From Fundamentals to Recent Advances
A Tutorial on Keyphrasification

Part 2.1 Deep Learning Methods for Keyphrase Extraction

Rui Meng, Debanjan Mahata, Florian Boudin

ECIR 2022

Part II
Neural Methods for Keyphrasification

Outline of Part II

Part I - Neural Keyphrase Extraction

Part II - Neural Keyphrase Generation

Part III - Hands-on Practice with OpenNMT-kpg and DLKP

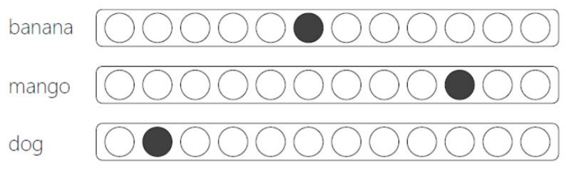
Where we are

	Traditional methods	Neural-based methods
extraction	statistical (e.g. TF.IDF) graph-based (e.g. TextRank) supervised feature-based (e.g. Kea) sequence labeling: CRF (Das Gollapalli et al., 2017)	Bi-LSTM-CRF (Alzaidy et al. 2019) embedding-based (e.g. Key2Vec) (Mahata et al., 2018)
generation	resource-based (e.g. Kea++) (Medelyan and Witten, 2006)	sequence-to-sequence (e.g. CopyRNN)

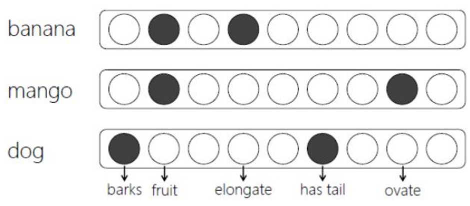
(Medelyan and Witten, 2006) Thesaurus based automatic keyphrase indexing. JCDL.
(Das Gollapalli et al., 2017) Incorporating expert knowledge into keyphrase extraction. AAAI.
(Mahata et al., 2018) Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. NAACL.
(Alzaidy et al. 2019) Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents. WWW.

Why Deep Learning?

- Distributed representation is natively better for representing semantics



(a) Local representation



(b) Distributed representation

Figure credit:
https://www.microsoft.com/en-us/research/uploads/prod/2018/04/NeuralIR-Nov2017.pdf

Why Deep Learning?

- Enables end-to-end learning
  - Get rid of manual feature engineering
  - Learn keyness/phraseness from data directly

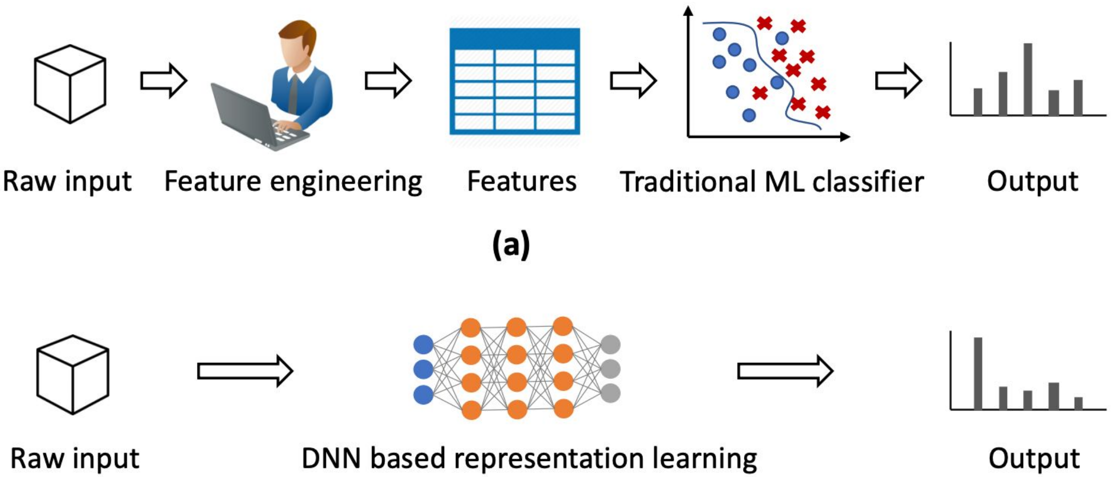
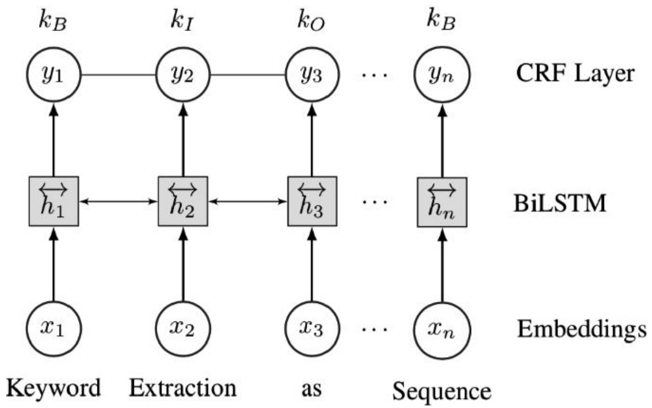


