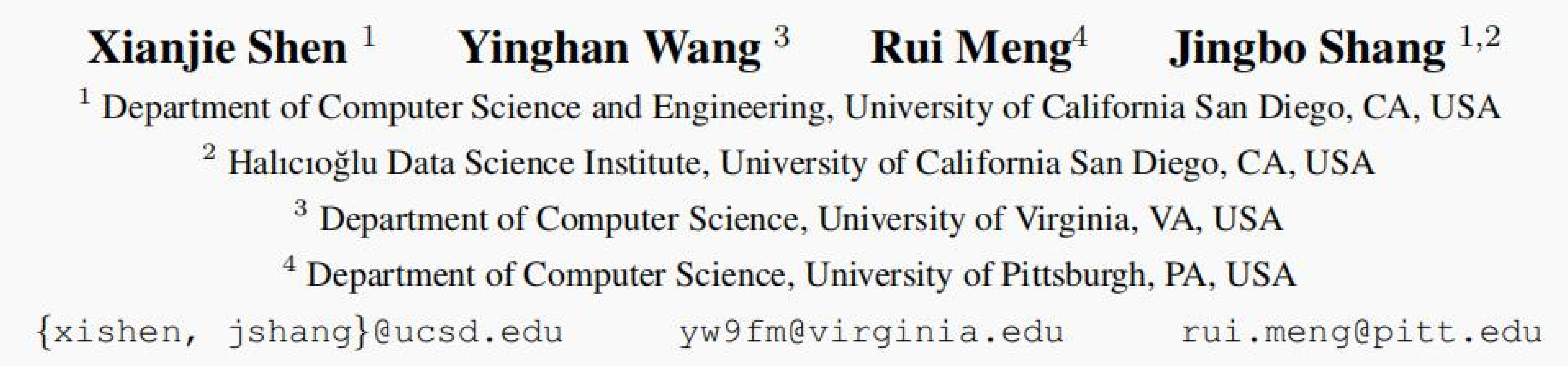
Unsupervised Deep Keyphrase Generation



Xianjie Shen [1] Yinghan Wang [3] Rui Meng[4] Jingbo Shang [1,2]

[1] Department of Computer Science and Engineering, University of California San Diego, CA, USA
[2] Halıcıoğlu Data Science Institute, University of California San Diego, CA, USA
[3] Department of Computer Science, University of Virginia, VA, USA
[4] Department of Computer Science, University of Pittsburgh, PA, USA

{xishen, jshang}@ucsd.edu yw9fm@virginia.edu rui.meng@pitt.edu

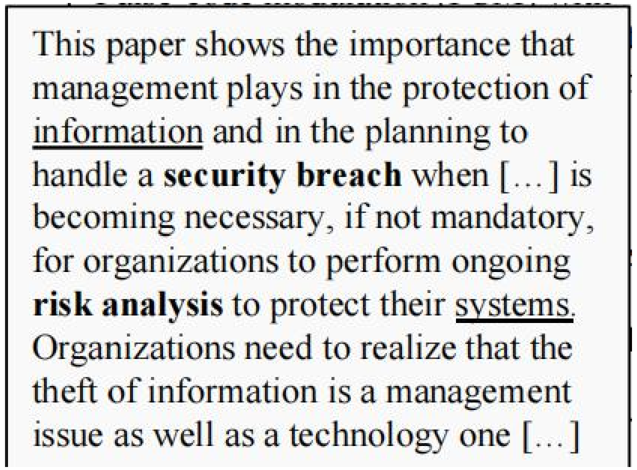
AAAI 2022

https://github.com/Jayshen0/Unsupervised-Deep-Keyphrase-Generation

Reported by Dongdong Hu

Introduction

- Extractive methods can only predict phrases that appear in the original document.
- absent keyphrases of a document can be present in other documents as present keyphrases.
- many absent keyphrases in fact appear in the original document in part as separate tokens



This paper shows the importance that management plays in the protection of information and in the planning to handle a **security breach** when […] is becoming necessary, if not mandatory, for organizations to perform ongoing **risk analysis** to protect their systems. Organizations need to realize that the theft of information is a management issue as well as a technology one […]

