

One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases

Xingdi Yuan^{†*} Tong Wang^{†*} Rui Meng^{†*} Khushboo Thaker[‡]
 Peter Brusilovsky[‡] Daqing He[‡] Adam Trischler[†]

[†]Microsoft Research, Montréal

[‡]School of Computing and Information, University of Pittsburgh
 {eric.yuan, tong.wang}@microsoft.com rui.meng@pitt.edu

Abstract

Different texts shall by nature correspond to different number of keyphrases. This desideratum is largely missing from existing neural keyphrase generation models. In this study, we address this problem from both modeling and evaluation perspectives.

We first propose a recurrent generative model that generates multiple keyphrases as delimiter-separated sequences. Generation diversity is further enhanced with two novel techniques by manipulating decoder hidden states. In contrast to previous approaches, our model is capable of generating diverse keyphrases and controlling number of outputs.

We further propose two evaluation metrics tailored towards the variable-number generation. We also introduce a new dataset (*STACKEX*) that expands beyond the only existing genre (i.e., academic writing) in keyphrase generation tasks. With both previous and new evaluation metrics, our model outperforms strong baselines on all datasets.

1 Introduction

Keyphrase generation is the task of automatically predicting keyphrases given a source text. Desired keyphrases are often multi-word units that summarize the high-level meaning and highlight certain important topics or information of the source text. Consequently, models that can successfully perform this task should be capable of not only distilling high-level information from a document, but also locating specific, important snippets therein.

To make the problem even more challenging, a keyphrase may or may not be a substring of the source text (i.e., it may be *present* or *absent*). Moreover, a given source text is usually associated with

Dataset	#Train	#Valid	#Test	Mean	Var	%Pre
KP20K	≈514k	≈20k	≈20k	5.3	14.2	63.3%
INSPEC	–	1500	500	9.6	22.4	78.5%
KRAPIVIN	–	1844	460	5.2	6.6	56.2%
NUS	–	–	211	11.5	64.6	51.3%
SEMEVAL	–	144	100	15.7	15.1	44.5%
STACKEX	≈298k	≈16k	≈16k	2.7	1.4	57.5%

Table 1: Statistics of various datasets. Mean and Var indicate the mean and variance of target phrase numbers, %Pre denotes percentage of present keyphrases.

a *set* of multiple keyphrases. Thus, keyphrase generation is an instance of the set generation problem, where both the size of the set and the size (i.e., the number of tokens in a phrase) of each element can vary depending on the source.

Similar to summarization, keyphrase generation is often formulated as a sequence-to-sequence (Seq2Seq) generation task in most prior studies (Meng et al., 2017; Chen et al., 2018a; Ye and Wang, 2018; Chen et al., 2018b). Conditioned on a source text, Seq2Seq models generate phrases individually or as a longer sequence jointed by delimiting tokens. Since standard Seq2Seq models generate only one sequence at a time, thus to generate multiple phrases, a common approach is to over-generate using beam search (Reddy et al., 1977) with a large beam width. Models are then evaluated by taking a *fixed* number of top predicted phrases (typically 5 or 10) and comparing them against the ground truth keyphrases.

Though this approach has achieved good empirical results, we argue that it suffers from two major limitations. Firstly, models that use beam search to generate multiple keyphrases generally lack the ability to determine the dynamic number of keyphrases needed for different source texts. Meanwhile, the parallelism in beam search also fails to model the inter-relation among the generated phrases, which can often result in diminished diver-

* These authors contributed equally. The order is determined by a fidget spinner.

sity in the output. Although certain existing models take output diversity into consideration during training (Chen et al., 2018a; Ye and Wang, 2018), the effort is significantly undermined during decoding due to the reliance on over-generation and phrase ranking with beam search.

Secondly, the current evaluation setup is rather problematic, since existing studies attempt to match a fixed number of outputs against a variable number of ground truth keyphrases. Empirically, the number of keyphrases can vary drastically for different source texts, depending on a plethora of factors including the length or genre of the text, the granularity of keyphrase annotation, etc. For the several commonly used keyphrase generation datasets, for example, the average number of keyphrases per data point can range from 5.3 to 15.7, with variances sometimes as large as 64.6 (Table 1). Therefore, using an arbitrary, fixed number k to evaluate entire datasets is not appropriate. In fact, under this evaluation setup, the F1 score for the oracle model on the KP20K dataset is 0.858 for $k = 5$ and 0.626 for $k = 10$, which apparently poses serious normalization issues as evaluation metrics.

To overcome these problems, we propose novel decoding strategies and evaluation metrics for the keyphrase generation task. The main contributions of this work are as follows:

1. We propose a Seq2Seq based keyphrase generation model capable of generating diverse keyphrases and controlling number of outputs.
2. We propose **new metrics** based on commonly used F_1 score under the hypothesis of *variable-size* outputs from models, which results in improved empirical characteristics over previous metrics based on a fixed k .
3. An additional contribution of our study is the introduction of a **new dataset** for keyphrase generation: STACKEX.

With its marked difference in genre, we expect the dataset to bring added heterogeneity to keyphrase generation evaluation.

2 Related Work

2.1 Keyphrase Extraction and Generation

Traditional keyphrase extraction has been studied extensively in past decades. In most existing literature, keyphrase extraction has been formulated as a two-step process. First, lexical features such

as part-of-speech tags are used to determine a list of phrase candidates by heuristic methods (Witten et al., 1999; Liu et al., 2011; Wang et al., 2016; Yang et al., 2017). Second, a ranking algorithm is adopted to rank the candidate list and the top ranked candidates are selected as keyphrases. A wide variety of methods were applied for ranking, such as bagged decision trees (Medelyan et al., 2009; Lopez and Romary, 2010), Multi-Layer Perceptron, Support Vector Machine (Lopez and Romary, 2010) and PageRank (Mihalcea and Tarau, 2004; Le et al., 2016; Wan and Xiao, 2008). Recently, Zhang et al. (2016); Luan et al. (2017); Golapalli et al. (2017) used sequence labeling models to extract keyphrases from text; Subramanian et al. (2017) used Pointer Networks to point to the start and end positions of keyphrases in a source text; Sun et al. (2019) leveraged graph neural networks to extract keyphrases.

The main drawback of keyphrase extraction is that sometimes keyphrases are absent from the source text, thus an extractive model will fail predicting those keyphrases. Meng et al. (2017) first proposed the CopyRNN, a neural model that both generates words from vocabulary and points to words from the source text. Based on the CopyRNN architecture, Chen et al. (2018a); Zhao and Zhang (2019) leveraged attention to help reducing duplication and improving coverage. Ye and Wang (2018) proposed semi-supervised methods by leveraging both labeled and unlabeled data for training. Chen et al. (2018b); Ye and Wang (2018) proposed to use structure information (e.g., title of source text) to improve keyphrase generation performance. Chan et al. (2019) introduced RL to the keyphrase generation task. Chen et al. (2019a) retrieved similar documents from training data to help producing more accurate keyphrases.