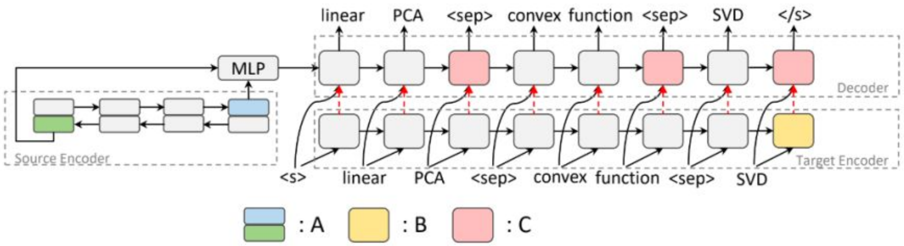
Figure credit: https://arxiv.org/pdf/1808.00033v1.pdf

Neural Keyphrasification

- Neural Keyphrase Extraction

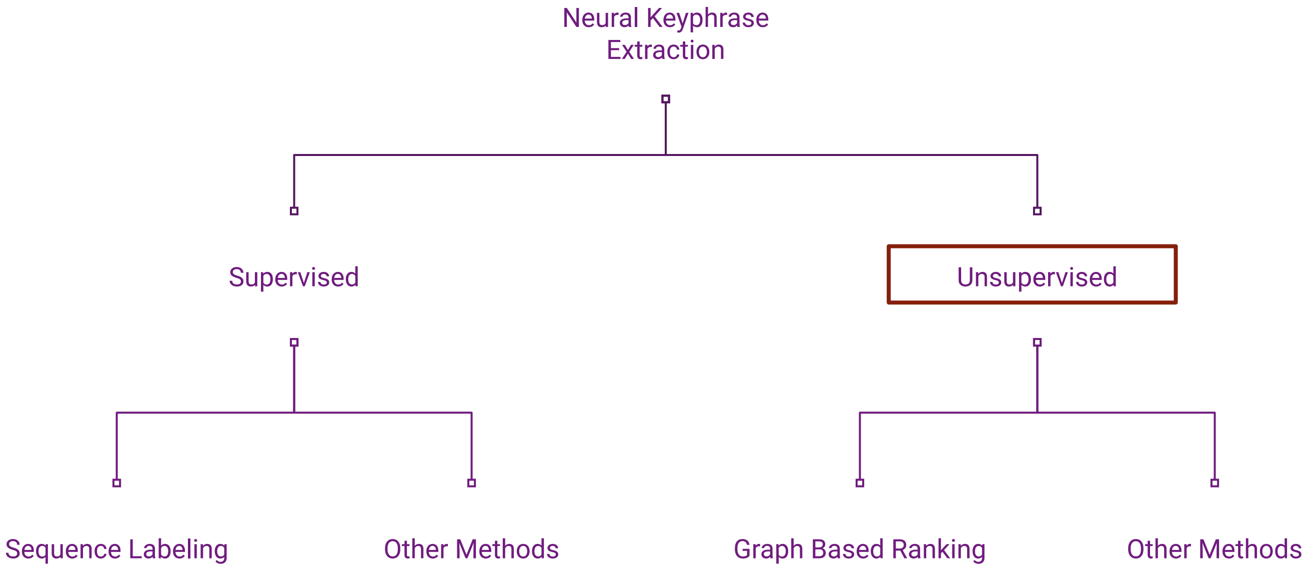


- Neural Keyphrase Generation

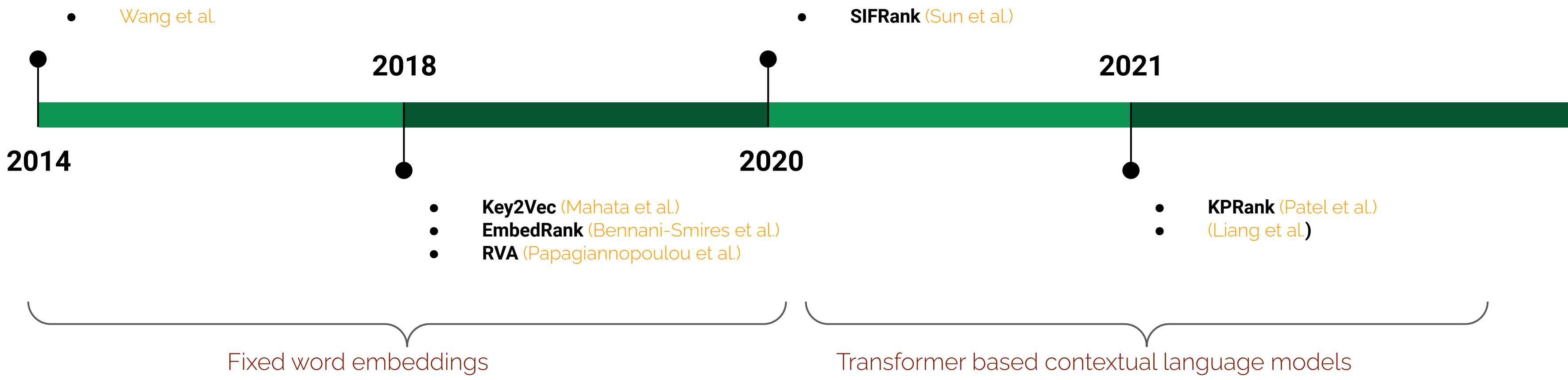


(Sahrawat, et al. 2019), Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings." arXiv.
(Yuan, et al. 2020) "One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases." ACL.