Method

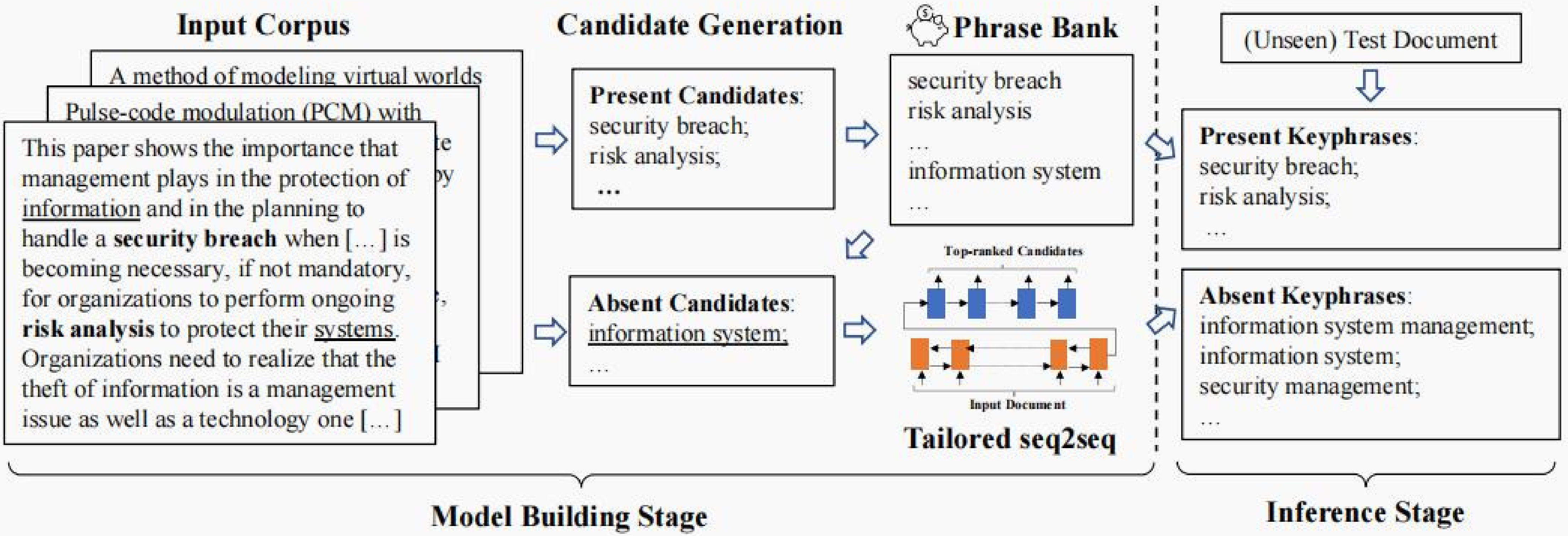
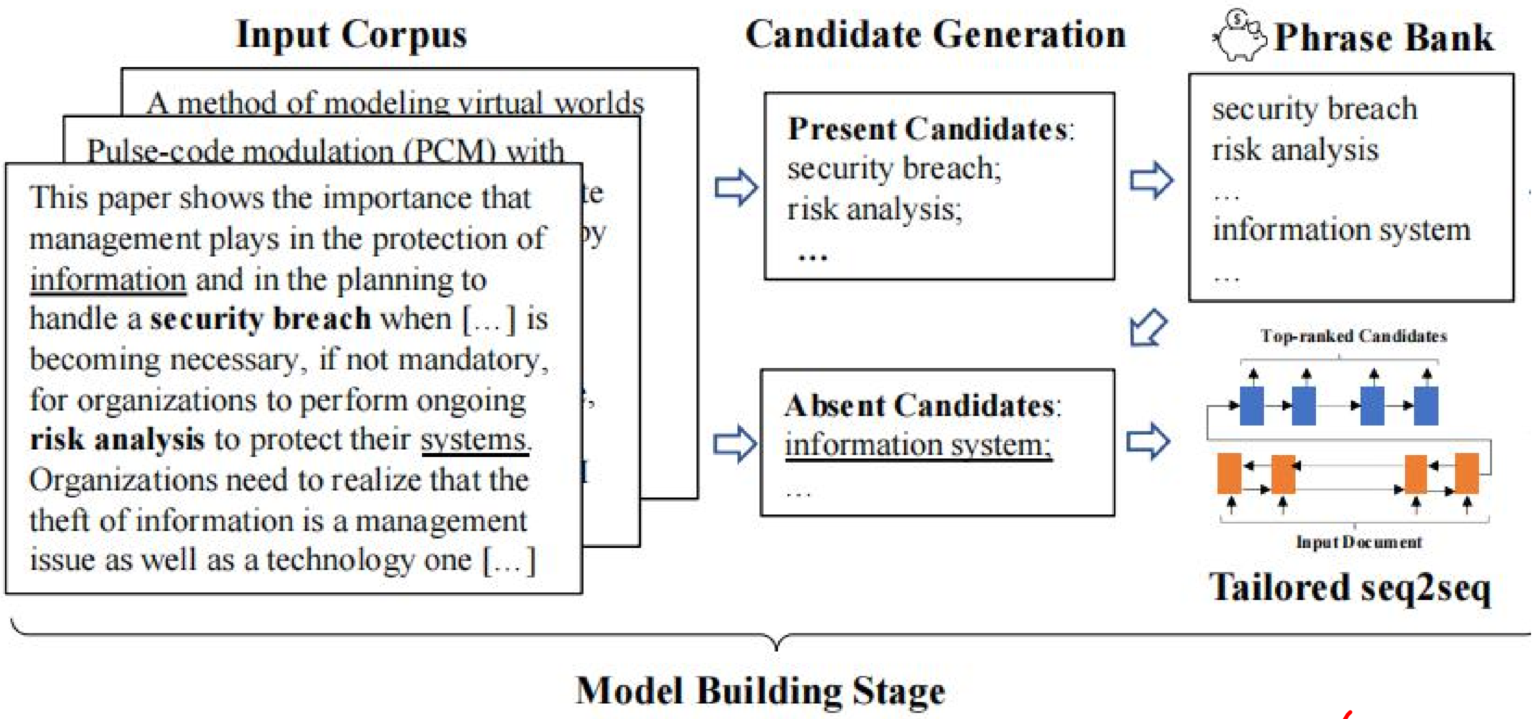