2.2 Sequence to Sequence Generation

Sequence to Sequence (Seq2Seq) learning was first introduced by Sutskever et al. (2014); together with the soft attention mechanism of (Bahdanau et al., 2014), it has been widely used in natural language generation tasks. Gülçehre et al. (2016); Gu et al. (2016) used a mixture of generation and pointing to overcome the problem of large vocabulary size. Paulus et al. (2017); Zhou et al. (2017) applied Seq2Seq models on summary generation tasks, while Du et al. (2017); Yuan et al. (2017) generated questions conditioned on docu-

ments and answers from machine comprehension datasets. Seq2Seq was also applied on neural sentence simplification (Zhang and Lapata, 2017) and paraphrase generation tasks (Xu et al., 2018).

3 Model Architecture

Given a piece of source text, our objective is to generate a variable number of multi-word phrases. To this end, we opt for the sequence-to-sequence (Seq2Seq) (Sutskever et al., 2014) framework as the basis of our model, combined with attention and pointer softmax mechanisms in the decoder.

Since each data example contains one source text sequence and multiple target phrase sequences (dubbed `ONE2MANY`, and each sequence can be of multi-word), two paradigms can be adopted for training Seq2Seq models. The first one (Meng et al., 2017) is to divide each `ONE2MANY` data example into multiple `ONE2ONE` examples, and the resulting models (e.g., CopyRNN) can generate one phrase at once and must rely on beam search technique to produce more unique phrases.

To enable models to generate multiple phrases and control the number to output, we propose the second training paradigm `ONE2SEQ`, in which we concatenate multiple phrases into a single sequence with a delimiter $\langle \text{sep} \rangle$, and this concatenated sequence is then used as the target for sequence generation during training. An overview of the model’s structure is shown in Figure 1.¹

Notations

In the following subsections, we use w to denote input text tokens, x to denote token embeddings, h to denote hidden states, and y to denote output text tokens. Superscripts denote time-steps in a sequence, and subscripts e and d indicate whether a variable resides in the encoder or the decoder of the model, respectively. The absence of a superscript indicates multiplicity in the time dimension. L refers to a linear transformation and L^f refers to it followed by a non-linear activation function f . Angled brackets, $\langle \rangle$, denote concatenation.

3.1 Sequence to Sequence Generation

We develop our model based on the standard Seq2Seq (Sutskever et al., 2014) model with attention mechanism (Bahdanau et al., 2014) and

pointer softmax (Gülçehre et al., 2016). Due to space limit, we describe this basic Seq2Seq model in Appendix A.

3.2 Mechanisms for Diverse Generation

There are usually multiple keyphrases for a given source text because each keyphrase represents certain aspects of the text. Therefore keyphrase diversity is desired for the keyphrase generation. Most previous keyphrase generation models generate multiple phrases by over-generation, which is highly prone to generate similar phrases due to the nature of beam search. Given our objective to generate variable numbers of keyphrases, we need to adopt new strategies for achieving better diversity in the output.

Recall that we represent variable numbers of keyphrases as delimiter-separated sequences. One particular issue we observed during error analysis is that the model tends to produce identical tokens following the delimiter token. For example, suppose a target sequence contains n delimiter tokens at time-steps t_1, \dots, t_n . During training, the model is rewarded for generating the same delimiter token at these time-steps, which presumably introduces much homogeneity in the corresponding decoder states $h_d^{t_1}, \dots, h_d^{t_n}$. When these states are subsequently used as inputs at the time-steps immediately following the delimiter, the decoder naturally produces highly similar distributions over the following tokens, resulting in identical tokens being decoded. To alleviate this problem, we propose two plug-in components for the sequential generation model.

3.2.1 Semantic Coverage

We propose a mechanism called *semantic coverage* that focuses on the semantic representations of generated phrases. Specifically, we introduce another uni-directional recurrent model GRU_{SC} (dubbed *target encoder*) which encodes decoder-generated tokens y^τ , where $\tau \in [0, t)$, into hidden states h_{SC}^t . This state is then taken as an extra input to the decoder GRU, modifying equation of the decoder GRU to:

$$h_d^t = \text{GRU}_d(\langle x_d^t, h_{\text{SC}}^t \rangle, h_d^{t-1}). \quad (1)$$

If the target encoder were to be updated with the training signal from generation (i.e., backpropagating error from the decoder GRU to the target encoder), the resulting decoder is essentially a 2-layer GRU with residual connections. Instead, inspired

¹We release the code, datasets and model outputs for reproducing our results in <https://github.com/memray/OpenNMT-kpg-release>.

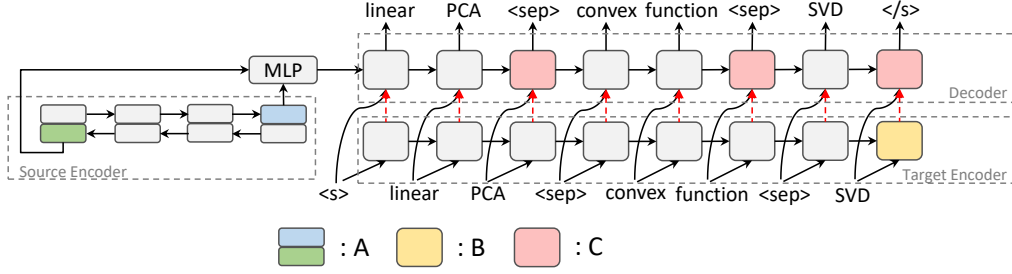


Figure 1: The architecture of the proposed model for improving keyphrase diversity. A represents last states of a bi-directional source encoder; B represents the last state of target encoder; C indicates decoder states where target tokens are either delimiters or end-of-sentence tokens. During orthogonal regularization, all C states are used; during target encoder training, we maximize mutual information between states A with B . Red dash arrow indicates a detached path, i.e., no back-propagation through such path.

by previous representation learning works (Logeswaran and Lee, 2018; van den Oord et al., 2018; Hjelm et al., 2018), we train the target encoder in an self-supervised fashion (Figure 1). Specifically, due to the autoregressive nature of the RNN-based decoder, we follow Contrastive Predictive Coding (CPC) (van den Oord et al., 2018), where a Noise-Contrastive Estimation (NCE) loss is used to maximize a lower bound on mutual information. That is, we extract target encoder’s final hidden state vector h_{SC}^M , where M is the length of target sequence, and use it as a general representation of the target phrases. We train by maximizing the mutual information between these phrase representations and the final state of the source encoder h_e^T as follows. For each phrase representation vector h_{SC}^M , we take the encodings $H_e^T = \{h_{e,1}^T, \dots, h_{e,N}^T\}$ of N different source texts, where $h_{e,true}^T$ is the encoder representation for the current source text, and the remaining $N - 1$ are negative samples (sampled at random) from the training data. The target encoder is trained to minimize the classification loss:

$$\mathcal{L}_{SC} = -\log \frac{g(h_{e,true}^T, h_{SC}^M)}{\sum_{i \in [1, N]} g(h_{e,i}^T, h_{SC}^M)}, \quad (2)$$

$$g(h_a, h_b) = \exp(h_a^\top B h_b)$$

where B is bi-linear transformation.

The motivation here is to constrain the overall representation of generated keyphrase to be semantically close to the overall meaning of the source text. With such representations as input to the decoder, the semantic coverage mechanism can potentially help to provide useful keyphrase information and guide generation.