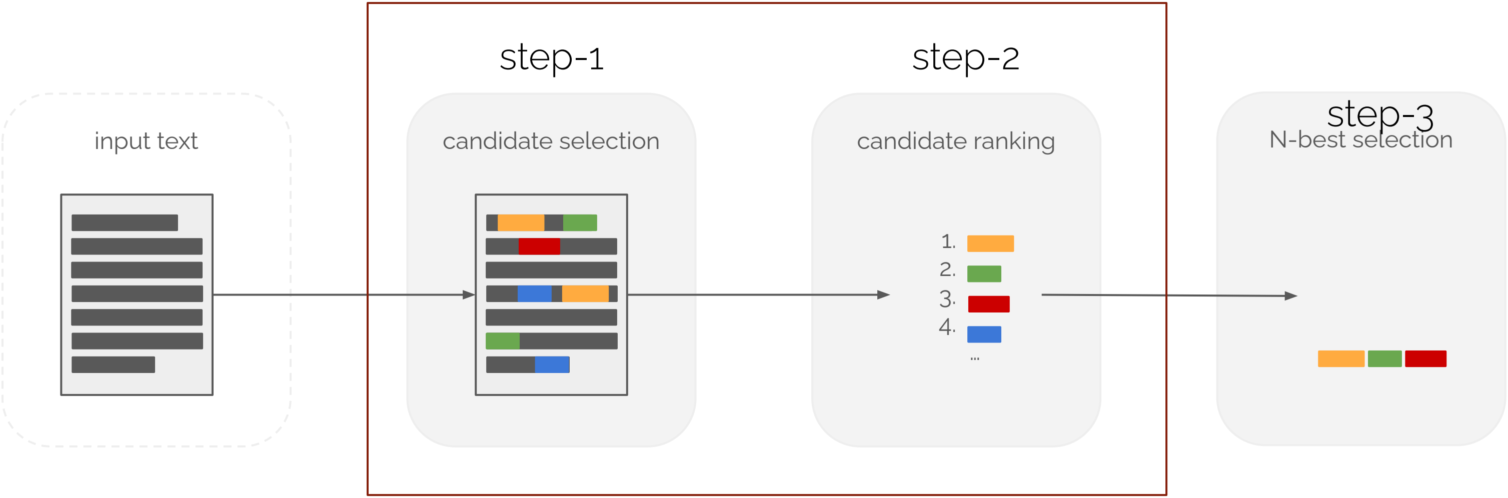
Taxonomy of Extractive Methods



Unsupervised Algorithms



Basic Framework



Use of Word/Phrase/Sentence/Document Embeddings

Apply Graph based ranking

Basic Steps

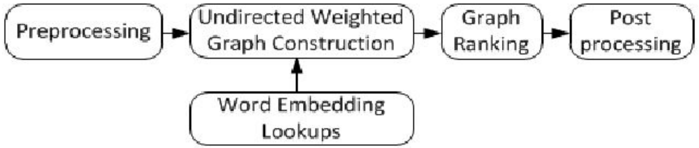
1. Candidate keyphrase selection - using some text processing heuristic

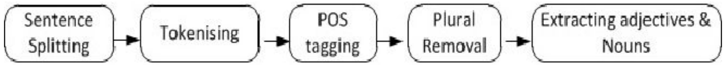
2. Reference representation - create a reference representation of the input document using newly trained or pre-trained word/phrase/sentence/document embeddings

3. Candidate keyphrase representation - get candidate keyphrase representations using newly trained or pre-trained word/phrase/sentence/document embeddings

