Automatic Keyphrase Extraction: A Survey of the State of the Art

Kazi Saidul Hasan and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{saidul,vince}@hlt.utdallas.edu

Zihao Li, 2020.09.28

Concept Definition

- ***What is a Keyphrase:*** a set of phrases that are related to the main topics discussed in a given document (Tomokiyo and Hurst, 2003; Liu et al., 2009b; Ding et al., 2011; Zhao et al., 2011);

- ***What is the Automatic Keyphrase Extraction:*** "the automatic selection of important and topical phrases from the body of a document" (Turney, 2000);

- What is a Keyword: International Encyclopedia of Information and Library Science defines "keyword" as "A word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document." Both single words (keywords) and phrases (keyphrases) may be referred to as "key terms";

- ***What is the difference between Keyphrase and Keyword:*** A keyphrase connotes a multi-word lexeme (e.g. computer science engineering, hard disk), whereas a keyword is a single word term (e.g. computer, disk);

Research Values

Document keyphrases have enabled fast and accurate searching for a given document from a large text collection, and have exhibited their potential in improving many natural language processing (NLP) and information retrieval (IR) tasks, such as text summarization (Zhang et al., 2004), text categorization (Hulth and Megyesi, 2006), opinion mining (Berend, 2011), and document indexing (Gutwin et al., 1999);

Canadian **Ben Johnson** left the **Olympics** today “in a complete state of shock,” accused of cheating with drugs in the world’s fastest **100-meter dash** and stripped of his **gold medal**. The prize went to American **Carl Lewis**. Many athletes accepted the accusation that Johnson used a muscle-building but dangerous and illegal anabolic steroid called **stanozolol** as confirmation of what they said they know has been going on in track and field. Two tests of Johnson’s urine sample proved positive and his denials of **drug use** were rejected today. “This is a blow for the Olympic Games and the Olympic movement,” said International Olympic Committee President Juan Antonio Samaranch.

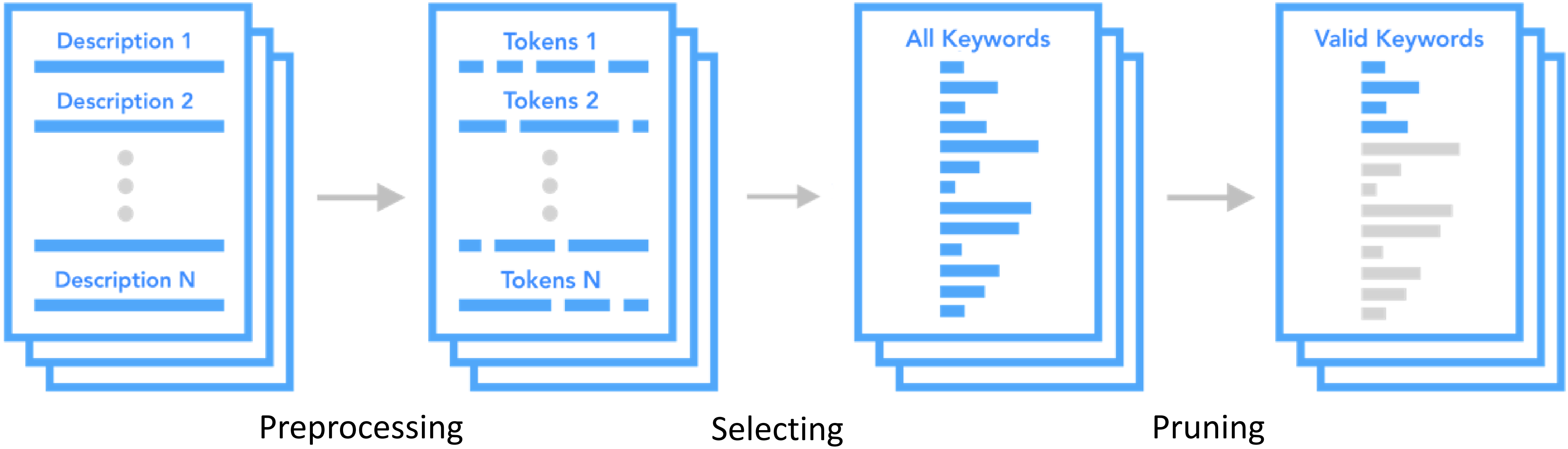
Figure 1: A news article on *Ben Johnson* from the DUC-2001 dataset. The keyphrases are boldfaced.

