Back to the Basics: A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction

Asahi Ushio
Federico Liberatore
Jose Camacho-Collados

https://github.com/asahi417/kex

Keyword Extraction

Extracting **keywords** in a document.

Keyword is a **representative phrase** of the document.

Unsupervised Method > Supervised Method

Input Text (from SemEval2017):
Video-oculography (VOG) is one of eye movement measurement methods. A key problem of VOG is to accurately estimate the pupil center. Then a pupil location method based on morphology and ...

Keyword Extraction Model

Keywords:

- sinusoid track test
- Video-Oculography
- wifi-based VOG system

Term-weighting Scheme

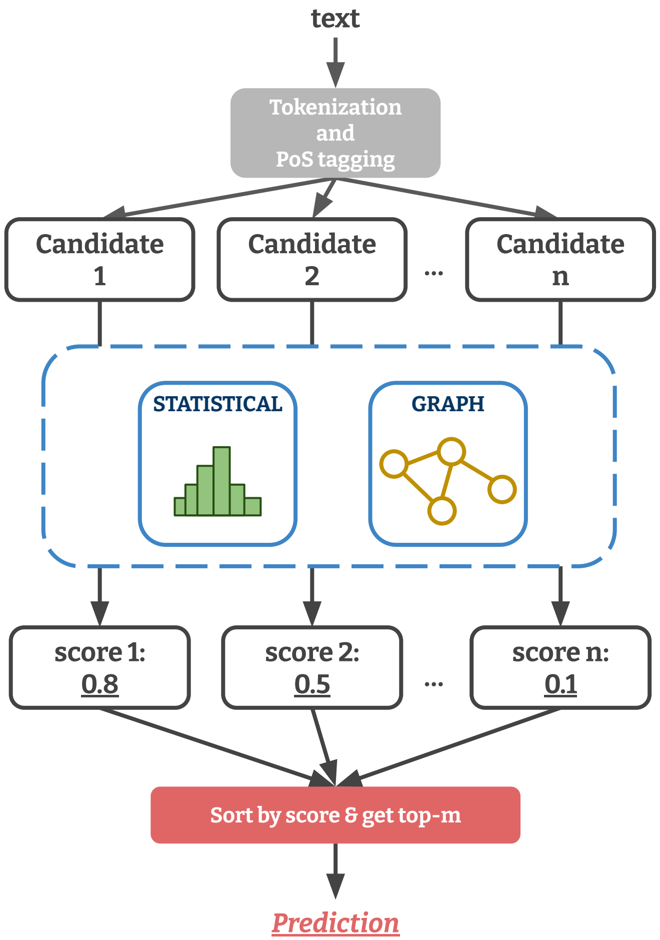
Keyword extraction is a **ranking task**.

Pipeline:

1. Candidate terms
2. Importance score for each term
   ⇒ **Term-weighting Scheme**
3. Top-N terms in terms of the score

Statistical vs **Graph-based**

- Statistics: Term Frequency, TF-IDF
- Graph-based
  - TextRank
  - TopicRank
  - PositionRank



Issues & Our Contribution

No **unified evaluation** in terms of each term-weighting scheme.

Few studies comparing statistical models (**only TF-IDF**).

Issues & Our Contribution

No **unified evaluation** in terms of each term-weighting scheme.

Few studies comparing statistical models (**only TF-IDF**).

Contributions

1. Unified evaluation of **11 models** (7 graph-based and 4 statistical model) over **15 public datasets** in English.
2. Propose new model class based on lexical specificity (**LexSpec**, **LexRank**).
3. Propose a simple extension of TextRank with TFIDF (**TFIDFRank**).

Lexical Specificity

What's lexical specificity?

- Hypergeometric distribution based probabilistic model of words from a text given a corpus (Lafon, 1980).
- The probability of a word *t* randomly appears *k* times in a text of size *n* from a corpus of size *N* containing the *word t* exactly *K* times.

Faster than TF-IDF to compute (Camacho-Collados et al. 2016).

Proposed Algorithms

- **LexSpec:** Lexical specificity as the importance score.
- **LexRank:** TextRank extension with lexical specificity as the bias term.