4. Rank candidate keyphrases - by using the similarity scores between reference and candidate keyphrase representations

Using Word Embeddings for Keyphrases Extraction



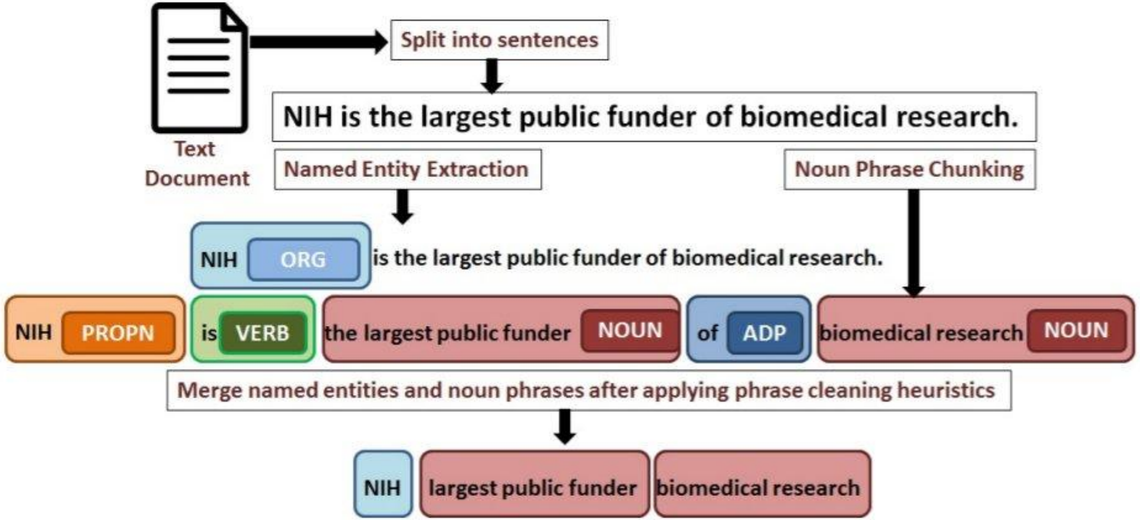


- SENNA Word embeddings are used as background knowledge
- Weighting scheme
  - Informativeness and phraseness scores of words
  - Word embeddings + local statistical information
- Undirected graph of words with edges determined by their co-occurrence
- Ranked using PageRank

Wang, R., Liu, W., & McDonald, C. (2014, July). Corpus-independent generic keyphrase extraction using word embedding vectors. In *Software engineering research conference* (Vol. 39, pp. 1-8).