The Pipeline of Keyphrase Extraction

A keyphrase extraction system typically operates in two steps:

- Candidate Selecting: extracting a list of words/phrases that serve as candidate keyphrases using some heuristics;

- Pruning: determining which of these candidate keyphrases are correct keyphrases using supervised or unsupervised approaches;

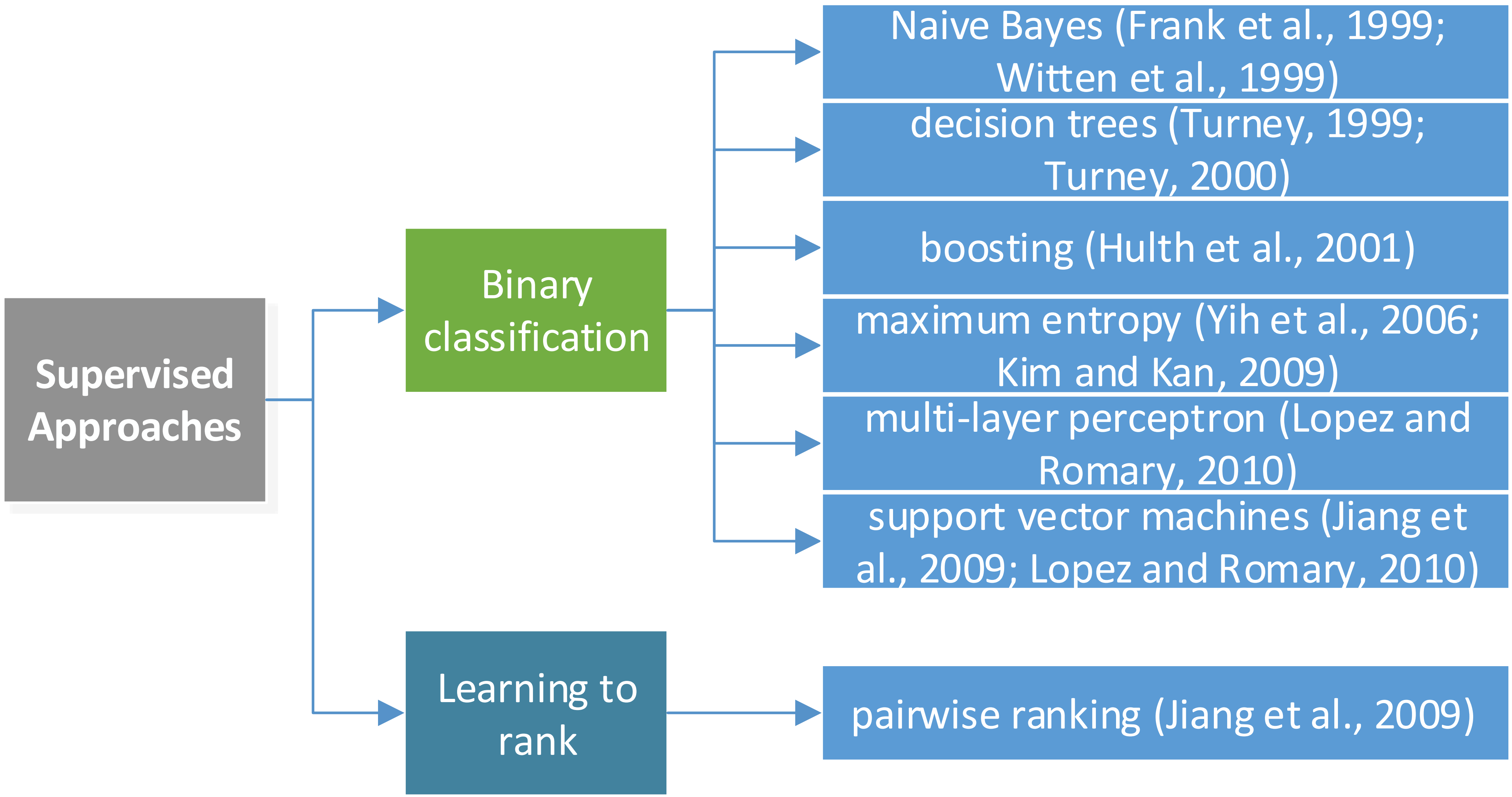


Selecting Candidate Words and Phrases

Applying heuristic rules to extract a set of phrases and words as candidate keyphrases:

- Using a stop word list to remove stop words (Liu et al., 2009b): meaningless words such as the, is, at, on, etc.;
- Allowing words with certain part-of-speech tags (POS) to be candidate keywords (Mihalcea and Tarau, 2004; Wan and Xiao, 2008b; Liu et al., 2009a): regard nouns or noun phrases as candidate of keyphrases;
- Allowing n-grams that appear in Wikipedia article titles to be candidate (Grineva et al., 2009): incorporating background knowledge;
- Extracting n-grams (Witten et al., 1999; Hulth, 2003; Medelyan et al., 2009) or noun phrases (Barker and Cornacchia, 2000; Wu et al., 2005) that satisfy pre-defined lexico-syntactic pattern(s) (Nguyen and Phan, 2009): AP<property>NP<class> (starry night);

Supervised Approaches (task reformulation)



Supervised Approaches (feature design)

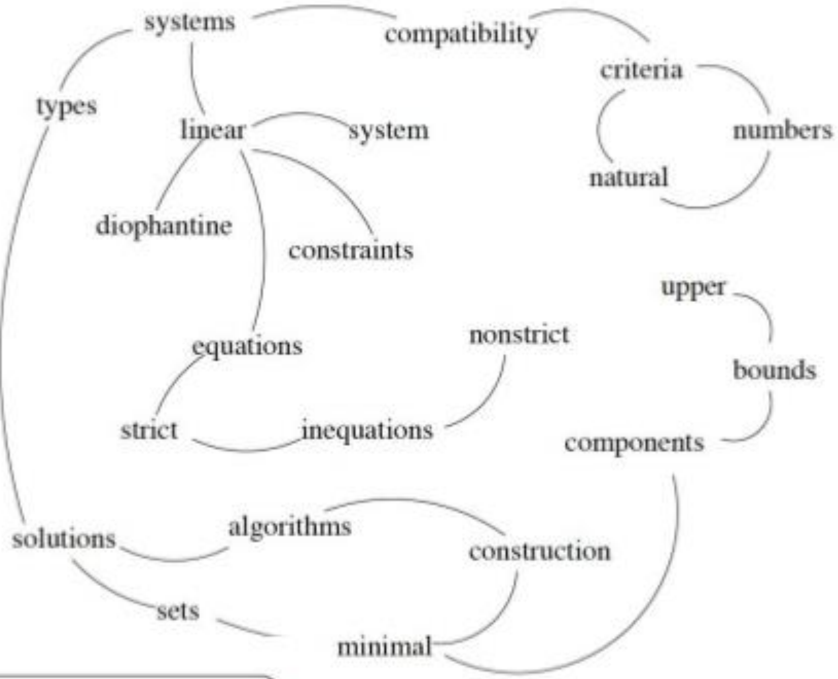
- Features
  - Within-Collection Features
    - Statistical features
      - Tf*idf (Salton and Buckley, 1988)
      - distance of a phrase
      - supervised keyphraseness
      - phrase length and spread
    - Structural features
      - how frequently a candidate keyphrase occurs in various sections of a scientific paper (e.g., introduction, conclusion) (Nguyen and Kan, 2007)
      - the location of a candidate keyphrase in a web page (e.g., whether it appears in the title) (Chen et al., 2005; Yih et al., 2006)
    - Syntactic features
      - PoS tag sequence
      - suffix sequence (Yih et al., 2006; Nguyen and Kan, 2007; Kim and Kan, 2009)
  - External Resource-Based Features
    - Wikipedia-based keyphraseness (Medelyan et al., 2009)
    - query log of a search engine (Yih et al., 2006)
    - coherence features (Turney, 2003)

Unsupervised Approaches (Graph-Based Ranking)

- Candidate keyphrase as node;
- Co-occurrence counts (Mihalcea and Tarau, 2004; Matsuo and Ishizuka, 2004) and semantic relatedness (Grineva et al., 2009) as weight of edge;
- Random walk extracts keyphrases;

TextRank (Mihalcea and Tarau, 2004)