3.2.2 Orthogonal Regularization

We also propose orthogonal regularization, which explicitly encourages the delimiter-generating decoder states to be different from each other. This is inspired by Bousmalis et al. (2016), who use orthogonal regularization to encourage representations across domains to be as distinct as possible. Specifically, we stack the decoder hidden states corresponding to delimiters together to form matrix $H = \langle h_d^{t_1}, \dots, h_d^{t_n} \rangle$ and use the following equation as the orthogonal regularization loss:

$$\mathcal{L}_{OR} = \left\| H^\top H \odot (\mathbf{1} - I_n) \right\|_2, \quad (3)$$

where H^\top is the matrix transpose of H , I_n is the identity matrix of rank n , \odot indicates element wise multiplication, $\|M\|_2$ indicates L^2 norm of each element in a matrix M . This loss function prefers orthogonality among the hidden states $h_d^{t_1}, \dots, h_d^{t_n}$ and thus improves diversity in the tokens following the delimiters.

3.2.3 Training Loss

We adopt the widely used negative log-likelihood loss in our sequence generation model, denoted as \mathcal{L}_{NLL} . The overall loss we use for optimization is:

$$\mathcal{L} = \mathcal{L}_{NLL} + \lambda_{OR} \cdot \mathcal{L}_{OR} + \lambda_{SC} \cdot \mathcal{L}_{SC}, \quad (4)$$

where λ_{OR} and λ_{SC} are hyper-parameters.

3.3 Decoding Strategies

According to different task requirements, various decoding methods can be applied to generate the target sequence y . Prior studies Meng et al. (2017); Yang et al. (2017) focus more on generating excessive number of phrases by leveraging beam

search to proliferate the output phrases. In contrast, models trained under `ONE2SEQ` paradigm are capable of determining the proper number of phrases to output. In light of previous research in psychology (Van Zandt and Townsend, 1993; Forster and Bednall, 1976), we name these two decoding/search strategies as Exhaustive Decoding and Self-terminating Decoding, respectively, due to their resemblance to the way humans behave in serial memory tasks. Simply speaking, the major difference lies in whether a model is capable of controlling the number of phrases to output. We describe the detailed decoding strategies used in this study as follows:

3.3.1 Exhaustive Decoding

As traditional keyphrase tasks evaluate models with a fixed number of top-ranked predictions (say F-score @5 and @10), existing keyphrase generation studies have to over-generate phrases by means of beam search (commonly with a large beam size, e.g., 150 and 200 in (Chen et al., 2018b; Meng et al., 2017), respectively), a heuristic search algorithm that returns K approximate optimal sequences. For the `ONE2ONE` setting, each returned sequence is a unique phrase itself. But for `ONE2SEQ`, each produced sequence contains several phrases and additional processes (Ye and Wang, 2018) are needed to obtain the final unique (ordered) phrase list.

It is worth noting that the time complexity of beam search is $O(Bm)$, where B is the beam width, and m is the maximum length of generated sequences. Therefore the exhaustive decoding is generally very computationally expensive, especially for `ONE2SEQ` setting where m is much larger than in `ONE2ONE`. It is also wasteful as we observe that less than 5% of phrases generated by `ONE2SEQ` models are unique.

3.3.2 Self-terminating Decoding

An innate characteristic of keyphrase tasks is that the number of keyphrases varies depending on the document and dataset genre, therefore dynamically outputting a variable number of phrases is a desirable property for keyphrase generation models². Since our model is trained to generate a variable number of phrases as a single sequence joined by delimiters, we can obtain multiple phrases by simply decoding a single sequence for each given

source text. The resulting model thus implicitly performs the additional task of dynamically estimating the proper size of the target phrase set: once the model believes that an adequate number of phrases have been generated, it outputs a special token $\langle /s \rangle$ to terminate the decoding process.

One notable attribute of the self-terminating decoding strategy is that, by generating a set of phrases in a single sequence, the model conditions its current generation on all previously generated phrases. Compared to the exhaustive strategy (i.e., phrases being generated independently by beam search in parallel), our model can model the dependency among its output in a more explicit fashion. Additionally, since multiple phrases are decoded as a single sequence, decoding can be performed more efficiently than exhaustive decoding by conducting greedy search or beam search on only the top-scored sequence.

4 Evaluating Keyphrase Generation

Formally, given a source text, suppose that a model predicts a list of unique keyphrases $\hat{\mathcal{Y}} = (\hat{y}_1, \dots, \hat{y}_m)$ ordered by the quality of the predictions \hat{y}_i , and that the ground truth keyphrases for the given source text is the oracle set \mathcal{Y} . When only the top k predictions $\hat{\mathcal{Y}}_{:k} = (\hat{y}_1, \dots, \hat{y}_{\min(k,m)})$ are used for evaluation, *precision*, *recall*, and F_1 score are consequently conditioned on k and defined as:

$$\begin{aligned} P@k &= \frac{|\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}|}{|\hat{\mathcal{Y}}_{:k}|}, & R@k &= \frac{|\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}|}{|\mathcal{Y}|}, \\ \mathbf{F}_1@k &= \frac{2 * P@k * R@k}{P@k + R@k}. \end{aligned} \quad (5)$$

As discussed in Section 1, the number of generated keyphrases used for evaluation can have a critical impact on the quality of the resulting evaluation metrics. Here we compare three choices of k and the implications on keyphrase evaluation for each choice:

- $\mathbf{F}_1@k$: where k is a pre-defined constant (usually 5 or 10). Due to the high variance of the number of ground truth keyphrases, it is often that $|\hat{\mathcal{Y}}_{:k}| \leq k < |\mathcal{Y}|$, and thus $R@k$ — and in turn $F_1@k$ — of an oracle model can be smaller than 1. This undesirable property is unfortunately prevalent in the evaluation metrics adopted by all existing keyphrase generation studies to our knowledge.

A simple remedy is to set k as a variable number which is specific to each data example. Here we define two new metrics:

²Note this is fundamentally different from other NLG tasks. In specific, the number of keyphrases is variable, the length of each keyphrase is also variable.