Key2Vec

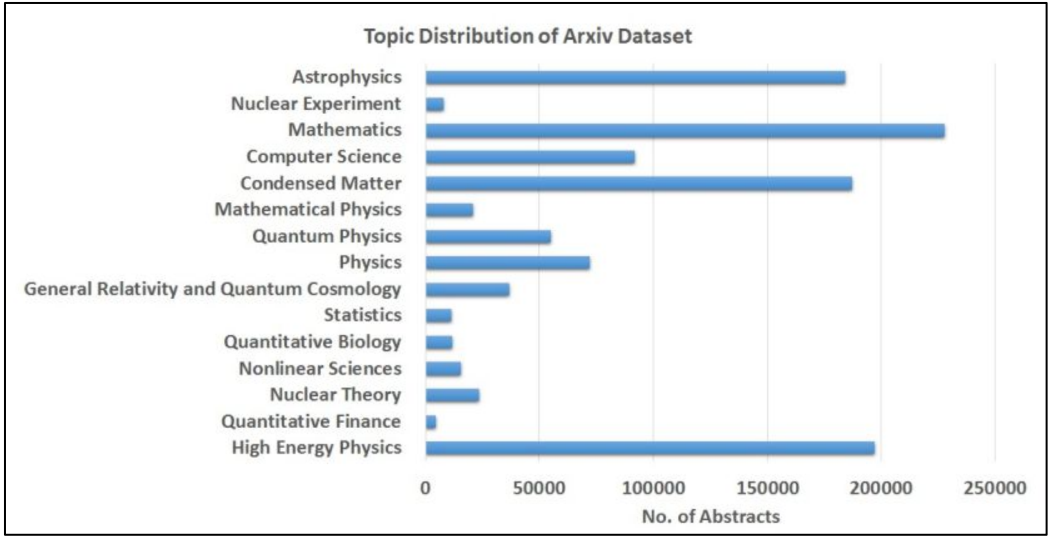
1. **Candidate selection**



Mahata, D., Kuriakose, J., Shah, R., & Zimmermann, R. (2018, June). Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 634-639).

Key2Vec - Learning Phrase Embeddings

1,149,244 arxiv abstracts → Text Processing → FastText PhraseEmbeddings



Phrase	Top 5 Similar Phrases
convolutional_neural_network	cnn, feature_representations, deep_convolutional_neural_network, deep_neural_network, scene_recognition
dark_matter	dm, dark_matter_particle, non-baryonic_dark_matter, dark_energy, self-interacting_dark_matter
natural_language_processing	nlp, language_processing, machine_translation, named_entity_recognition, sense_disambiguation
rnn	blstm, long_short-term_memory, lstms, handwritten_documents, recurrent_neural_network, lstm
svm	support_vector_machine, support_vector_machines, random_forest, svms, naive_bayes

Key2Vec - Theme Vector

Title: Identification of states of complex systems with estimation of admissible measurement errors on the basis of fuzzy information.

Abstract: The problem of identification of states of complex systems on the basis of fuzzy values of informative attributes is considered. Some estimates of a maximally admissible degree of measurement error are obtained that make it possible, using the apparatus of fuzzy set theory, to correctly identify the current state of a system.

complex systems

0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7

admissible measurement errors

0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6

fuzzy information

0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Theme Vector

0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7

$\oplus$

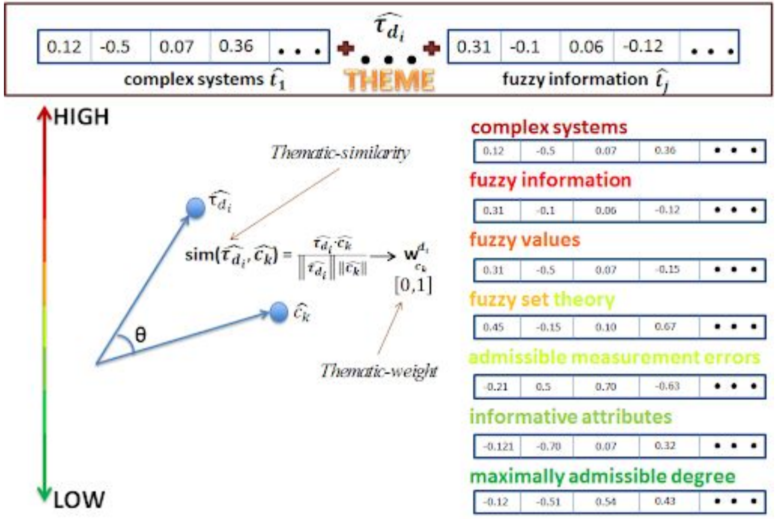
0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6

