Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
Technical University of Munich

PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction

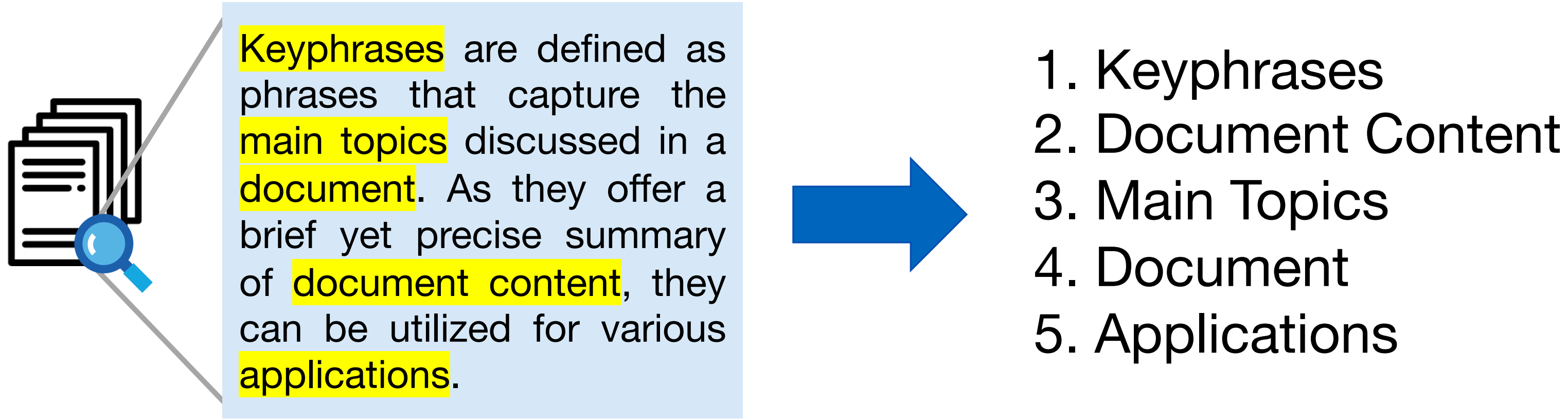
Tim Schopf, Simon Klimek, and Florian Matthes

tim.schopf@tum.de, simon.klimek@tum.de, matthes@tum.de

Motivation

- Provide a quick overview of the content of a text.
- Keyphrases consist of several compound words and can concisely reflect the semantic context of a text.
- Unsupervised keyphrase extraction approaches do not require labeled training data and are mostly domain independent.

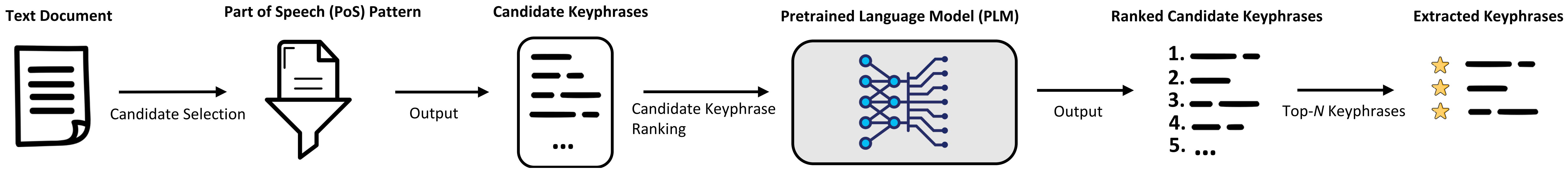
Keyphrase Extraction



Keyphrases are defined as phrases that capture the main topics discussed in a document. As they offer a brief yet precise summary of document content, they can be utilized for various applications.

