta$) $\mathcal{D}_{p+s} \leftarrow \mathcal{D}_p \cup \mathcal{D}_s$ *Shuffle* \mathcal{D}_{p+s} randomly*Update* θ on \mathcal{D}_{p+s} until converge**function** FINE-TUNE(\mathcal{D}_p, θ)*Set* θ as the best parameters from PRE-TRAIN*Update* θ on \mathcal{D}_p until converge

quence, we concatenate the terms from TF-IDF and then from TextRank, following the corresponding ranking order. We set K as 5 in our experiments.

Self-learning Algorithm. Inspired by prior work (Zhang and Zong, 2016; Sennrich et al., 2016), we adopt self-learning algorithm to boost training data. Concretely, we first build a baseline model by training the seq2seq model on the labeled corpus \mathcal{D}_p . Then the trained baseline model is utilized to generate synthetic keyphrase sequence \mathbf{y}' given a unlabeled document \mathbf{x}' . We adopt beam search to generate the synthetic keyphrase sequences and beam size is set as 10. The top one beam is selected.

Training Procedure. After the synthetic data $\mathcal{D}_s = \{\mathbf{x}'_{(i)}, \mathbf{y}'_{(i)}\}_{i=1}^M$ is obtained by either of the aforementioned methods, we mix labeled data \mathcal{D}_p with \mathcal{D}_s to train the seq2seq model. As described in Algorithm 2, we combine \mathcal{D}_p with \mathcal{D}_s into \mathcal{D}_{p+s} and shuffle \mathcal{D}_{p+s} randomly, then we pre-train the model on \mathcal{D}_{p+s} , in which no network parameters are frozen during the training process. The best performing model is selected based on validation set, then fine-tuned on \mathcal{D}_p until converge.

4.2 Multi-task Learning with Auxiliary Task

The second approach to leverage unlabeled documents is to employ a multi-task learning framework which combines the main task of keyphrase generation with an auxiliary task through parameter sharing strategy. Similar to the model structure in Zhang and Zong (2016), our main task and the auxiliary task share an encoder network but have different decoders. Multi-task learning will benefit from the source-side information to improve the model generality of encoder.

In most domains such as scientific papers or news articles, a document usually contains a title that summarizes the core topics or ideas, with a

Dataset	TRAIN	VALID
<i>Small-scale</i>		
Document-Keyphrase	40,000	5,000
Document-SyntheticKeyphrase	400,000	N/A
Document-Title	400,000	15,000
<i>Large-scale</i>		
Document-Keyphrase	130,000	5,000

Avg. #Tokens in Train	LABELLED	SYN.	MTL
<i>Small-scale</i>			
Document	176.3	175.9	165.5
Keyphrase Sequence	23.3	23.5	N/A
Title	N/A	N/A	10.4

Table 1: Statistics of datasets used in our experiments.

similar spirit as keyphrases. We thus choose title generation as auxiliary task, which has been studied as a summarization problem (Rush et al., 2015; Colmenares et al., 2015). Let $\mathcal{D}'_u = \{\mathbf{x}'_{(i)}, \mathbf{q}_{(i)}\}_{i=1}^M$ denote the dataset which is assigned with titles for unlabeled data \mathcal{D}_u , the loss function of multi-task learning is factorized as:

$$\mathcal{L}(\theta^e, \theta_1^d, \theta_2^d) = - \sum_{i=1}^N \log p(\mathbf{y}_{(i)} | \mathbf{x}_{(i)}; \theta^e, \theta_1^d) - \sum_{i=1}^M \log p(\mathbf{q}_{(i)} | \mathbf{x}'_{(i)}; \theta^e, \theta_2^d) \quad (7)$$

where θ^e indicates encoder parameters; θ_1^d and θ_2^d are the decoder parameters.