$\oplus$

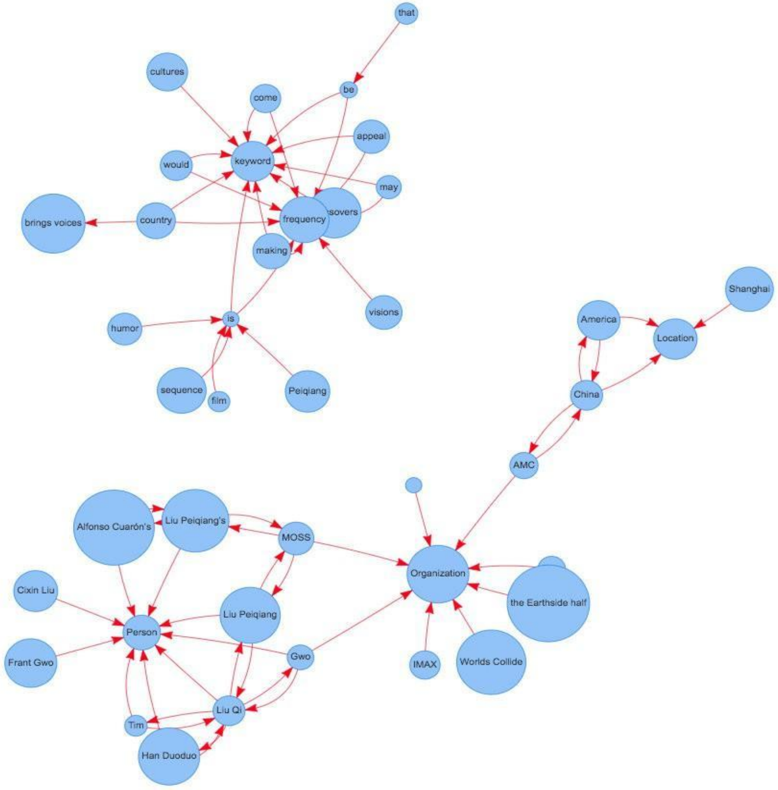
0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Key2Vec - Candidate Selection and Scoring

- complex systems
- fuzzy information
- fuzzy values
- fuzzy set theory
- admissible measurement errors
- informative attributes
- maximally admissible degree



Key2Vec - Ranking



Edge Weights

$$semantic(c_j^{d_i}, c_k^{d_i}) = \frac{1}{1 - cosine(c_j^{d_i}, c_k^{d_i})}$$

$$cooccur(c_j^{d_i}, c_k^{d_i}) = PMI(c_j^{d_i}, c_k^{d_i})$$

$$sr(c_j^{d_i}, c_k^{d_i}) = semantic(c_j^{d_i}, c_k^{d_i}) \times cooccur(c_j^{d_i}, c_k^{d_i})$$

Theme Biased PageRank

$$R(c_j^{d_i}) = (1-d)w_{c_j}^{d_i} + d \times \sum_{c_k^{d_i} \in \varepsilon(c_j^{d_i})} \left(\frac{sr(c_j^{d_i}, c_k^{d_i})}{|out(c_k^{d_i})|}\right) R(c_k^{d_i})$$

Steps for Neural Unsupervised Keyphrase Extraction

Input Document/Text

Ranked Keyphrases

Candidate Selection	Reference Selection	Candidate and Document Representation	Candidate Ranking
❖ Noun Phrases ❖ Named Entities ❖ N-grams	❖ Title ❖ Title + Abstract ❖ Document	❖ Fasttext ❖ Word2Vec ❖ Glove ❖ Doc2Vec ❖ BERT ❖ ELMO ❖ SIF ❖ SciBERT	❖ PageRank ❖ Biased PageRank ❖ Cosine Similarity ❖ Boundary Aware Centrality