1. Keyphrases
2. Document Content
3. Main Topics
4. Document
5. Applications

PatternRank



Noun Phrase:

(Zero or more adjectives followed by one or more nouns.)

$$\{ADJ\} * \{\mathrm{NOUN}\} +$$

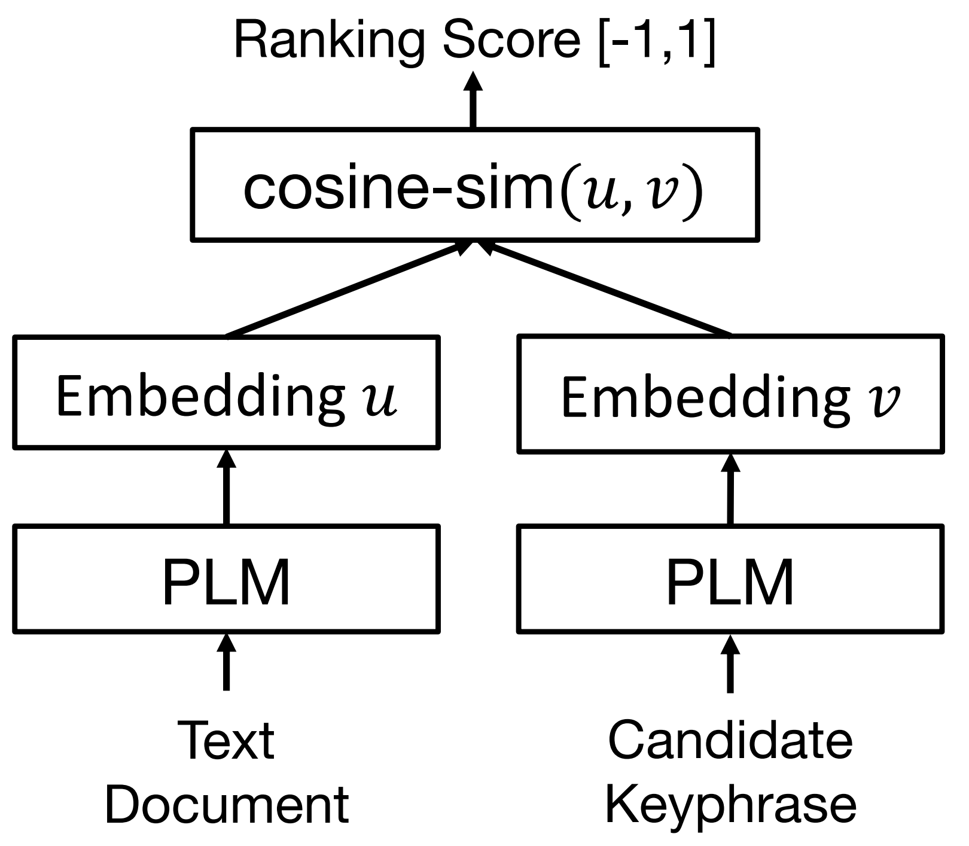
Part of Speech Pattern:

(Arbitrary parts-of-speech separated by a hyphen, followed by zero or more nouns OR zero or more verb, followed by zero or more adjectives, followed by one or more nouns.)

$$((\{.*\}\{HYPH\}\{.*\})\{NOUN\}*)|$$
$$((\{VBG\}|\{VBN\})?\{ADJ\}*\{NOUN\}+)$$

SCAN ME

PLM-based Candidate Keyphrase Ranking:



Approach

1. The input consists of a single text document which is being word tokenized.
2. The word tokens are then tagged with part-of-speech tags.
3. Tokens whose tags match a previously defined part-of-speech pattern are selected as candidate keyphrases.
4. Then, a pretrained language model embeds the entire text document as well as all candidate keyphrases as semantic vector representations.
5. Subsequently, the cosine similarities between the document representation and the candidate keyphrase representations are computed and the candidate keyphrases are ranked in descending order based on the computed similarity scores.
6. Finally, the top-N ranked keyphrases, which are most representative of the input document, are extracted.

Evaluation

Method	F_1@5	F_1@10	F_1@20
YAKE	15.37	18.50	19.65
SingleRank	21.97	28.55	30.80
KeyBERT	7.82	10.30	11.76
PatternRank $_{NP}$	23.92	29.66	29.19
PatternRank $_{PoS}$	**24.35**	**30.99**	**31.37**

- Inspec dataset consisting of 2,000 English computer science abstracts. Each abstract has assigned a set of gold keyphrases.
- Evaluation based on exact match of extracted keyphrases and gold keyphrases.

Contact

Tim Schopf
Technical University of Munich
E-Mail: tim.schopf@tum.de
Phone: +49 89 289-17105