	Kp20K			Inspec			Krapivin			NUS			SemEval		
Model	$F_1@5$	$F_1@10$	$F_1@O$	$F_1@5$	$F_1@10$	$F_1@O$	$F_1@5$	$F_1@10$	$F_1@O$	$F_1@5$	$F_1@10$	$F_1@O$	$F_1@5$	$F_1@10$	$F_1@O$
Abstractive Neural															
CopyRNN (Meng et al.)	32.8	25.5	-	<u>29.2</u>	<u>33.6</u>	-	30.2	25.2	-	34.2	31.7	-	29.1	29.6	-
CopyRNN*	31.7	27.3	33.5	24.4	28.9	29.0	30.5	26.6	32.5	37.6	<u>35.2</u>	40.6	31.8	31.8	<u>31.7</u>
CorrRNN (Chen et al.)	-	-	-	-	-	-	31.8	27.8	-	35.8	33.0	-	<u>32.0</u>	<u>32.0</u>	-
ParaNetT + CoAtt (Zhao and Zhang)	36.0	28.9	-	29.6	35.7	-	<u>32.9</u>	28.2	-	36.0	35.0	-	31.1	31.2	-
catSeqTG-2RF1 [†] (Chan et al.)	32.1	-	<u>35.7</u>	25.3	-	28.0	30.0	-	<u>34.8</u>	<u>37.5</u>	-	25.5	28.7	-	29.8
KG-KE-KR-M [†] (Chen et al.)	31.7	28.2	38.8	25.7	28.4	<u>31.4</u>	27.2	25.0	31.7	28.9	28.6	38.4	20.2	22.3	30.3
CatSeq (Ours)	31.4	27.3	31.9	29.0	30.0	<u>30.7</u>	30.7	27.4	32.4	35.9	34.9	<u>38.3</u>	30.2	30.6	31.0
CatSeqD (Ours)	34.8	<u>29.8</u>	<u>35.7</u>	27.6	33.3	33.1	32.5	<u>28.5</u>	37.1	37.4	36.6	40.6	32.7	35.2	35.7
Extractive IR															
Tfidf (Hasan and Ng)	7.2	9.4	6.3	16.0	24.4	20.8	6.7	9.3	6.8	11.2	14.0	12.2	8.8	14.7	11.3
TextRank (Mihalcea and Tarau)	18.1	15.1	18.4	28.6	33.9	33.5	18.5	16.0	21.1	23.0	21.6	23.8	21.7	22.6	22.9
KEA (Witten et al.)	4.6	4.4	5.1	2.2	2.2	2.2	1.8	1.7	1.7	7.3	7.1	8.1	6.8	6.5	6.6
Maui (Medelyan et al.)	0.5	0.5	0.4	3.5	4.6	3.9	0.5	0.7	0.6	0.4	0.6	0.6	1.1	1.4	1.1
Extractive Neural															
DivGraphPointer (Sun et al.)	36.8	29.2	-	38.6	41.7	-	46.0	40.2	-	40.1	38.9	-	36.3	29.7	-
w/ Additional Data															
Semi-Multi (Ye and Wang)	32.8	26.4	-	32.8	31.8	-	32.3	25.4	-	36.5	32.6	-	31.9	31.2	-
TG-Net (Chen et al.)	37.2	31.5	-	31.5	38.1	-	34.9	29.5	-	40.6	37.0	-	31.8	32.2	-

Table 2: Performance of present keyphrase prediction on scientific publications datasets. Best/second-best performing score in each column is highlighted with bold/underline. We also list results from literature where models that are not directly comparable (i.e., models leverage additional data and pure extractive models). Note model names with [†] represent its $F_1@O$ is computed by us using existing works’ released keyphrase predictions.³

- $F_1@O$: O denotes the number of oracle (ground truth) keyphrases. In this case, $k = |Y|$, which means for each data example, the number of predicted phrases taken for evaluation is the same as the number of ground truth keyphrases.

- $F_1@M$: M denotes the number of predicted keyphrases. In this case, $k = |\hat{Y}|$ and we simply take all the predicted phrases for evaluation without truncation.

By simply extending the *constant* number k to different variables accordingly, both $F_1@O$ and $F_1@M$ are capable of reflecting the nature of *variable* number of phrases for each document, and a model can achieve the maximum F_1 score of 1.0 if and only if it predicts the exact same phrases as the ground truth. Another merit of $F_1@O$ is that it is independent from model outputs, therefore we can use it to compare existing models.

5 Datasets and Experiments

In this section, we report our experiment results on multiple datasets and compare with existing models. We use **CatSeq** to refer to the delimiter-concatenated sequence-to-sequences model de-

scribed in Section 3; **CatSeqD** refers to the model augmented with orthogonal regularization and semantic coverage mechanism.

To construct target sequences for training **CatSeq** and **CatSeqD**, ground truth keyphrases are sorted by their order of first occurrence in the source text. Keyphrases that do not appear in the source text are appended to the end. This order may guide the attention mechanism to attend to source positions in a smoother way. Implementation details can be found in Appendix D. As for the pre-processing and evaluation, we follow the same steps as in (Meng et al., 2017). More details are provide in Appendix E for reproducing our results.

We include a set of existing models (Meng et al., 2017; Chen et al., 2018a; Chan et al., 2019; Zhao and Zhang, 2019; Chen et al., 2019a) as baselines, they all share same behavior of abstractive keyphrase generation with our proposed model. Specially for computing existing model’s scores with our proposed new metrics ($F_1@O$ and $F_1@M$), we implemented our own version of CopyRNN (Meng et al., 2017) based on their open sourced code, denoted as CopyRNN*. We also report the scores of models from Chan et al. and Chen et al. based on their publicly released outputs.

We also include a set of models that use similar strategies but can not directly compare with.

³We acknowledge that $F_1@O$ scores of Chan et al. (2019) and Chen et al. (2019a) might be not completely comparable with ours. This is due to additional post-processing and filtering methods might have been applied in different work. We elaborate the data pre-processing and evaluation protocols used in this work in Appendix E.

	Present			Absent	
Model	$F_1@5$	$F_1@10$	$F_1@O$	$R@10$	$R@50$
Tfidf	8.0	8.9	5.2	-	-
TextRank	12.1	10.1	11.6	-	-
KEA	4.9	4.8	5.3	-	-
Maui	35.8	23.3	51.8	-	-
CopyRNN*	44.2	30.3	66.2	48.8	66.0
CatSeq	<u>48.3</u>	45.5	63.5	40.7	42.2
CatSeqD	48.7	43.9	<u>65.6</u>	54.8	<u>65.7</u>

Table 3: Model performance on STACKEX dataset.

This includes four non-neural extractive models: Tfidf (Hasan and Ng, 2010), TextRank (Mihalcea and Tarau, 2004), KEA (Witten et al., 1999), and Maui (Medelyan et al., 2009); one neural extractive model (Sun et al., 2019); and two neural models that use additional data (e.g., title) (Ye and Wang, 2018; Chen et al., 2019b).

In Section 5.3, we apply the self-terminating decoding strategy. Since no existing model supports such decoding strategy, we only report results from our proposed models. They can be used for comparison in future studies.

5.1 Experiments on Scientific Publications

Our first dataset consists of a collection of scientific publication datasets, namely KP20K, INSPEC, KRAPIVIN, NUS, and SEMEVAL, that have been widely used in existing literature (Meng et al., 2017; Chen et al., 2018a; Ye and Wang, 2018; Chen et al., 2018b; Chan et al., 2019; Zhao and Zhang, 2019; Chen et al., 2019a; Sun et al., 2019). KP20K, for example, was introduced by Meng et al. (2017) and comprises more than half a million scientific publications. For each article, the abstract and title are used as the source text while the author keywords are used as target. The other four datasets contain much fewer articles, and thus used to test transferability of our model.

We report our model’s performance on the *present*-keyphrase portion of the KP20K dataset in Table 2.⁴ To compare with previous works, we provide compute $F_1@5$ and $F_1@10$ scores. The new proposed $F_1@O$ metric indicates consistent ranking with $F_1@5/10$ for most cases. Due to its target number sensitivity, we find that its value is closer to $F_1@5$ for KP20K and KRAPIVIN where average target keyphrases is less and closer to $F_1@10$ for the other three datasets.

⁴We show experiment results on absent data in Appendix B.

	KP20K		STACKEX	
Model	$F_1@O$	$F_1@M$	$F_1@O$	$F_1@M$
Greedy Search				
CatSeq	<u>33.1</u>	32.4	<u>59.2</u>	56.3
CatSeqD	33.4	33.9	59.6	59.3
Top Ranked Sequence in Beam Search				
CatSeq	24.3	25.1	52.4	52.7
CatSeqD	31.9	<u>33.4</u>	56.5	<u>57.0</u>

Table 4: $F_1@O$ and $F_1@M$ when generating variable number of keyphrases (self-terminating decoding).