EmbedRank

1. Candidate keyphrase selection - only those phrases that consists of zero or more adjectives followed by one or multiple nouns

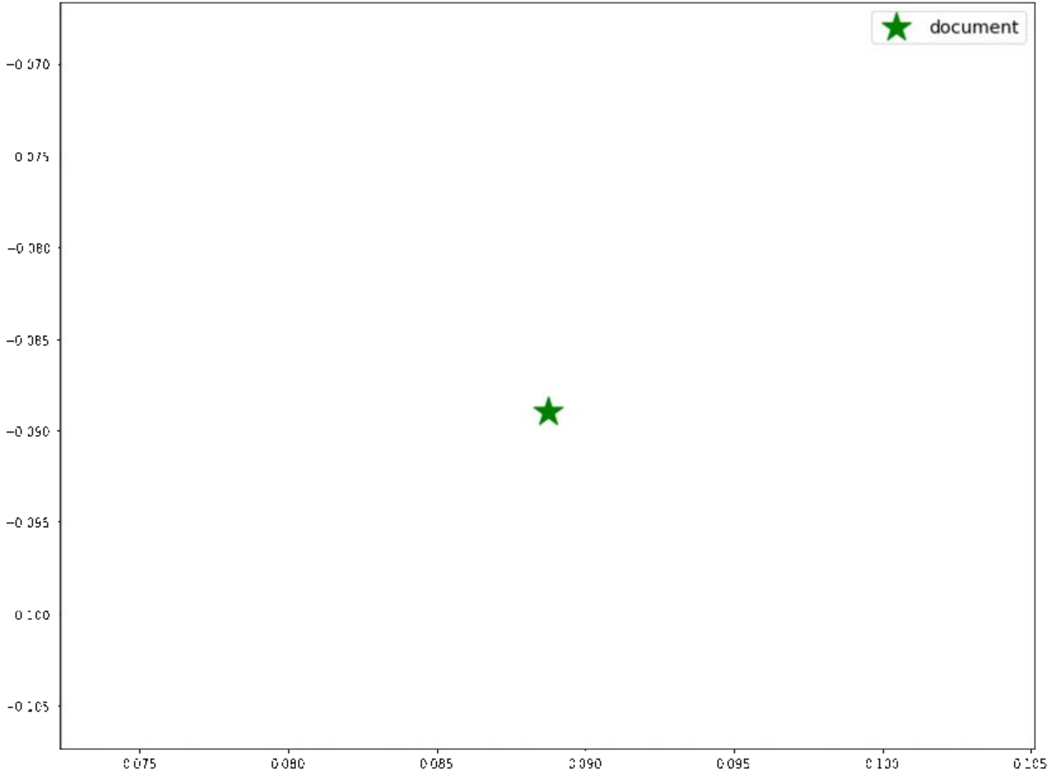
2. Reference representation - represents adjectives and nouns in the document using sent2vec and doc2vec

3. Candidate keyphrase representation - gets embeddings of each candidate phrase using sent2vec and doc2vec

4. Candidate ranking - ranks the candidate keyphrases according to their cosine distance to the document representation

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:1801.04470*.