Training Procedure. We adopt a simple alternating training strategy to switch training between the main task and the auxiliary task. Specifically, we first estimate parameters on auxiliary task with \mathcal{D}'_u for one epoch, then train model on the main task with \mathcal{D}_p (labeled dataset) for T epochs. We follow this training procedure for several times until the model of our main task converges. We set T as 3.

5 Experiments

5.1 Datasets

Our major experiments are conducted on scientific articles which have been studied in previous work (Hulth, 2003; Nguyen and Kan, 2007; Meng et al., 2017). We use the dataset from Meng et al. (2017) which is collected from various online digital libraries, e.g. ScienceDirect, ACM Digital Library, Wiley, and other portals.

As indicated in Table 1, we construct a relatively small-scale labeled dataset with 40K document-keyphrase³ pairs, and a large-scale

³Here keyphrase refers to the keyphrase sequence.

Model	INSPEC		KRAPIVIN		NUS		SEMEVAL		KP20K	
	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10
<i>Comparisons</i>										
TF-IDF	0.223	0.304 [†]	0.113	0.143	0.139	0.181	0.120	0.184 [†]	0.105	0.130
TEXTRANK	0.229	0.275	0.173	0.147	0.195	0.190	0.172	0.181	0.181	0.150
SINGLERANK	0.214	0.297	0.096	0.137	0.145	0.169	0.132	0.169	0.099	0.124
EXPANDRANK	0.211	0.295	0.096	0.136	0.137	0.162	0.135	0.163	N/A	N/A
MAUI	0.041	0.033	0.243	0.208 [†]	0.249	0.261 [†]	0.045	0.039	0.265	0.227 [†]
KEA	0.109	0.129	0.120	0.131	0.068	0.081	0.027	0.027	0.180	0.163
SEQ2SEQ-COPY	0.269	0.234	0.274	0.207	0.345	0.282	0.278	0.226	0.291	0.215
<i>Semi-supervised</i>										
SYN.UNSUPER.	0.326*	0.334*	0.283	0.239*	0.356	0.317*	0.306	0.294*	0.300*	0.245*
SYN.SELF-LEARN.	0.310*	0.301*	0.289	0.236*	0.339	0.305	0.295	0.282*	0.301*	0.240*
MULTI-TASK	0.326*	0.309*	0.296	0.240*	0.354	0.320*	0.322	0.289*	0.308*	0.243*

Table 2: Results of *present* keyphrase generation with metrics F₁@5 and F₁@10. * marks numbers that are significantly better than SEQ2SEQ-COPY ($p < 0.01$, F -test). Due to the high time perplexity, no result is reported by ExpandRank on KP20K, as done in Meng et al. (2017).

dataset of 400K unlabeled documents. Each document contains an abstract and a title of the paper. In multi-task learning setting, the auxiliary task is to generate the title from the abstract. For the validation set, we collect 5K document-keyphrase pairs for the process of pre-training and fine-tuning for methods based on synthetic data construction. For multi-task learning, we use the same 5K document-keyphrase pairs for the main task training, another 15K document-title pairs for the auxiliary task. We further conduct experiments on a 130K large-scale labeled dataset, which includes the small-scale labeled data.

Similar to Meng et al. (2017), we test our model on five widely used public datasets from the scientific domain: INSPEC (Hulth, 2003), NUS (Nguyen and Kan, 2007), KRAPIVIN (Krapivin et al., 2009), SEMEVAL-2010 (Kim et al., 2010) and KP20K (Meng et al., 2017).

5.2 Experimental Settings

Data preprocessing is implemented as in (Meng et al., 2017). The texts are first tokenized by NLTK (Bird and Loper, 2004) and lowercased, then the numbers are replaced with <digit>. We set maximal length of source text as 200, 40 for target text. Encoder and decoder both have a vocabulary size of 50K. The word embedding size is set to 128. Embeddings are randomly initialized and learned during training. The size of hidden vector is 512. Dropout rate is set as 0.3. Maximal gradient normalization is 2. Adagrad (Duchi et al., 2011) is adopted to train the model with learning rate of 0.15 and the initial accumulator rate is 0.1.