From the result we can see that our **CatSeqD** outperform *existing abstractive models* on most of the datasets. Our implemented CopyRNN* achieves better or comparable performance against the original model, and on NUS and SemEval the advantage is more salient.

As for the proposed models, both **CatSeq** and **CatSeqD** yield comparable results to CopyRNN, indicating that ONE2SEQ paradigm can work well as an alternative option for the keyphrase generation task. **CatSeqD** outperforms **CatSeq** on all metrics, suggesting the semantic coverage and orthogonal regularization help the model to generate higher quality keyphrases and achieve better generalizability. To our surprise, on the metric $F_1@10$ for KP20K and KRAPIVIN (average number of keyphrases is only 5), where high-recall models like CopyRNN are more favored, **CatSeqD** is still able to outperform ONE2ONE baselines, indicating that the proposed mechanisms for diverse generation are effective.

5.2 Experiments on The STACKEX Dataset

Inspired by the StackLite tag recommendation task on Kaggle, we build a new benchmark based on the public StackExchange data⁵. We use questions with titles as source, and user-assigned tags as target keyphrases. We provide details regarding our data collection in Appendix C.

Since oftentimes the questions on StackExchange contain less information than in scientific publications, there are fewer keyphrases per data point in STACKEX (statistics are shown in Table 1). Furthermore, StackExchange uses a tag recommendation system that suggests topic-relevant tags to users while submitting questions; therefore, we are more likely to see general terminology such as

⁵<https://archive.org/details/stackexchange>, we choose 19 computer science related topics from Oct. 2017 dump.

Model	KP20K	Inspec	Krapivin	NUS	SemEval
CatSeq	31.9	30.7	32.3	38.3	31.0
+ Orth. Reg.	31.1	29.3	31.0	36.5	29.5
+ Sem. Cov.	32.9	32.1	34.5	40.2	32.9
CatSeqD	35.7	33.1	37.1	40.6	35.7

Table 5: Ablation study with $F_1@O$ scores on five scientific publication datasets.

Linux and **Java**⁶. This characteristic challenges models with respect to their ability to distill major topics of a question rather than selecting specific snippets from the text.

We report our models’ performance on **STACKEX** in Table 3. Results show **CatSeqD** performs the best in general; on the *absent*-keyphrase generation tasks, it outperforms **CatSeq** by a large margin.

5.3 Generating Variable Number Keyphrases

One key advantage of our proposed model is the capability of predicting the number of keyphrases conditioned on the given source text. We thus conduct a set of experiments on **KP20K** and **STACKEX** present keyphrase generation tasks, as shown in Table 4, to study such behavior. We adopt the self-terminating decoding strategy (Section 3.3), and use both $F_1@O$ and $F_1@M$ (Section 4) to evaluate.

In these experiments, we use beam search as in most Natural Language Generation (NLG) tasks, i.e., only use the top ranked prediction sequence as output. We compare the results with greedy search. Since no existing model is capable of generating variable number of keyphrases, in this subsection we only report performance on such setting from **CatSeq** and **CatSeqD**.

From Table 4 we observe that in the variable number generation setting, greedy search outperforms beam search consistently. This may be because beam search tends to generate short and similar sequences. We can also see the resulting $F_1@O$ scores are generally lower than results reported in previous subsections, this suggests an over-generation decoding strategy may still benefit from achieving higher recall.

6 Analysis and Discussion

6.1 Ablation Study

We conduct an ablation experiment to study the effects of orthogonal regularization and semantic coverage mechanism on **CatSeq**. As shown in Table 5, semantic coverage provides significant boost to **CatSeq**’s performance on all datasets. Orthogonal regularization hurts performance when is solely applied to **CatSeq** model. Interestingly, when both components are enabled (**CatSeqD**), the model outperforms **CatSeq** by a noticeable margin on all datasets, this suggests the two components help keyphrase generation in a synergistic way. One future direction is to apply orthogonal regularization directly on target encoder, since the regularizer can potentially diversify target representations at phrase level, which may further encourage diverse keyphrase generation in decoder.

6.2 Visualizing Diversified Generation

To verify our assumption that target encoding and orthogonal regularization help to boost the diversity of generated sequences, we use two metrics, one quantitative and one qualitative, to measure diversity of generation.

First, we simply calculate the average *unique* predicted phrases produced by both **CatSeq** and **CatSeqD** in experiments shown in Section 5.1 (beam size is 50). The resulting numbers are 20.38 and 89.70 for **CatSeq** and **CatSeqD** respectively. Second, from the model running on the **KP20K** validation set, we randomly sample 2000 decoder hidden states at k steps following a delimiter ($k = 1, 2, 3$) and apply an unsupervised clustering method (t-SNE (van der Maaten and Hinton, 2008)) on them. From the Figure 2 we can see that hidden states sampled from **CatSeqD** are easier to cluster while hidden states sampled from **CatSeq** yield one mass of vectors with no obvious distinct clusters. Results on both metrics suggest target encoding and orthogonal regularization indeed help diversifying generation of our model.

6.3 Qualitative Analysis

To illustrate the difference of predictions between our proposed models, we show an example chosen from the **KP20K** validation set in Appendix F. In this example there are 29 ground truth phrases. Neither of the models is able to generate all of the

⁶One example is shown in Appendix F.

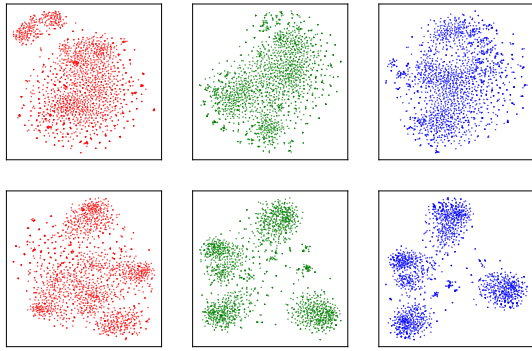


Figure 2: t-SNE results on decoder hidden states. Up-per row: **CatSeq**; lower row: **CatSeqD**; column k shows hidden states sampled from tokens at k steps following a delimiter.

keyphrases, but it is obvious that the predictions from **CatSeq** all start with “test”, while predictions from **CatSeqD** are diverse. This to some extent verifies our assumption that without the target encoder and orthogonal regularization, decoder states following delimiters are less diverse.

7 Conclusion and Future Work

We propose a recurrent generative model that sequentially generates multiple keyphrases, with two extra modules that enhance generation diversity. We propose new metrics to evaluate keyphrase generation. Our model shows competitive performance on a set of keyphrase generation datasets, including one introduced in this work. In future work, we plan to investigate how target phrase order affects the generation behavior, and further explore set generation in an order invariant fashion.

Acknowledgments

This work is supported by the National Science Foundation under grant No. 1525186. This research was also supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. The authors thank the anonymous ACL reviewers for their helpful feedback and suggestions.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain separation networks. *CoRR*, abs/1608.06019.

Chan, H. P., Chen, W., Wang, L., and King, I. (2019). Neural keyphrase generation via reinforcement learning with adaptive rewards. *arXiv preprint arXiv:1906.04106*.

Chen, J., Zhang, X., Wu, Y., Yan, Z., and Li, Z. (2018a). Keyphrase generation with correlation constraints. *CoRR*, abs/1808.07185.