Input Corpus
A method of modeling virtual worlds
Pulse-code modulation (PCM) with
This paper shows the importance that management plays in the protection of information and in the planning to handle a security breach when [...] is becoming necessary, if not mandatory, for organizations to perform ongoing risk analysis to protect their systems. Organizations need to realize that the theft of information is a management issue as well as a technology one [...]
Candidate Generation
Present Candidates:
security breach;
risk analysis;
...
Absent Candidates:
information system;
...
Phrase Bank
security breach
risk analysis
...
information system
...
Top-ranked Candidates
Input Document
Tailored seq2seq
(Unseen) Test Document
Present Keyphrases:
security breach;
risk analysis;
...
Absent Keyphrases:
information system management;
information system;
security management;
...
Model Building Stage
Inference Stage

Figure 1: An overview of our proposed AutoKeyGen framework with a part of real example. The full version of the example can be found in our case study.

Method



A method of modeling virtual worlds

Pulse-code modulation (PCM) with

This paper shows the importance that management plays in the protection of information and in the planning to handle a **security breach** when [...] is becoming necessary, if not mandatory, for organizations to perform ongoing **risk analysis** to protect their systems. Organizations need to realize that the theft of information is a management issue as well as a technology one [...]

Present Candidates:
security breach;
risk analysis;
...

Absent Candidates:
information system;
...

security breach
risk analysis
...
information system
...

$$\mathbf{x} = [x_1, \ldots, x_{|\mathbf{x}|}]$$

$$\mathcal{Y}^P = \{\mathbf{y}_1^p, \ldots, \mathbf{y}_{|\mathcal{Y}^P|}^p\}$$

$$\mathcal{Y}^A = \{\mathbf{y}_1^a, \ldots, \mathbf{y}_{|\mathcal{Y}^A|}^a\}.$$

$$\mathcal{Y} =< \mathcal{Y}^P, \mathcal{Y}^A >$$

Embedding Similarity

$$\text{Semantic}(\mathbf{x}, c) = \frac{||E(\mathbf{x}) \cdot E(c)||}{||E(\mathbf{x})|| \cdot ||E(c)||}$$

有错

lexical-level similarity

$$\text{Lexical}(\mathbf{x}, c) = \frac{\text{TF}(c, \mathbf{x})}{|\mathbf{x}|} log \frac{|\mathcal{D}|}{\text{DF}(c, \mathcal{D})}$$

$$\text{RankScore}(\mathbf{x}, c) = \sqrt{\text{Semantic}(\mathbf{x}, c) \cdot \text{Lexical}(\mathbf{x}, c)}$$

where $|\mathbf{x}|$ is the number of word in document $\mathbf{x}$, $\text{TF}(c, \mathbf{x})$ is the term frequency of c in $\mathbf{x}$, $\text{DF}(c, \mathcal{D})$ is the document frequency of c in $\mathcal{D}$.