For synthetic data construction, we first use batch size of 64 for model pre-training and then re-

duce to 32 for model fine-tuning. For both training stages, after 8 epochs, learning rate be decreased with a rate of 0.5. For multi-task learning, batch size is set to 32 and learning rate is reduced to half after 20 training epochs. To build baseline seq2seq model, we set batch size as 32 and decrease learning rate after 8 epochs. For self-learning algorithm, beam size is set to 10 to generate target sequences for unlabeled data and the top one is retained.

5.3 Comparisons with Baselines

Evaluation Metrics. Following (Liu et al., 2011; Meng et al., 2017), we adopt top-N macro-averaged *precision*, *recall* and *F-measure* (F_1) for model evaluation. *Precision* means how many top-N extracted keywords are correct and *recall* means how many target keyphrases are extracted in top-N candidates. Porter Stemmer is applied before comparisons.

Baselines. We train a baseline seq2seq model on the small-scale labeled dataset. We further compare with four unsupervised learning methods: TF-IDF, TextRank (Mihalcea and Tarau, 2004), SingleRank (Wan and Xiao, 2008), ExpandRank (Wan and Xiao, 2008), and two supervised learning methods of Maui (Medelyan et al., 2009) and KEA (Witten et al., 1999). We follow Meng et al. (2017) for baselines setups.

Results. Here we show results for *present* and *absent* keyphrase generation separately⁴. From the results of present keyphrase generation as shown in Table 2, although trained on small-scale

⁴Recall that *present* means the keyphrase appears in the document, otherwise, it is *absent*.

Dataset	SEQ.	Semi-supervised		
		SYN.UN.	SYN.SELF.	MULTI.
INSPEC	0.012	0.018	0.013	0.022
KRAPIVIN	0.01	0.008	0.018	0.021
NUS	0.002	0.011	0.003	0.013
SEM EVAL	0.001	0.008	0.003	0.006
KP20K	0.01	0.016	0.013	0.021

Table 3: Results of *absent* keyphrase generation based on Recall@10.

labeled corpora, our baseline seq2seq model with copying mechanism still achieves better $F_1@5$ scores, compared to other baselines. This demonstrates that a baseline seq2seq model has learned the mapping patterns from source text to target keyphrases to some degree. However, small-scale labeled data still hinders the potential of seq2seq model according to the poor performance of $F_1@10$. By leveraging large-scale unlabeled data, our semi-supervised learning methods achieve significant improvement over seq2seq baseline, as well as exhibit the best performance in both $F_1@5$ and $F_1@10$ on almost all datasets.

We further compare seq2seq based models on the task of generating keyphrases beyond input article vocabulary. Illustrated by Table 3, semi-supervised learning significantly improves the absent generation performance, compared to the baseline seq2seq. Among our models, the multi-task learning method is more effective at generating absent keyphrases than the two methods by leveraging synthetic data. The main reason may lie in that synthetic keyphrases potentially introduce noisy annotations, while the decoder in multi-task learning setting focuses on learning from gold-standard keyphrases. We can also see that the overall performances by all models are low, due to the intrinsic difficulty of absent keyphrase generation (Meng et al., 2017). Moreover, we only employ 40K labeled data for training, which is rather limited for training. Besides, we believe better evaluation methods should be used instead of exact match, e.g., by considering paraphrases. This will be studied in the future work.

5.4 Effect of Synthetic Keyphrase Quality

In this section, we conduct experiments to further study the effect of synthetic keyphrase quality on model performance. Two sets of experiments are undertaken, one for evaluating unsupervised learning and one for self-learning algorithm.