Chen, W., Chan, H. P., Li, P., Bing, L., and King, I. (2019a). An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. *arXiv preprint arXiv:1904.03454*.

Chen, W., Gao, Y., Zhang, J., King, I., and Lyu, M. R. (2018b). Title-guided encoding for keyphrase generation. *CoRR*, abs/1808.08575.

Chen, W., Gao, Y., Zhang, J., King, I., and Lyu, M. R. (2019b). Title-guided encoding for keyphrase generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6268–6275.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *CoRR*, abs/1705.00106.

Forster, K. I. and Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, 4(1):53–61.

Gollapalli, S. D., Li, X., and Yang, P. (2017). Incorporating expert knowledge into keyphrase extraction. In *AAAI*, pages 3180–3187. AAAI Press.

Gu, J., Lu, Z., Li, H., and Li, V. O. K. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *CoRR*, abs/1603.06393.

Gülçehre, Ç., Ahn, S., Nallapati, R., Zhou, B., and Bengio, Y. (2016). Pointing the unknown words. *CoRR*, abs/1603.08148.

Hasan, K. S. and Ng, V. (2010). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 365–373. Association for Computational Linguistics.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

- Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.
- Le, T. T. N., Nguyen, M. L., and Shimazu, A. (2016). Unsupervised keyphrase extraction: Introducing new kinds of words to keyphrases. *29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016*.
- Liu, Z., Chen, X., Zheng, Y., and Sun, M. (2011). Automatic keyphrase extraction by bridging vocabulary gap. *ACL*.
- Logeswaran, L. and Lee, H. (2018). An efficient framework for learning sentence representations. *CoRR*, abs/1803.02893.
- Lopez, P. and Romary, L. (2010). Humb: Automatic key term extraction from scientific articles in grobidp. *the 5th International Workshop on Semantic Evaluation*.
- Luan, Y., Ostendorf, M., and Hajishirzi, H. (2017). Scientific information extraction with semi-supervised neural tagging. *CoRR*, abs/1708.06075.
- Medelyan, O., Frank, E., and Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. *EMNLP*.
- Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. In *ACL*.
- Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.
- Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.
- Reddy, D. R. et al. (1977). Speech understanding systems: A summary of results of the five-year research effort. *Department of Computer Science. Carnegie-Mell University, Pittsburgh, PA*, 17.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*
- Subramanian, S., Wang, T., Yuan, X., and Trischler, A. (2017). Neural models for key phrase detection and question generation. *CoRR*, abs/1706.04560.
- Sun, Z., Tang, J., Du, P., Deng, Z.-H., and Nie, J.-Y. (2019). Divgraphpointer: A graph pointer network for extracting diverse keyphrases. *arXiv preprint arXiv:1905.07689*.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.
- van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.
- van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.
- Van Zandt, T. and Townsend, J. T. (1993). Self-terminating versus exhaustive processes in rapid visual and memory search: An evaluative review. *Perception & Psychophysics*, 53(5):563–580.
- Wan, X. and Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. *AAAI*.
- Wang, M., Zhao, B., and Huang, Y. (2016). Ptr: Phrase-based topical ranking for automatic keyphrase extraction in scientific publications. *ICONIP 2016*.
- Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). Kea: Practical automatic keyphrase extraction. In *DL '99*.
- Xu, Q., Zhang, J., Qu, L., Xie, L., and Nock, R. (2018). D-page: Diverse paraphrase generation. *CoRR*.
- Yang, Z., Hu, J., Salakhutdinov, R., and Cohen, W. W. (2017). Semi-supervised qa with generative domain-adaptive nets. In *ACL*.
- Ye, H. and Wang, L. (2018). Semi-supervised learning for neural keyphrase generation. *CoRR*, abs/1808.06773.
- Yuan, X., Wang, T., Gülçehre, Ç., Sordoni, A., Bachman, P., Subramanian, S., Zhang, S., and Trischler, A. (2017). Machine comprehension by text-to-text neural question generation. *CoRR*, abs/1705.02012.
- Zhang, Q., Wang, Y., Gong, Y., and Huang, X. (2016). Keyphrase extraction using deep recurrent neural networks on twitter. In *EMNLP*.
- Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. *CoRR*, abs/1703.10931.
- Zhao, J. and Zhang, Y. (2019). Incorporating linguistic constraints into keyphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5224–5233.
- Zhou, Q., Yang, N., Wei, F., and Zhou, M. (2017). Selective encoding for abstractive sentence summarization. *CoRR*, abs/1704.07073.

A Sequence to Sequence Generation

A.1 The Encoder-Decoder Model

Given a source text consisting of N words w_e^1, \dots, w_e^N , the encoder converts their corresponding embeddings x_e^1, \dots, x_e^N into a set of N real-valued vectors $h_e = (h_e^1, \dots, h_e^N)$ with a bidirectional GRU (Cho et al., 2014):

$$\begin{aligned} h_{e,\text{fwd}}^t &= \text{GRU}_{e,\text{fwd}}(x_e^t, h_{e,\text{fwd}}^{t-1}), \\ h_{e,\text{bwd}}^t &= \text{GRU}_{e,\text{bwd}}(x_e^t, h_{e,\text{bwd}}^{t+1}), \\ h_e^t &= \langle h_{e,\text{fwd}}^t, h_{e,\text{bwd}}^t \rangle. \end{aligned} \quad (6)$$

Dropout (Srivastava et al., 2014) is applied to both x_e and h_e for regularization.

The decoder is a uni-directional GRU, which generates a new state h_d^t at each time-step t from the word embedding x_d^t and the recurrent state h_d^{t-1} :

$$h_d^t = \text{GRU}_d(x_d^t, h_d^{t-1}).^7 \quad (7)$$

The initial state h_d^0 is derived from the final encoder state h_e^N by applying a single-layer feed-forward neural net (FNN):

$$h_d^0 = L_0^{\tanh}(h_e^N). \quad (8)$$

Dropout is applied to both the embeddings x_d and the GRU states h_d .

A.2 Attentive Decoding

When generating token y^t , in order to better incorporate information from the source text, an attention mechanism (Bahdanau et al., 2014) is employed to infer the importance $\alpha^{t,i}$ of each source word w_e^i given the current decoder state h_d^t . This importance is measured by an energy function with a 2-layer FNN:

$$\text{energy}(h_d^t, h_e^i) = L_1(L_2^{\tanh}(\langle h_d^t, h_e^i \rangle)). \quad (9)$$

The output over all decoding steps t thus define a distribution over the source sequence:

$$\alpha^t = \text{softmax}(\text{energy}(h_d^t, h_e)). \quad (10)$$

These attention scores are then used as weights for a refined representation of the source encodings, which is then concatenated to the decoder state h_d^t to derive a generative distribution p_a :

$$p_a(y^t) = L_3^{\text{softmax}}(L_4^{\tanh}(\langle h_d^t, \sum_i \alpha^{t,i} \cdot h_e^i \rangle)), \quad (11)$$

⁷During training (with teacher forcing), w_d^t is the ground truth target token at previous time-step $t-1$; during evaluation, $w_d^t = y^{t-1}$, is the prediction at the previous time-step.

where the output size of L_3 equals to the target vocabulary size. Subscript a indicates the *abstractive* nature of p_a since it is a distribution over a prescribed vocabulary.