Method

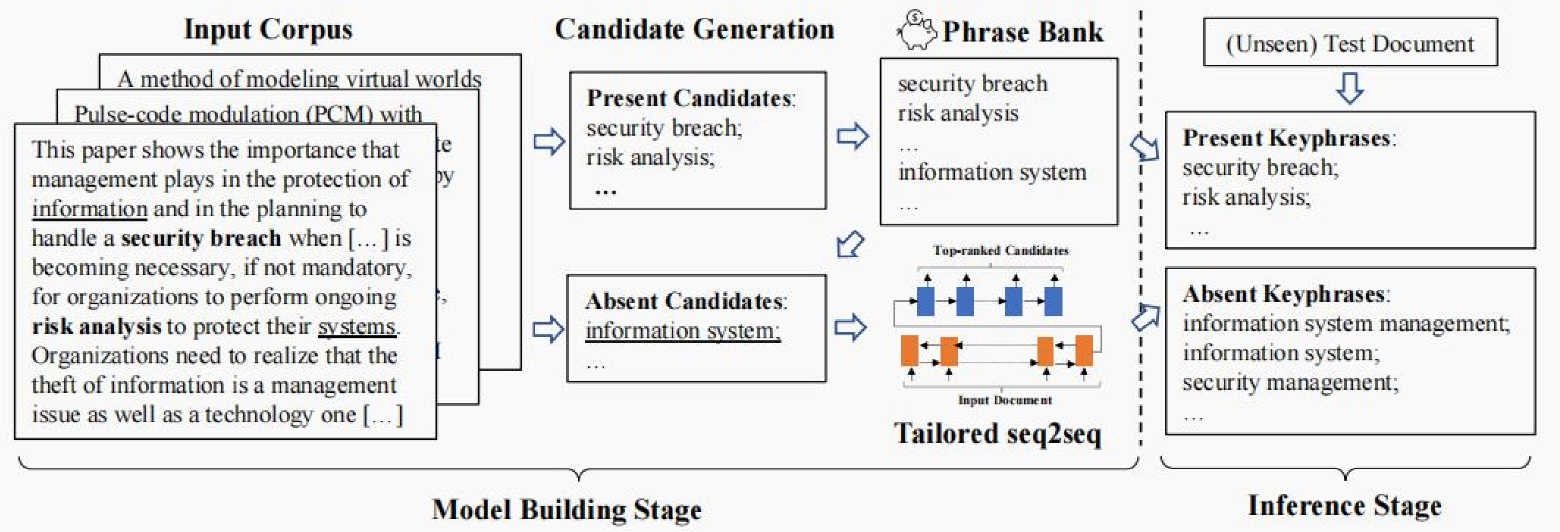

Input Corpus
A method of modeling virtual worlds
Pulse-code modulation (PCM) with
This paper shows the importance that management plays in the protection of information and in the planning to handle a security breach when [...] is becoming necessary, if not mandatory, for organizations to perform ongoing risk analysis to protect their systems. Organizations need to realize that the theft of information is a management issue as well as a technology one [...]
Candidate Generation
Present Candidates: security breach; risk analysis; ...
Absent Candidates: information system; ...
Phrase Bank
security breach
risk analysis
...
information system
...
Top-ranked Candidates
Input Document
Tailored seq2seq
(Unseen) Test Document
Present Keyphrases: security breach; risk analysis; ...
Absent Keyphrases: information system management; information system; security management; ...
Model Building Stage
Inference Stage

Figure 1: An overview of our proposed AutoKeyGen framework with a part of real example. The full version of the example can be found in our case study.

Classical Encoder-Decoder Model.

$$
\begin{aligned}
\mathbf{h}_{enc}^{t} &= f_{enc}(\mathbf{h}_{enc}^{t-1}, x^{t}),\\
\mathbf{c} &= q(h_{enc}^{1}, h_{enc}^{2}, \ldots, h_{enc}^{|\mathbf{x}|}),\\
\mathbf{h}_{dec}^{t} &= f_{dec}(\mathbf{h}_{dec}^{t-1}, o^{t-1}, \mathbf{c})
\end{aligned}
$$

$$
p_g(y^t | y^{1,\ldots,t-1}, \mathbf{x}) = f_{out}(y^{t-1}, \mathbf{h}_{dec}^{t}, \mathbf{c})
$$

o^{t-1} is the predicted output of decoder at time $t-1$; and $\mathbf{c}$ is the context vector derived from all the hidden states of encoder though a non-linear function q.