For self-learning algorithm, we further generate

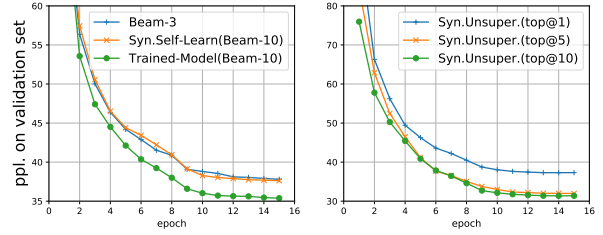


Figure 4: *Pre-training* curves with perplexity on validation set with various options for synthetic keyphrases construction. Left is for options for self-learning algorithm and right is for unsupervised learning methods.

synthetic keyphrases using following options:

- **Beam-size-3:** Based on our baseline model trained with labeled data, we use beam search with a smaller beam size of 3 to generate synthetic data⁵.
- **Trained-model:** We adopt the model which has been trained with self-learning algorithm on 40K labeled data and 400K unlabeled data, to generate the top one keyphrase sequence with beam size of 10.

For unsupervised learning method, we originally merge top-K ($K = 5$) keyphrases from TF-IDF and TextRank, here we use options where K is set as 1 or 10 to extract keyphrases:

- **Top@1:** Using TF-IDF or TextRank, we only keep top 1 extraction from each, then take the union of the two.
- **Top@10:** Similarly, we keep top 10 extracted terms from TF-IDF or TextRank, then take the union.

As illustrated in Figure 4, when models are pre-trained with synthetic keyphrases of better quality, results by “Trained-model” consistently produce better performance (i.e., lower perplexity). Similar phenomenon can be observed when “top@5” and “top@10” are applied for extraction in unsupervised learning setting. Furthermore, after models are pre-trained and then fine-tuned, the results in Figure 5 show that the difference among baselines becomes insignificant—the quality of synthetic keyphrases have limited effect on final scores. The reason might be that though synthetic

⁵We also experiment with greedy search (i.e. beam size of 1), however, unknown words are frequently generated.

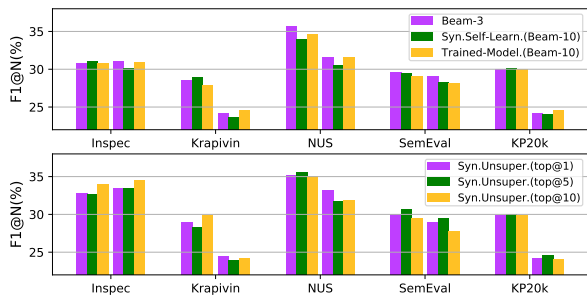


Figure 5: Effect of synthetic data quality on *present* keyphrase generation (models are pre-trained and fine-tuned) based on $F_1@5$ (left three columns) and $F_1@10$ (right three columns), on five datasets. The upper is for self-learning algorithm and the bottom is for unsupervised learning method.

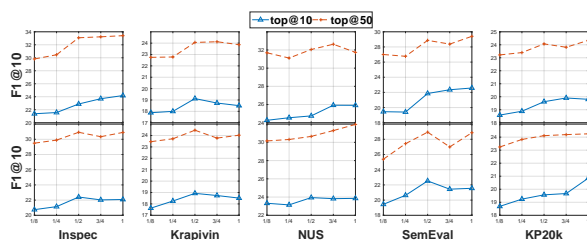


Figure 6: Effect of various amounts of unlabeled data for training on *present* keyphrase generation with $F_1@10$. Upper is for synthetic data construction method with unsupervised learning. Bottom is for multi-task learning algorithm.

keyphrases potentially introduce noisy information for decoder training, the encoder is still well trained. In addition, after fine-tuning on labeled data, the decoder acquires additional knowledge, thus leading to better performance and minimal difference among the options.