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

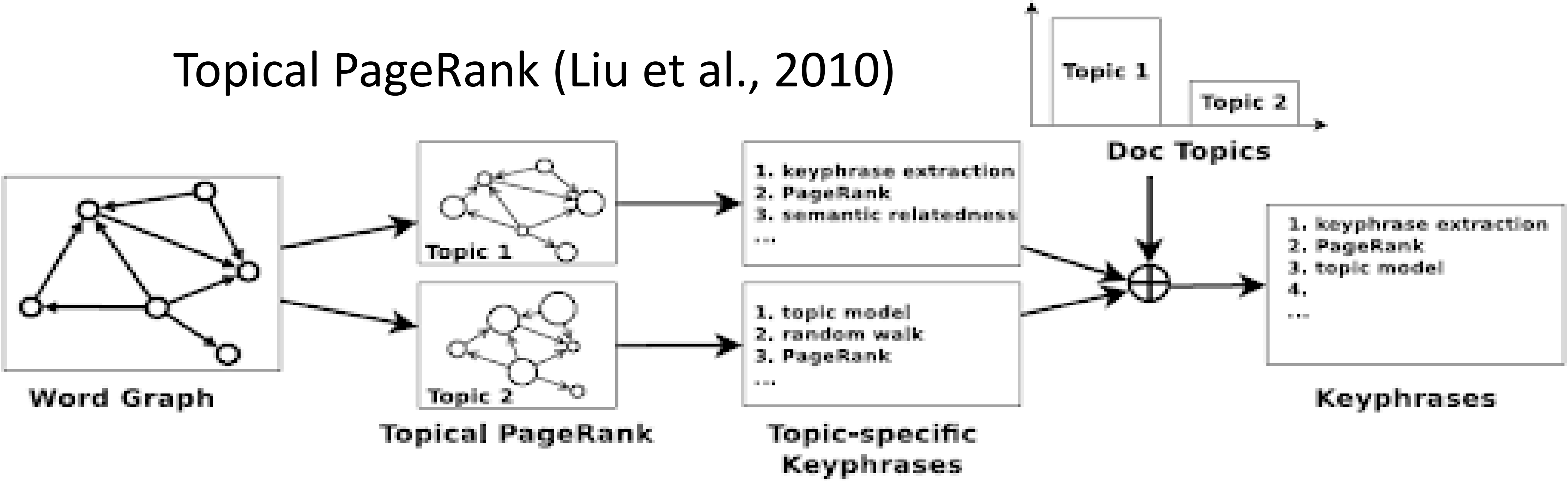


Keywords assigned by TextRank:
linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:
linear constraints; linear diophantine equations; minimal generating sets; non–strict inequations; set of natural numbers; strict inequations; upper bounds

Unsupervised Approaches (Topic-Based Clustering)

Topical PageRank (Liu et al., 2010)



- Defining several random jump probability as topic-specific preference value:

$$R_z(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j, w_i)}{O(w_j)} R_z(w_j) + (1-\lambda) p_z(w_i), s.t. \sum_{w \in V} p_z(w_i) = 1$$

- Combining topic distribution(LDA) with topic-specific keyphrases to acquire final keyphrases:

$$R(p) = \sum_{z=1}^{K} R_z(p) \times pr(z \mid d), R_z(p) = \sum_{w_i \in p} R_z(w_i)$$

Unsupervised Approaches (Language Modeling)

Assumption:

- a sentence is important if it contains important words, and important words appear in important sentences;
- an important sentence is connected to other important sentences;
- an important word is linked to other important words;

1. Compute and normalize the scores of sentences:

$$\boldsymbol{u}^{(n)} = \alpha \tilde{\boldsymbol{U}}^{T} \boldsymbol{u}^{(n\text{-}1)} + \beta \hat{\boldsymbol{W}}^{T} \boldsymbol{v}^{(n\text{-}1)},$$

$$\boldsymbol{u}^{(n)} = \boldsymbol{u}^{(n)} / \left\| \boldsymbol{u}^{(n)} \right\|_1$$

2. Compute and normalize the scores of words:

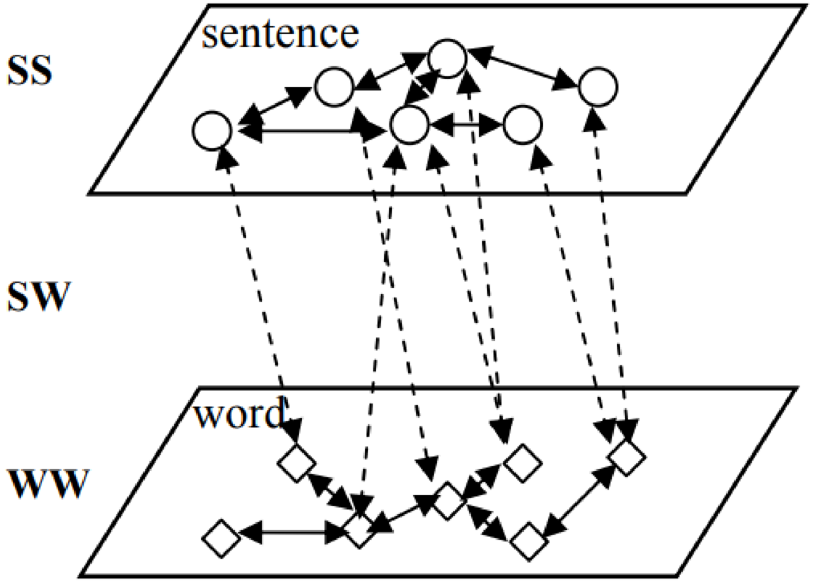
$$\boldsymbol{v}^{(n)} = \alpha \tilde{\boldsymbol{V}}^{T} \boldsymbol{v}^{(n\text{-}1)} + \beta \tilde{\boldsymbol{W}}^{T} \boldsymbol{u}^{(n\text{-}1)},$$