EXPERIMENTS

Experimental Setup

Datasets: 15 datasets diverse in domain/type.

- English.
- Number of keywords is not fixed.

Metric:

- Precision@5
- Mean Reciprocal Rank (MRR)

Models:

- 7 graph-based models
- 4 statistical models

Data	Size	Domain	Type
KPCrowd	500	-	news
Inspec	2000	CS	abstract
Krapivin2009	2304	CS	article
Nguyen2007	209	-	article
PubMed	500	BM	article
Schutz2008	1231	BM	article
SemEval2010	243	CS	article
SemEval2017	493	-	paragraph
citeulike180	183	BI	article
fao30	30	AG	article
fao780	779	AG	article
theses100	100	-	article
kdd	755	CS	abstract
wiki20	20	CS	report
www	1330	CS	abstract

Result (Precision@5)

LexRank & TFIDFRank achieve the best average metric!

Metric	Dataset	Statistical				Graph-based						
		FirstN	TF	Lex Spec	TFIDF	Text Rank	Single Rank	Position Rank	Lex Rank	TFIDF Rank	Single TPR	Topic Rank
P@5	KPCrowd	35.8	25.3	**39.0**	**39.0**	30.6	30.5	31.8	32.0	32.1	26.9	37.0
	Inspec	31.0	18.9	31.0	31.5	33.2	**33.8**	32.7	32.9	33.3	30.4	31.3
	Krapivin2009	**16.7**	0.1	8.7	7.6	6.6	9.1	14.3	9.7	9.7	7.4	8.5
	Nguyen2007	17.8	0.2	17.2	15.9	13.1	17.3	**20.6**	18.6	18.6	14.0	13.3
	PubMed	9.8	3.6	7.5	6.7	10.1	**10.6**	10.1	8.9	8.8	9.3	7.8
	Schutz2008	16.9	1.6	39.0	38.9	34.0	36.5	18.3	38.9	39.4	14.5	**46.6**
	SemEval2010	15.1	1.5	14.7	12.9	13.4	17.4	**23.2**	16.8	16.6	12.8	16.5
	SemEval2017	30.1	17.0	45.7	**47.2**	41.5	43.0	40.5	46.0	46.4	34.3	36.5
	citeulike180	6.6	9.5	18.0	15.2	23.0	23.9	20.3	23.2	**24.4**	23.7	16.7
	fao30	17.3	16.0	24.0	20.7	26.0	30.0	24.0	29.3	29.3	**32.7**	24.7
	fao780	9.3	3.2	11.7	10.5	12.4	14.3	13.2	13.2	13.1	**14.5**	12.0
	kdd	11.7	7.0	11.2	11.6	10.6	11.5	11.9	11.6	**11.9**	9.4	10.7
	theses100	5.6	0.9	**10.7**	9.4	6.6	7.8	9.3	10.6	9.1	8.3	8.1
	wiki20	13.0	13.0	17.0	21.0	13.0	19.0	14.0	22.0	**23.0**	19.0	16.0
	www	12.2	8.1	11.9	12.2	10.6	11.2	**12.6**	11.6	11.7	10.2	11.2
	AVG	16.6	8.4	20.5	20.0	19.0	21.1	19.8	21.7	**21.8**	17.8	19.8

Result (MRR)

LexRank & TFIDFRank achieve the best average metric.

LexSpec is also competitive.