Tailored Seq2Seq Generative Model.

we encourage the decoder model to generate words that appear in the input document $\mathbf{x}$. More specifically, we double the probabilities of the words occurred in the input document.

Experiments

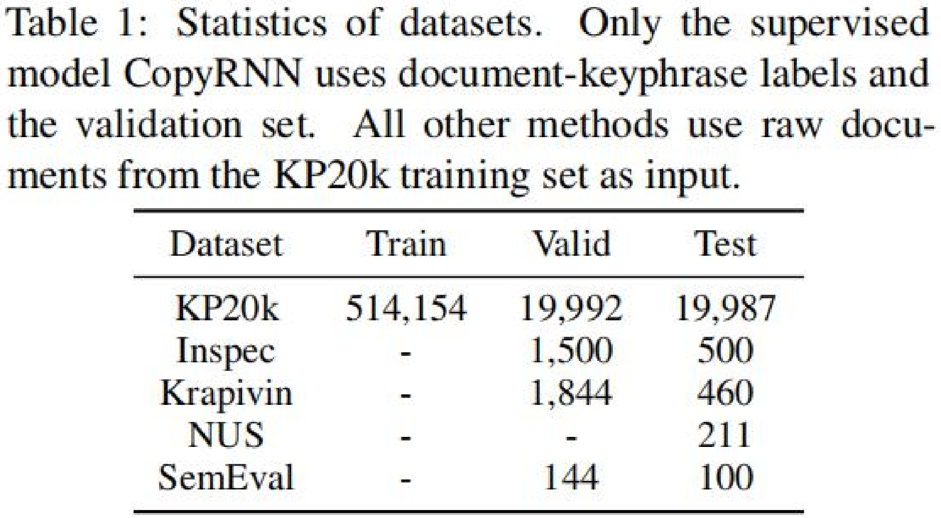


Table 1: Statistics of datasets. Only the supervised model CopyRNN uses document-keyphrase labels and the validation set. All other methods use raw documents from the KP20k training set as input.

Dataset	Train	Valid	Test
KP20k	514,154	19,992	19,987
Inspec	-	1,500	500
Krapivin	-	1,844	460
NUS	-	-	211
SemEval	-	144	100