$$\boldsymbol{v}^{(n)} = \boldsymbol{v}^{(n)} / \left\| \boldsymbol{v}^{(n)} \right\|_1$$

where $\boldsymbol{u}^{(n)}$ and $\boldsymbol{v}^{(n)}$ denote the vectors computed at the n-th iteration.

Denote:

U, V, W: each entry corresponding to the weight of a link in SS-Graph, WW-Graph and SW-Graph respectively;

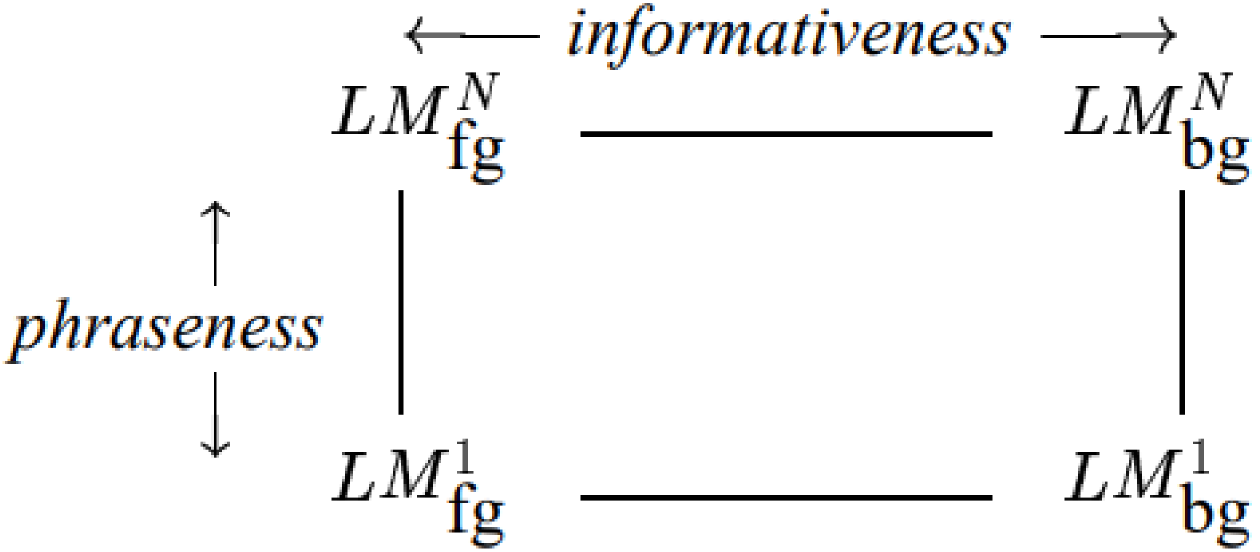


SS-Relationship: It reflects the homogeneous relationships between sentences, usually computed by their content similarity.

WW-Relationship: It reflects the homogeneous relationships between words, usually computed by knowledge-based approach or corpus-based approach.

SW-Relationship: It reflects the heterogeneous relationships between sentences and words, usually computed as the relative importance of a word in a sentence.