5.5 Effect of Amount of Unlabeled Data

In this section, we further evaluate whether varying the amount of unlabeled data will affect model performance. We conduct experiments based on methods of synthetic data construction with unsupervised learning and multi-task learning. We further carry out experiments with randomly selected 50K(1/8), 100K(1/4), 200K(1/2) and 300K(3/4) unlabeled documents from the pool of 400K unlabeled data. After models being trained, we adopt beam search to generate keyphrase sequences with beam size of 50. We keep top N keyphrase sequences to yield the final keyphrases using Algorithm 1. $F_1@10$ is adopted to illustrate

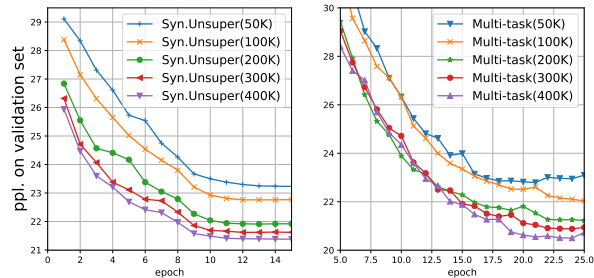


Figure 7: Perplexity on validation set with varying amounts of unlabeled data for training. Left is for fine-tuning procedure based on models trained with synthetic data constructed with unsupervised learning. Right is for multi-task learning procedure with performance on the main task.

Model	F_1	Model	F_1
<i>Our Models</i>		<i>Unsupervised</i>	
SEQ2SEQ	0.056	TF-IDF	0.270
SYN.UNSUPER.	0.083	TEXTRANK	0.097
SYN.SELF-LEARN.	0.065	SINGLERANK	0.256
MULTI-TASK	0.109	EXPANDRANK	0.269

Table 4: Results of keyphrase generation for news from DUC dataset with F_1 . Results of unsupervised learning methods are adopted from Hasan and Ng (2010).

the model performances. N is set as 10 or 50.

The *present* keyphrase generation results are shown in Figure 6, from which we can see that when increasing the amount of unlabeled data, model performance is further improved. This is because additional unlabeled data can provide with more evidence on linguistic or context features and thus make the model, especially the encoder, have better generalizability. This finding echoes with the training procedure illustrated in Figure 7, where more unlabeled data uniformly leads to better performance. Therefore, we believe that leveraging more unlabeled data for model training can boost model performance.

5.6 A Pilot Study for Cross-Domain Test

Up to now, we have demonstrated the effectiveness of leveraging unlabeled data for in-domain experiments, but is it still effective when being tested on a different domain? We thus carry out a pilot cross-domain test on news articles. The widely used DUC dataset (Wan and Xiao, 2008) is utilized, consisting of 308 articles with 2,048 labeled keyphrases.

The experimental results are shown in Table 4 which indicate that: 1) though trained on scientific papers, our models still have the ability to gener-

Model	INSPEC		KRAPIVIN		NUS		SEMEVAL		KP20K	
	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10	F ₁ @5	F ₁ @10
SEQ2SEQ-COPY	0.34	0.329	0.308	0.251	0.36	0.327	0.301	0.285	0.318	0.251
<i>Semi-supervised</i>										
SYN.UNSUPER.	0.338	0.34	0.316	0.255	0.365	0.335	0.337	0.308	0.322	0.261*
SYN.SELF-LEARN.	0.33	0.326	0.304	0.255	0.359	0.336	0.304	0.304	0.321	0.263*
MULTI-TASK	0.328	0.318	0.323	0.254	0.365	0.326	0.319	0.312	0.328*	0.264*

Table 5: Results of *present* keyphrase generation on large-scale labeled data with F₁@5 and F₁@10. * indicates significant better performance than SEQ2SEQ-COPY with $p < 0.01$ (F -test).

ate keyphrases for news articles, illustrating that our models have learned some universal features between the two domains; and 2) semi-supervised learning by leveraging unlabeled data improves the generation performances more, indicating that our proposed method is reasonably effective when being tested on cross-domain data. Though unsupervised methods are still superior, for future work, we can leverage unlabeled out-of-domain corpora to improve cross-domain keyphrase generation performance, which could be a promising direction for domain adaption or transfer learning.