Metric	Dataset	Statistical: FirstN	Statistical: TF	Statistical: LexSpec	Statistical: TFIDF	Graph-based: TextRank	Graph-based: SingleRank	Graph-based: PositionRank	Graph-based: LexRank	Graph-based: TFIDFRank	Graph-based: SingleTPR	Graph-based: TopicRank
MRR	KPCrowd	60.1	45.5	**73.6**	72.4	62.4	61.6	64.0	65.8	65.2	50.2	60.7
	Inspec	57.3	33.0	52.4	52.8	51.4	52.4	57.1	53.3	53.7	50.5	**57.8**
	Krapivin2009	**36.1**	1.3	22.9	21.0	18.1	22.2	31.4	23.6	23.8	19.1	21.8
	Nguyen2007	43.0	2.8	38.1	41.2	30.8	34.6	**43.2**	36.4	37.9	29.8	33.7
	PubMed	23.1	13.3	23.5	21.4	**31.7**	30.5	30.6	26.9	26.3	26.0	19.8
	Schutz2008	24.6	8.6	76.6	**76.7**	68.9	70.9	38.5	75.5	76.3	33.7	67.3
	SemEval2010	**49.7**	4.5	35.8	34.6	32.9	35.5	47.8	35.3	36.4	28.7	35.9
	SemEval2017	52.0	32.7	68.6	**68.7**	61.4	63.5	62.4	67.3	67.2	54.3	63.7
	citeulike180	20.9	23.6	55.5	47.7	58.2	62.6	51.0	63.0	**65.7**	62.5	40.3
	fao30	31.1	38.3	61.8	49.1	60.2	70.0	48.6	66.1	67.0	**74.6**	50.6
	fao780	17.0	8.5	39.0	35.9	36.1	38.6	35.9	**39.5**	38.9	38.4	31.6
	kdd	26.1	13.0	27.0	27.8	24.5	26.5	28.1	27.9	**28.8**	18.3	26.2
	theses100	15.1	3.1	**32.5**	31.6	23.2	26.3	24.9	31.6	31.1	26.1	26.9
	wiki20	27.5	27.7	**52.7**	47.7	40.1	45.7	31.1	52.2	46.5	39.6	35.5
	www	29.7	17.1	30.5	**30.6**	26.5	27.6	30.4	29.2	30.1	21.7	27.9
	AVG	34.2	18.2	46.0	44.0	41.8	44.6	41.7	46.2	**46.3**	38.2	40.0

Wilcoxon Rank Test

Consider 117,447 documents from all datasets individually.

Wilcoxon rank test results in following groups:

- *TFIDFRank*
- *LexRank, LexSpec*
- *SingleRank, TFIDF*
- *PositionRank, TopicRank*
- *TextRank*
- *FirstN*
- *SingleTPR*
- *TF*

	Method	P@5	MRR
Statistical	FirstN	18.8	37.1
	TF	7.9	16.1
	LexSpec	20.8	42.9
	TFIDF	20.5	42.2
Graph-based	TextRank	19.5	39.2
	SingleRank	21.0	41.2
	PositionRank	20.0	40.9
	LexRank	21.4	42.9
	TFIDFRank	21.6	43.3
	SingleTPR	16.4	33.2
	TopicRank	21.0	40.3

Findings:

- **TFIDFRank** is the best among the groups.
- **LexSpec** slightly but consistently outperforms TFIDF.

Conclusion

- **Proposed new algorithms** (TFIDFRank, LexSpec, and LexRank) and show their efficacy in the experiments.

- Conducted a comprehensive keyword extraction experiments over **15 datasets with 11 models**.

- **Conducted statistical analyses** over the experimental result and provided insights into the performance of each model.

Release of `kex` Library

We release python package **`kex`** (install via **`pip install kex`**), a keyword extraction library including all the models explained in our paper.

Please check our project page https://github.com/asahi417/kex !!

```
>>> import kex
>>> model = kex.SingleRank()  # any algorithm listed above
>>> sample = '''
We propose a novel unsupervised keyphrase extraction approach th
It starts by training word embeddings on the target document to
uses the minimum covariance determinant estimator to model the d
assumption that these vectors come from the same distribution, i
expressed by the dimensions of the learned vector representation
detected as outliers of this dominant distribution. Empirical re
of-the-art and recent unsupervised keyphrase extraction methods.
'''
>>> model.get_keywords(sample, n_keywords=2)
[{'stemmed': 'non-keyphras word vector',
  'pos': 'ADJ NOUN NOUN',
  'raw': ['non-keyphrase word vectors'],
  'offset': [[47, 49]],
  'count': 1,
  'score': 0.06874471825637762,
  'n_source_tokens': 112},
 {'stemmed': 'semant regular word',
  'pos': 'ADJ NOUN NOUN',
  'raw': ['semantic regularities words'],
  'offset': [[28, 32]],
  'count': 1,
  'score': 0.06001468574146248,
  'n_source_tokens': 112}]
```

🌳 Thank you! 🌳