Experiments

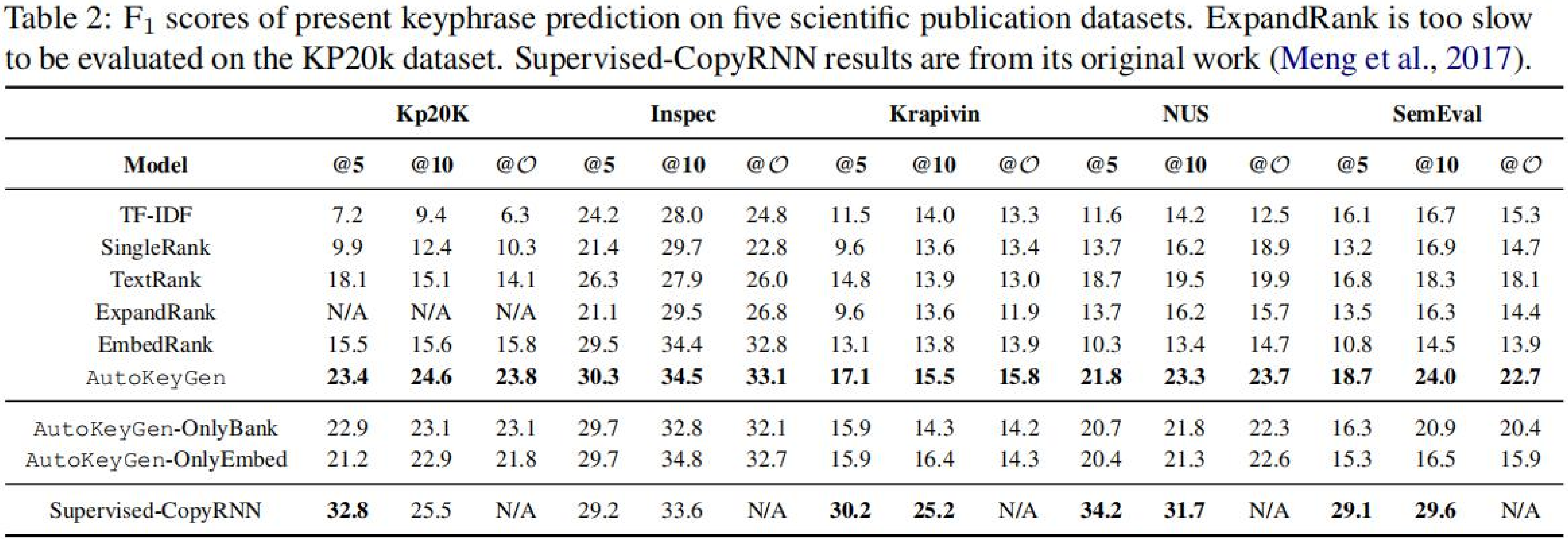


Table 2: F_1 scores of present keyphrase prediction on five scientific publication datasets. ExpandRank is too slow to be evaluated on the KP20k dataset. Supervised-CopyRNN results are from its original work (Meng et al., 2017).

	Kp20K			**Inspec**			**Krapivin**			**NUS**			**SemEval**		
Model	**@5**	**@10**	**@$\mathcal{O}$**	**@5**	**@10**	**@$\mathcal{O}$**	**@5**	**@10**	**@$\mathcal{O}$**	**@5**	**@10**	**@$\mathcal{O}$**	**@5**	**@10**	**@$\mathcal{O}$**
TF-IDF	7.2	9.4	6.3	24.2	28.0	24.8	11.5	14.0	13.3	11.6	14.2	12.5	16.1	16.7	15.3
SingleRank	9.9	12.4	10.3	21.4	29.7	22.8	9.6	13.6	13.4	13.7	16.2	18.9	13.2	16.9	14.7
TextRank	18.1	15.1	14.1	26.3	27.9	26.0	14.8	13.9	13.0	18.7	19.5	19.9	16.8	18.3	18.1
ExpandRank	N/A	N/A	N/A	21.1	29.5	26.8	9.6	13.6	11.9	13.7	16.2	15.7	13.5	16.3	14.4
EmbedRank	15.5	15.6	15.8	29.5	34.4	32.8	13.1	13.8	13.9	10.3	13.4	14.7	10.8	14.5	13.9
AutoKeyGen	**23.4**	**24.6**	**23.8**	**30.3**	**34.5**	**33.1**	**17.1**	**15.5**	**15.8**	**21.8**	**23.3**	**23.7**	**18.7**	**24.0**	**22.7**
AutoKeyGen-OnlyBank	22.9	23.1	23.1	29.7	32.8	32.1	15.9	14.3	14.2	20.7	21.8	22.3	16.3	20.9	20.4
AutoKeyGen-OnlyEmbed	21.2	22.9	21.8	29.7	34.8	32.7	15.9	16.4	14.3	20.4	21.3	22.6	15.3	16.5	15.9
Supervised-CopyRNN	**32.8**	25.5	N/A	29.2	33.6	N/A	**30.2**	**25.2**	N/A	**34.2**	**31.7**	N/A	**29.1**	**29.6**	N/A