Unsupervised Approaches (Language Modeling)



$$P(\mathbf{w}) = \prod_{i=1}^{n} P(w_i | w_1 w_2 \ldots w_{i-1})$$

- Phraseness: the extent to which a word sequence can be treated as a phrase;
- Informativeness: the extent to which a word sequence captures the central idea of the document it appears in;

Phraseness of **w** is how much we lose information by assuming independence of each word by applying the unigram model, instead of the N-gram model.

$$\delta_{\mathbf{w}}(LM_{\text{fg}}^{N} \parallel LM_{\text{fg}}^{1}) \tag{8}$$

Informativeness of **w** is how much we lose information by assuming the phrase is drawn from the background model instead of the foreground model.

$$\delta_{\mathbf{w}}(LM_{\text{fg}}^{N} \parallel LM_{\text{bg}}^{N}), \text{ or} \tag{9}$$

$$\delta_{\mathbf{w}}(LM_{\text{fg}}^{1} \parallel LM_{\text{bg}}^{1}) \tag{10}$$

Combined The following is considered to be a mixture of phraseness and informativeness.

$$\delta_{\mathbf{w}}(LM_{\text{fg}}^{N} \parallel LM_{\text{bg}}^{1}) \tag{11}$$

SOTA (The State of the Art)

Dataset	Approach and System [Supervised?]	Score		
		P	R	F
Abstracts (*Inspec*)	Topic clustering (Liu et al., 2009b) [×]	35.0	66.0	45.7
Blogs	Topic community detection (Grineva et al., 2009) [×]	35.1	61.5	44.7
News (DUC -2001)	Graph-based ranking for extended neighborhood (Wan and Xiao, 2008b) [×]	28.8	35.4	31.7
Papers (SemEval -2010)	Statistical, semantic, and distributional features (Lopez and Romary, 2010) [✓]	27.2	27.8	27.5

Challenge

- ***Length:*** The difficulty of the task increases with the length of the input document as longer documents yield more candidate keyphrases. Consequently, it is harder to extract keyphrases from scientific papers, technical reports, and meeting transcripts than abstracts, emails, and news articles;

- ***Structural consistency:*** In a structured document, there are certain locations where a keyphrase is most likely to appear. In contrast, the lack of structural consistency in other types of structured documents (e.g., web pages, which can be blogs, forums, or reviews) may render structural information less useful;

- ***Topic change:*** An observation commonly exploited in keyphrase extraction from scientific articles and news articles is that keyphrases typically appear not only at the beginning (Witten et al., 1999) but also at the end (Medelyan et al., 2009) of a document. This observation does not necessarily hold for conversational text (e.g., meetings, chats), however;

Challenge

- ***Overgeneration and Redundancy:*** are a major type of precision error, contributing to 36–59% of the overall error. Overgeneration errors occur when a system correctly predicts a candidate as a keyphrase because it contains a word that appears frequently in the associated document, but at the same time erroneously outputs other candidates as keyphrases because they contain the same word. Redundancy errors occur when a system correctly identifies a candidate as a keyphrase, but at the same time outputs a semantically equivalent candidate (e.g., its alias) as a keyphrase. For instance, *Olympic(s)* and *Olympic movement and Olympic games*;

- ***Infrequency:*** are a major type of recall error contributing to 24-27% of the overall error. Infrequency errors occur when a system fails to identify a keyphrase owing to its infrequent presence in the associated document (Liu et al., 2011). Like, *100-meter dash* and *gold medal*;

- ***Evaluation:*** are a type of recall error contributing to 7–10% of the overall error;

Conclusion and Future Directions

- ***Incorporating background knowledge.*** While much recent work has focused on algorithmic development, keyphrase extractors need to have a deeper "understanding" of a document in order to reach the next level of performance. Such an understanding can be facilitated by the incorporation of background knowledge;

- ***Handling long documents.*** While it may be possible to design better algorithms to handle the large number of candidates in long documents, we believe that employing sophisticated features, especially those that encode background knowledge, will enable keyphrases and non-keyphrases to be distinguished more easily even in the presence of a large number of candidates;

- ***Improving evaluation schemes.*** To more accurately measure the performance of keyphrase extractors, they should not be penalized for evaluation errors. We have suggested several possibilities as to how this problem can be addressed;

Reference

- Bharti S K, Babu K S. Automatic keyword extraction for text summarization: A survey[J]. arXiv preprint arXiv:1704.03242, 2017.

- Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.

- Liu Z, Huang W, Zheng Y, et al. Automatic keyphrase extraction via topic decomposition[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. 2010: 366-376.

- Wan X, Yang J, Xiao J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction[C]//Proceedings of the 45th annual meeting of the association of computational linguistics. 2007: 552-559.

- Tomokiyo T, Hurst M. A language model approach to keyphrase extraction[C]//Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment. 2003: 33-40.

- Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: a literature review[J]. International Journal of Computer Applications, 2015, 109(2).

Q&A

Thanks