Dataset	SEQ.	SYN.UN.	<i>Semi-supervised</i>	
			SYN.SELF.	MULTI.
INSPEC	0.021	0.024	0.032	0.033
KRAPIVIN	0.02	0.031	0.043	0.047
NUS	0.009	0.026	0.024	0.036
SEMEVAL	0.011	0.014	0.015	0.02
KP20K	0.021	0.034	0.039	0.046

Table 6: Performance on *absent* keyphrase generation by Recall@10 with large-scale labeled training data.

5.7 Training on Large-scale Labeled Data

Finally, it would be interesting to study whether unlabeled data can still improve performance when the model is trained on a larger scaled labeled data. We conduct experiments on a larger labeled dataset with 130K pairs, along with the 400K unlabeled data. Here the baseline seq2seq model is built on the 130K dataset.

From the *present* keyphrase generation results in Table 5, it can be seen that unlabeled data is still helpful for model training on a large-scale labeled dataset. This implies that we can also leverage unlabeled data to enhance generation performance even in a resource-rich setting. Referring to the *absent* keyphrase generation results shown in Table 6, semi-supervised learning also boosts the scores. From Table 6, training on large-scale labeled data, *absent* generation is significantly improved, compared to being trained on a small-scale labeled data (see Table 3).

6 Conclusion and Future Work

In this paper, we presented a semi-supervised learning framework that leverages unlabeled data for keyphrase generation built upon seq2seq models. We introduced synthetic keyphrases construction algorithm and multi-task learning to effectively leverage abundant unlabeled documents. Extensive experiments demonstrated the effectiveness of our methods, even in scenario where large-scale labeled data is available.

For future work, we will 1) leverage unlabeled data to study domain adaptation or transfer learning for keyphrase generation; and 2) investigate novel models to improve *absent* keyphrase generation when limited labeled data is available based on semi-supervised learning.

Acknowledgements

This research is based upon work supported in part by National Science Foundation through Grants IIS-1566382 and IIS-1813341, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of 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. We thank three anonymous reviewers for their insightful suggestions on various aspects of this work.

References

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 1162–1170.

- Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, July 21-26, 2004 - Poster and Demonstration*.
- Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 543–551.
- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- Carlos A. Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. HEADS: headline generation as sequence prediction using an abstract feature-rich space. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 133–142.
- Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087.
- Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.
- Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 668–673.
- Çağlar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.
- Khaled M Hammouda, Diego N Matute, and Mohamed S Kamel. 2005. Corephrase: Keyphrase extraction for document clustering. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 265–274. Springer.
- Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 365–373.
- Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1262–1273.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2003, Sapporo, Japan, July 11-12, 2003*.
- Su Nam Kim and Min-Yen Kan. 2009. Re-examining automatic keyphrase extraction approaches in scientific articles. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, MWE@IJCNLP 2009, Singapore, August 6, 2009*, pages 9–16.
- Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 21–26.
- Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction. Technical report, University of Trento.
- Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009a. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 620–628. Association for Computational Linguistics.
- Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009b. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pages 620–628.

- Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.
- Zhiyuan Liu, Xinxiong Chen, Yabin Zheng, and Maosong Sun. 2011. Automatic keyphrase extraction by bridging vocabulary gap. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011*, pages 135–144.
- Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 366–376.
- Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009c. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 257–266.
- Patrice Lopez and Laurent Romary. 2010. HUMB: automatic key term extraction from scientific articles in GROBID. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 248–251.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.
- Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1318–1327.
- Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. 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 582–592.
- 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.
- Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, 10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007, Proceedings*, pages 317–326.
- Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 383–391.
- Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. 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 2121–2130.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389.
- 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.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.
- Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336.
- Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 855–860.

- Lu Wang and Claire Cardie. 2013. Domain-Independent Abstract Generation for Focused Meeting Summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.
- Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 47–57.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.
- Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. KEA: practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA*, pages 254–255.
- Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1854–1864.
- Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1535–1545.
- Yongzheng Zhang, A. Nur Zincir-Heywood, and Evangelos E. Milios. 2004. World wide web site summarization. *Web Intelligence and Agent Systems*, 2(1):39–53.