Experiments

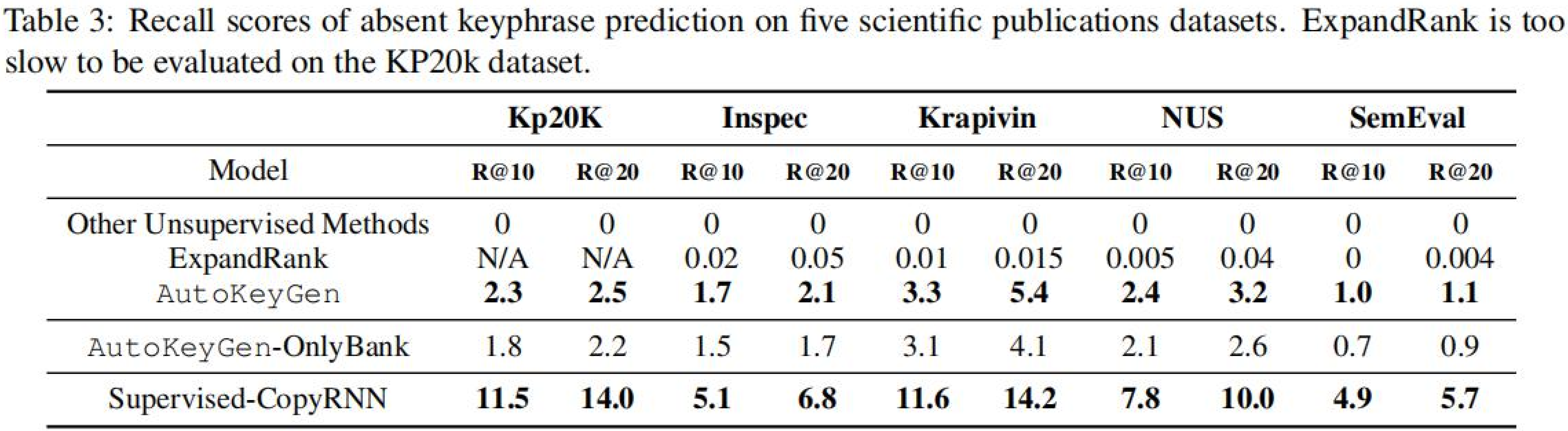
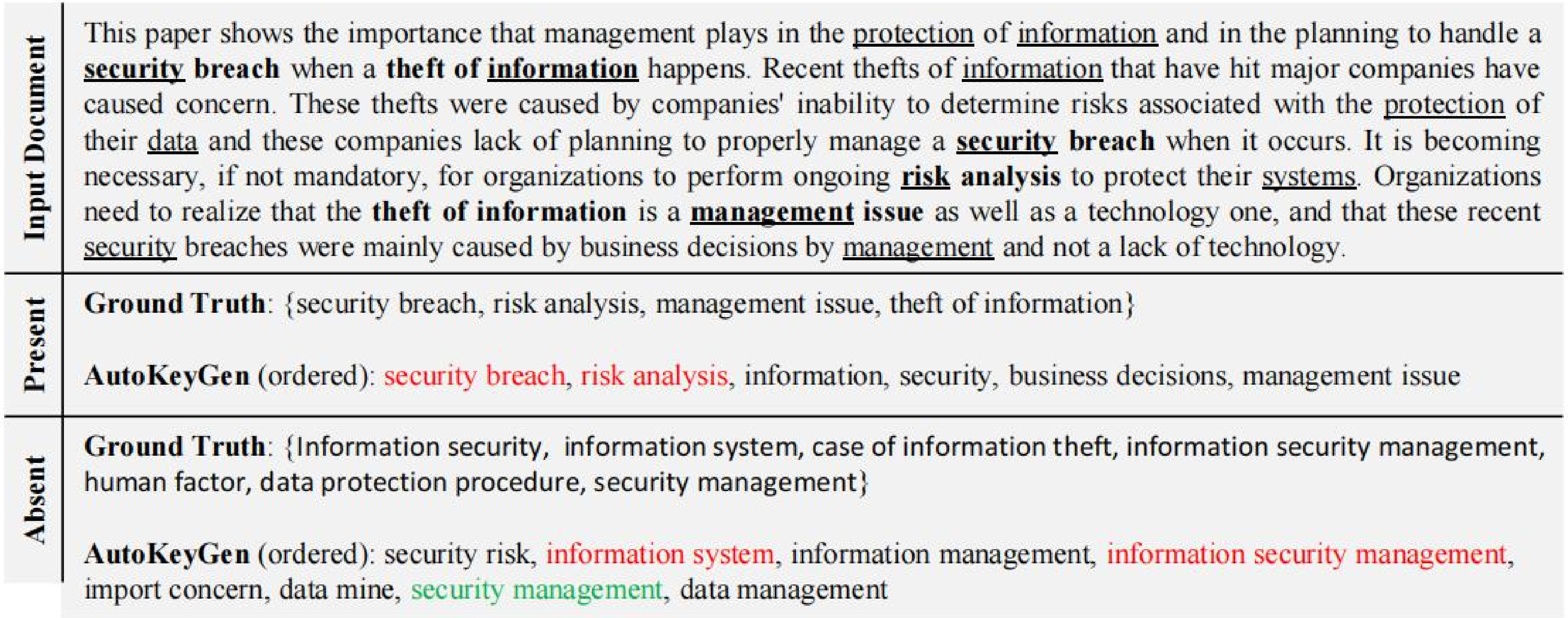


Table 3: Recall scores of absent keyphrase prediction on five scientific publications datasets. ExpandRank is too slow to be evaluated on the KP20k dataset.