A.3 Pointer Softmax

We employ the pointer softmax (Gülçehre et al., 2016) mechanism to switch between generating a token y^t (from a vocabulary) and pointing (to a token in the source text). Specifically, the pointer softmax module computes a scalar switch s^t at each generation time-step and uses it to interpolate the abstractive distribution $p_a(y^t)$ over the vocabulary (see Equation 11) and the extractive distribution $p_x(y^t) = \alpha^t$ over the source text tokens:

$$p(y^t) = s^t \cdot p_a(y^t) + (1 - s^t) \cdot p_x(y^t), \quad (12)$$

where s^t is conditioned on both the attention-weighted source representation $\sum_i \alpha^{t,i} \cdot h_e^i$ and the decoder state h_d^t :

$$s^t = L_5^{\text{sigmoid}}(\tanh(L_6(\sum_i \alpha^{t,i} \cdot h_e^i) + L_7(h_d^t))). \quad (13)$$

B Experiment Results on KP20K Absent Subset

Generating *absent* keyphrases on scientific publication datasets is a rather challenging problem. Existing studies often achieve seemingly good performance by measuring recall on tens and sometimes hundreds of keyphrases produced by exhaustive decoding with a large beam size — thus completely ignoring precision.

We report the models' Recall@10/50 scores on the *absent* portion of five scientific paper datasets in Table 6 to be in line with previous studies.

The absent keyphrase prediction highly prefers recall-oriented models, therefore CopyRNN with beam size of 200 is innately proper for this task setting. However, from the results we observe that with the help of exhaustive decoding and diverse mechanisms, **CatSeqD** is able to perform comparably to CopyRNN model, and it generally works better for top predictions. Even though the trend of models' performance somewhat matches what we observe on the *present* data, we argue that it is hard to compare different models' performance on such scale. We argue that **STACKEX** is better testbeds for *absent* keyphrase generation.

	Kp20K		Inspec		Krapivin		NUS		SemEval	
Model	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CopyRNN (Meng et al., 2017)	<u>11.5</u>	18.9	<u>5.1</u>	10.1	<u>11.6</u>	19.5	<u>7.8</u>	14.4	4.9	7.5
CopyRNN* (Meng et al., 2017)	3.3	8.7	4.0	8.3	4.0	8.1	2.4	8.1	0.5	2.6
CatSeq (ours)	6.0	6.2	2.8	2.9	7.0	7.4	3.7	3.1	2.5	2.5
CatSeqD (ours)	11.7	<u>15.1</u>	5.2	<u>7.1</u>	12.0	<u>14.5</u>	8.4	<u>11.0</u>	4.6	<u>6.3</u>

Table 6: Performance of absent keyphrase prediction on scientific publications datasets. Best/second-best performing score in each column is highlighted with bold/underline.

C STACKEX Data Collection

We download the public data dump from <https://archive.org/details/stackexchange>, and choose 19 computer science related topics from Oct. 2017 dump. We select computer science forums (CS/AI), using “title” + “body” as source text and “tags” as the target keyphrases. After removing questions without valid tags, we collect 330,965 questions. We thus randomly select 16,000 for validation, and another 16,000 as test set. Note some questions in StackExchange forums contain large blocks of code, resulting in long texts (sometimes more than 10,000 tokens after tokenization), this is difficult for most neural models to handle. Consequently, we truncate texts to 300 tokens and 1,000 tokens for training and evaluation splits respectively.

D Implementation Details

Implementation details of our proposed models are as follows. In all experiments, the word embeddings are initialized with 100-dimensional random matrices. The number of hidden units in both the encoder and decoder GRU are 150. The number of hidden units in target encoder GRU is 150. The size of vocabulary is 50,000. In all experiments, we use a dropout rate of 0.1.

The numbers of hidden units in MLPs described in Section 3 are as follows. During negative sampling, we randomly sample 16 samples from the same batch, thus target encoding loss in Equation 2 is a 17-way classification loss. In **CatSeqD**, we select both λ_{OR} and λ_{SC} in Equation 4 from [0.01, 0.03, 0.1, 0.3, 1.0] using validation sets. The selected values are listed in Table 7.

We use *Adam* (Kingma and Ba, 2014) as the step rule for optimization. The learning rate is $1e^{-3}$. The model is implemented using *PyTorch* (Paszke et al., 2017) and *OpenNMT* (Klein et al., 2017).

For exhaustive decoding, we use a beam size of 50 and a maximum sequence length of 40.

Experiment Setting	λ_{OR}	λ_{SC}
Table 2	1.0	0.03
Table 3	0.03	0.1
Table 4, KP20K Greedy	1.0	0.3
Table 4, KP20K Top Rank	1.0	0.3
Table 4, STACKEX Greedy	1.0	0.3
Table 4, STACKEX Top Rank	1.0	0.3
Table 5, CatSeq + Orth. Reg.	0.3	0.0
Table 5, CatSeq + Sem. Cov.	0.0	0.03
Table 5, CatSeqD	Same as Table 2	
Table 6	Same as Table 2	

Table 7: Semantic coverage and orthogonal regularization coefficients.

Following Meng et al. (2017), lowercase and stemming are performed on both the ground truth and generated keyphrases during evaluation.

We leave out 2,000 data examples as validation set for both KP20K and STACKEX and use them to identify optimal checkpoints for testing. And all the scores reported in this paper are from checkpoints with best performances ($F_1@O$) on validation set.

In Section 6.2, we use the default parameters for t-SNE in sklearn (learning rate is 200.0, number of iterations is 1000, as defined in ⁸).

E Dataset and Evaluation Details

We strictly follow the data pre-processing and evaluation protocols provided by Meng et al. (2017).

We pre-process both document texts and ground-truth keyphrases, including word segmentation, lowercasing and replacing all digits with symbol <digit>. In the datasets, examples with empty ground-truth keyphrases are removed.

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

Dataset	#Doc	#KP	#PreDoc	#PreKP	#AbsDoc	#AbsKP
KP20K	19,987	105,181	19,048	66,595	16,357	38,586
INSPEC	500	4,913	497	3,858	381	1,055
KRAPIVIN	460	2,641	437	1,485	417	1,156
NUS	211	2,461	207	1,263	195	1,198
SEMEVAL	100	1,507	100	671	99	836
STACKEX	16,000	43,131	13,475	24,809	10,984	18,322
DUC	308	2,484	308	2,421	38	63

Table 8: Statistics on number of documents and keyphrases of each test set. #Doc#KP denotes the number of documents/ground-truth keyphrases in the dataset. #PreKP/#AbsKP denotes the number of present/absent ground-truth keyphrases, and #PreDoc/#AbsDoc denotes the number of documents that contain at least one present/absent ground-truth keyphrase.

We evaluate models’ performance on predicting present and absent phrases separately. Specifically, we first lowercase the text, then we determine the presence of each ground-truth keyphrase by checking whether it is a sub-string of the source text (we use Porter Stemmer⁹). To evaluate present phrase performance, we compute Precision/Recall/F1-score (see 14-16 for formulas) for each document taking only present ground-truth keyphrases as target and ignore the absent ones.

$$P@k = \frac{\#(correct@k)}{\min\{k, \#(pred)\}} \quad (14)$$

$$R = \frac{\#(correct@k)}{\#(target)} \quad (15)$$

$$F_1@k = \frac{2 * P@k * R}{P@k + R} \quad (16)$$

where $\#(pred)$ and $\#(target)$ are the number of predicted and ground-truth keyphrases respectively; and $\#(correct@k)$ is the number of correct predictions among the first k results.

