TopicRank

Graph-Based Topic Ranking for Keyphrase Extraction

Adrien Bougouin and Florian Boudin and Béatrice Daille

Ygor Gallina

PhD Student at LS2N, Team TALN
Directress: DAILLE Béatrice
Supervisor: BOUDIN Florian
Departmental funding

Table of contents

1. Cleaning the reference

2. Abstractive and Extractive models

Cleaning the reference

Cleaning the reference i

Assumption

The model has low performance for generating absent keyphrases because the reference contain noise.

Cleaning the training reference should result in better performances as the reference is simpler to model.

	Test		Cleaned Test	
	F@5	MAP	F@5	MAP
CopySci	*27.6*	*28.5*	**28.3** +0.7	**29.5** +1.0
CopySci_clean	27.2	25.9	28.0 +0.8	27.1 +1.2

Table 1: Performance comparison of CopyRNN models trained on cleaned and raw reference

Cleaning the reference ii

		Test		Cleaned Test	
		F@5	MAP	F@5	MAP
Absent	CopySci	4.8	*6.0*	4.9 +0.1	**6.1** +0.1
	CopySci_clean	*4.9*	4.7	**5.1** +0.2	4.8 +0.1
Present	CopySci	*28.3*	*29.5*	**33.6** +5.3	**44.2** +14.7
	CopySci_clean	28.0	27.1	33.3 +5.3	40.9 +13.8

Table 2: Performance comparison of CopyRNN models trained on cleaned and raw reference

Cleaning the reference iii

CopySci_clean	Rank	CopySci
graph algorithm	1	\\(k\\) -separ problem
treewidth	2	treewidth
approxim algorithm	3	\\(k\\) -separ
bound treewidth	4	approxim algorithm
weakli chordal	5	complex

Table 3: Example of output of CopySci_clean and CopySci models

Cleaning the reference iv

Future work

Compute the percentage of noisy keyphrase generated by the CopySci model to see how often this happend.

Understand why the performance are not better with a cleaned model

1. The (simple) cleaning is not enough (there is still noise in the references)
2. The meta-parameters (epoch number) should be tweaked

Cleaning the reference v

Conclusion

The model trained on cleaned reference is less performant than the original model.

Evaluating on the cleaned reference always improves the performance.

The noise does not affect the generation of absent keyphrases. The model is robust to the noise.

Abstractive and Extractive models

Abstractive and Extractive models i

Assumption

In order to understand whether the tasks of generating absent (abstractive) and present (extractive) keyphrases are two different task and whether they **complement** or **restrain** each other. We train a model to generate keyphrases that do not appear in the ouput and another to generate keyphrases that do appear in the input.

[1] train a model using Graph Convolutionnal Networks that can only perform extraction and it outperforms the state of the art (DivGraphPointer F@5 36.8 vs. CopyRNN F@5 32.8).

The basic idea would be to split the keyphrase generation into two subtask : keyphrase extraction and keyphrase generation.

Abstractive and Extractive models ii

Model	All	Pres	Abs
CopySci	**27.6**	**33.0**	*4.8*
CopySci Pres	*27.4*	*32.7*	1.2
CopySci Abs	8.5	21.7	**4.9**

(a) F-1 Score @ 5

Model/P@5	All	Pres	Abs
CopySci	**28.2**	**28.2**	**3.7**
CopySci Pres	*28.1*	*28.0*	1.0
CopySci Abs	8.5	20.5	**3.7**

(b) Precison @ 5

Model/R@5	All	Pres	Abs
CopySci	**29.6**	**48.9**	*8.5*
CopySci Pres	*29.3*	*48.5*	1.9
CopySci Abs	9.3	29.0	**8.6**

(c) Recall @ 5

(d) Performances of keyphrase generation on unfiltered (All), only present (Pres) and only absent (Abs) KP20k reference for model trained on only present (CopySci Pres), absent (CopySci Abs) and all (CopySci) keyphrases

Abstractive and Extractive models iii

Conclusion

The two tasks don't seem to help each other their scores are just combined.

The Absent model does not generate better than the generic model. The model seem to be limiting the generation.

The Absent model generates present keyphrases. That is linked to the fact that "absent keyphrases" can be copied from the input (cf. Table 5).

Title	Hybrid	Analytical	Modeling	of	Pending	Cache	Hits , Data ...
Present		analytical	modeling				
Absent					pending		hit

Table 5: Example of Present and Absent keyphrases

Abstractive and Extractive models iv

Future works

- Can these conclusion apply to more than this dataset ?
  Do the same work with NYTime dataset to verify the hypothethis.
- How do the keyphrases differ between models ?
  By watching the outputs the kp of the absent model seemed (sometime) more generic than the extracted ones (cf. Table 6).
- How to propose a model specialized in generate Absent Keyphrases ?
- Redefine the task of generating to be more abstract (cf. Table 5)

Abstractive and Extractive models v

CopySci Abs	Copy Sci Pres
data mine	topolog similar
machin learn	random walk
cluster	heterogen network
data priorit	proteinprotein interact network
cluster analysi	priorit
pattern recognit	phenotyp data
bioinformat	diseas gene
graph theori	candid diseas gene
unsupervis learn	phenotyp
diseas priorit	multipl network

Table 6: Example of output of CopySci Abs and CopySci Pres models

References i

Z. Sun, J. Tang, P. Du, Z.-H. Deng, and J.-Y. Nie.
DivGraphPointer: A Graph Pointer Network for Extracting Diverse Keyphrases.
arXiv:1905.07689 [cs], May 2019.
arXiv: 1905.07689.