	Kp20K		**Inspec**		**Krapivin**		**NUS**		**SemEval**	
Model	**R@10**	**R@20**	**R@10**	**R@20**	**R@10**	**R@20**	**R@10**	**R@20**	**R@10**	**R@20**
Other Unsupervised Methods	0	0	0	0	0	0	0	0	0	0
ExpandRank	N/A	N/A	0.02	0.05	0.01	0.015	0.005	0.04	0	0.004
AutoKeyGen	**2.3**	**2.5**	**1.7**	**2.1**	**3.3**	**5.4**	**2.4**	**3.2**	**1.0**	**1.1**
AutoKeyGen-OnlyBank	1.8	2.2	1.5	1.7	3.1	4.1	2.1	2.6	0.7	0.9
Supervised-CopyRNN	**11.5**	**14.0**	**5.1**	**6.8**	**11.6**	**14.2**	**7.8**	**10.0**	**4.9**	**5.7**

Experiments



Input Document	This paper shows the importance that management plays in the protection of information and in the planning to handle a **security breach** when a **theft of information** happens. Recent thefts of information that have hit major companies have caused concern. These thefts were caused by companies' inability to determine risks associated with the protection of their data and these companies lack of planning to properly manage a **security breach** when it occurs. It is becoming necessary, if not mandatory, for organizations to perform ongoing **risk analysis** to protect their systems. Organizations need to realize that the **theft of information** is a **management issue** as well as a technology one, and that these recent security breaches were mainly caused by business decisions by management and not a lack of technology.
Present	**Ground Truth**: {security breach, risk analysis, management issue, theft of information} **AutoKeyGen** (ordered): security breach, risk analysis, information, security, business decisions, management issue
Absent	**Ground Truth**: {Information security, information system, case of information theft, information security management, human factor, data protection procedure, security management} **AutoKeyGen** (ordered): security risk, information system, information management, information security management, import concern, data mine, security management, data management

Figure 2: A case study of `AutoKeyGen` from the NUS test set. Present keyphrases are marked bold in the input document. Tokens in the input document related to absent keyphrases are underlined. Correctly predicted keyphrases are highlighted in red. The green one is a correct phrase predicted by our generating module, which is omitted by noun phrase extraction method.

Experiments

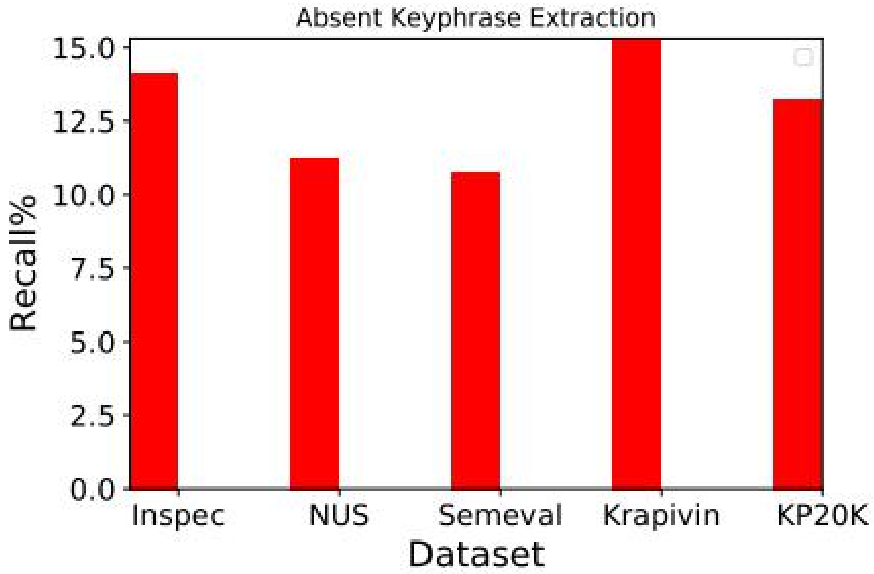


Figure 3: The recall of absent keyphrases using all the phrases in phrase bank on five datasets.

Thanks