We report the macro-averaged scores over documents that have at least one present ground-truth phrases (corresponding to the column #PreDoc in Table 8), and similarly to the case for absent phrase evaluation.

F Examples of KP20K and STACKEX with Model Prediction

See Table 9 and Figure 3.

⁹<https://www.nltk.org/api/nltk.stem.html#module-nltk.stem.porter>

Source	<p>Integration of a Voice Recognition System in a Social Robot</p> <p>Human-robot interaction (HRI) (1) is one of the main fields in the study and research of robotics. Within this field, dialogue systems and interaction by voice play an important role. When speaking about human-robot natural dialogue we assume that the robot has the capability to accurately recognize what the human wants to transmit verbally and even its semantic meaning, but this is not always achieved. In this article we describe the steps and requirements that we went through in order to endow the personal social robot Maggie , developed at the University Carlos III of Madrid, with the capability of understanding the natural language spoken by any human. We have analyzed the different possibilities offered by current software/hardware alternatives by testing them in real environments. We have obtained accurate data related to the speech recognition capabilities in different environments, using the most modern audio acquisition systems and analyzing not so typical parameters such as user age, gender, intonation, volume, and language. Finally, we propose a new model to classify recognition results as accepted or rejected, based on a second automatic speech recognition (ASR) opinion. This new approach takes into account the precalculated success rate in noise intervals for each recognition framework, decreasing the rate of false positives and false negatives.</p>
CatSeq	voice recognition system ; social robot ; human robot interaction ; voice recognition ; hri ; speech recognition ; automatic speech recognition ; noise intervals ; noise ; human robot ; automatic speech ; natural language
CatSeqD	human robot interaction ; voice recognition ; social robotics ; social robots ; integration ; speech recognition ; hri ; social robot ; robotics ; voice recognition system ; recognition ; asr ; automatic speech recognition ;
Ground Truth	asr ; automatic speech recognition ; dialogue ; human robot interaction ; maggie ; social robot ; speech recognition ; voice recognition ;

Table 9: Example from KP20K validation set, and predictions generated by CatSeq and CatSeqD models.

Creating a function when user clicks on element to toggle another element

[Ask Question](#)

Asked 5 years, 6 months ago Active 5 years, 6 months ago Viewed 1k times

I use a lot of "when user clicks on element then another element will be displayed as visible/hidden... aka toggle".

2

So I'm trying to make a general function to abide the DRY principal.

▼

Before jumping into the code, I just wanted to state that there are still bugs. I just want to know before continuing if writing a function like the one below is necessary or if I should just ignore the DRY in this case and have lots of separate toggle functions. Or if it's necessary, would there be a more efficient method?

★

[JSFiddle](#)

*Brief Description of Code

-There are 4 parameters

- The first two are mandatory, the last two are optional
- The index parameter is a string of numbers ex. "1.0","1"
- The reason I need the last two parameters is because I have elements that can't be easily targeted. For example, they don't have an #id. So to grab those generic elements, I use `getElementsByTagName`
- Then I'll use the index parameter to get the element in the array.
- index parameter: The number before '.' is the index of the clickElement and the number after '.' is the index of the toggleElement. `clickElement = 1. toggleElement = 0.`

```
function toggleClick(clickElement, toggleElement, condition, index) {
    var toggleIndex;
    var clickIndex;

    if(condition === false){
        if(index.indexOf('.') != -1){
            //indexOf returns -1 if '.' is not found.
            // Checks to see if two index is entered
            indexArray = index.split('.');
            clickIndex = indexArray[0];
            toggleIndex = indexArray[1];
            clickElement = document.getElementsByTagName(clickElement)[clickIndex];
            toggleElement = document.getElementsByTagName(toggleElement)[toggleIndex];
        }else{
            //If there is no '.', that means only one index was entered.
            // By function requirement, it should be the index of clickElement

            clickElement = document.getElementsByTagName(clickElement)[index];
        }
    }
    $(clickElement).click(function() {
        $(toggleElement).toggle();
    });
}
```

This is one of my first functions, and so I need some guidance on whether I'm writing bad code. Yes, this might be broad because there may be many different ways to write this function. But I just need to know if there's a more efficient way or if my way is okay.

[javascript](#) [jquery](#) [event-handling](#)

Blog

[Tis the Season for Hats! Join Us for Winter Bash 2019](#)

[This Week #StackOverflowKnows About Infinity, Internet-Speak, and Password...](#)

Featured on Meta

[The context of "Lacks concrete context": let's fix it](#)



More ways to Q&A:
Try Stack Overflow for Business
with secure single sign-on for
your entire team

Want a javascript job?

Lead JavaScript Developer

CenturyLink Ashburn, VA

[javascript](#) [css](#)

Front End Web Developer - Entry/Mid-Level

National Security Agency Fort Meade, MD

\$47K - \$108K

[javascript](#) [user-experience](#)

[View more jobs on Stack Overflow](#)

Related

- 2 Accordion JS plugin, CSS3 webkit
- 2 Optimization of recursive JQuery images swap function
- 6 Tracking (uniform spacing) between 2D elements
- 4 Javascript equalise heights of groups of elements
- 3 Creating keyboard shortcuts for highlighting elements of given classes and toggle through them

Figure 3: Example from the STACKEX dataset, we show the screenshot of the original web page to better present the example. Note the input to the model is the entire question (including the code), we removed the format information in the dataset. Also note on the bottom of the screenshot it shows the 3 keyphrases (in this example all absent) which we collected as the ground-truth keyphrases in our dataset.

Ground Truth: [javascript](#) ; [jquery](#) ; [event handling](#)

CatSeq Prediction: [javascript](#); [c#](#); [jquery](#); [php](#); [linq](#); [comparative review](#); [ecmascript 6](#); [asp](#) . [js](#); [beginner](#); [strings](#); [performance](#); [datetime](#)

CatSeqD Prediction: [javascript](#) ; [jquery](#) ; [performance](#) ; [event handling](#) ; [array](#) ; [twitter bootstrap](#) ; [beginner](#) ; [algorithm](#) ; [indexarray](#) ; [optimization](#) ; [event programming](#) ; [datetime](#) ; [comparative review](#) ; [ecmascript 6](#) ; [indexof](#) ; [dry](#) ; [php](#) ; [r](#) ; [java](#) ; [coffeescript](#) ; [combinatorics](#) ; [dom](#) ; [html](#) ; [event tracking](#) ; [strings](#) ; [python](#) ; [ruby](#) ; [natural language processing](#) ; [animation](#) ; [angular . js](#) ; [homework](#) ; [parameters](#) ; [jquery ui](#) ; [functional programming](#) ; [google app engine](#) ; [. net](#) ; [python 2 . 7](#) ; [c#](#) ; [php5](#) ; [validation](#) ; [regex](#) ; [parsing](#) ; [formatting](#) ; [hash table](#) ; [object oriented](#) ; [web scraping](#) ; [python 3 . x](#) ; [python 3 . x programming](#) ; [python 2 . net](#) ; [python 2 . 6](#) ; [python 2 . sql](#) ; [mysql](#) ; [object oriented design](#) ; [actionscript](#)