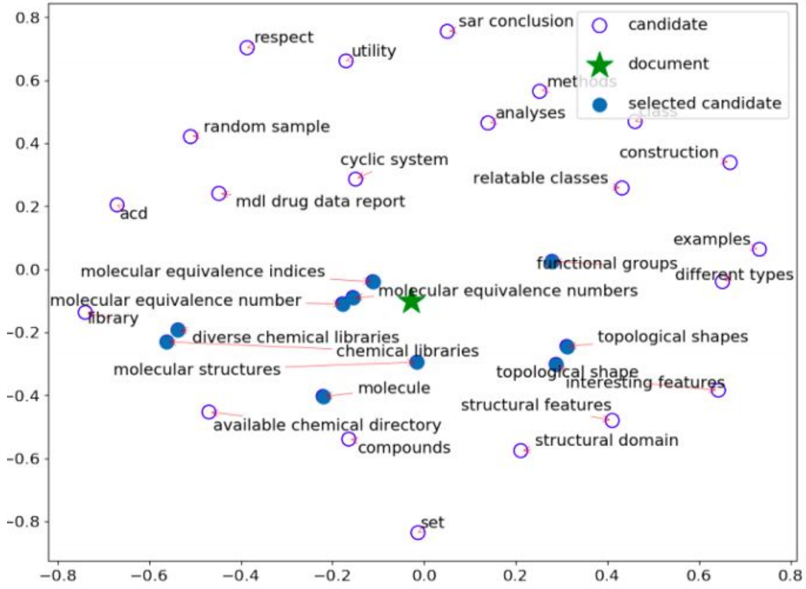
EmbedRank



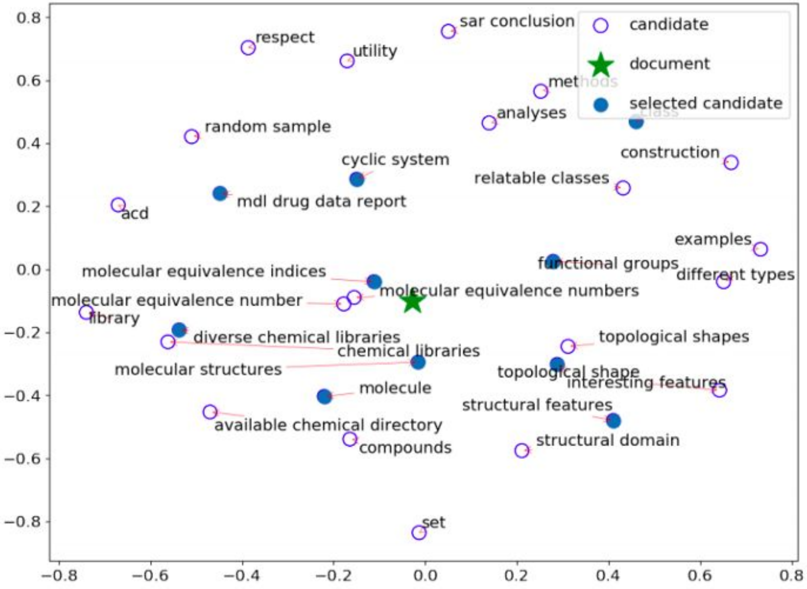
- Embeds the document and the candidate keyphrases in the same embedding space

- Performance obtained using doc2vec and sent2vec were comparable. However, doc2vec was slower than sent2vec

EmbedRank++



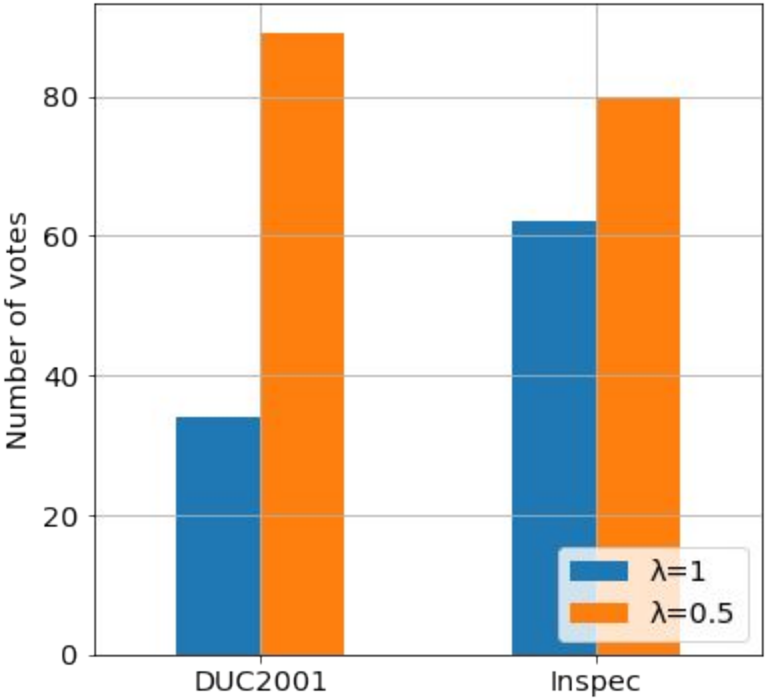
(a) EmbedRank (without diversity)



(b) EmbedRank++ (with diversity)

Maximal Marginal Relevance (MMR)

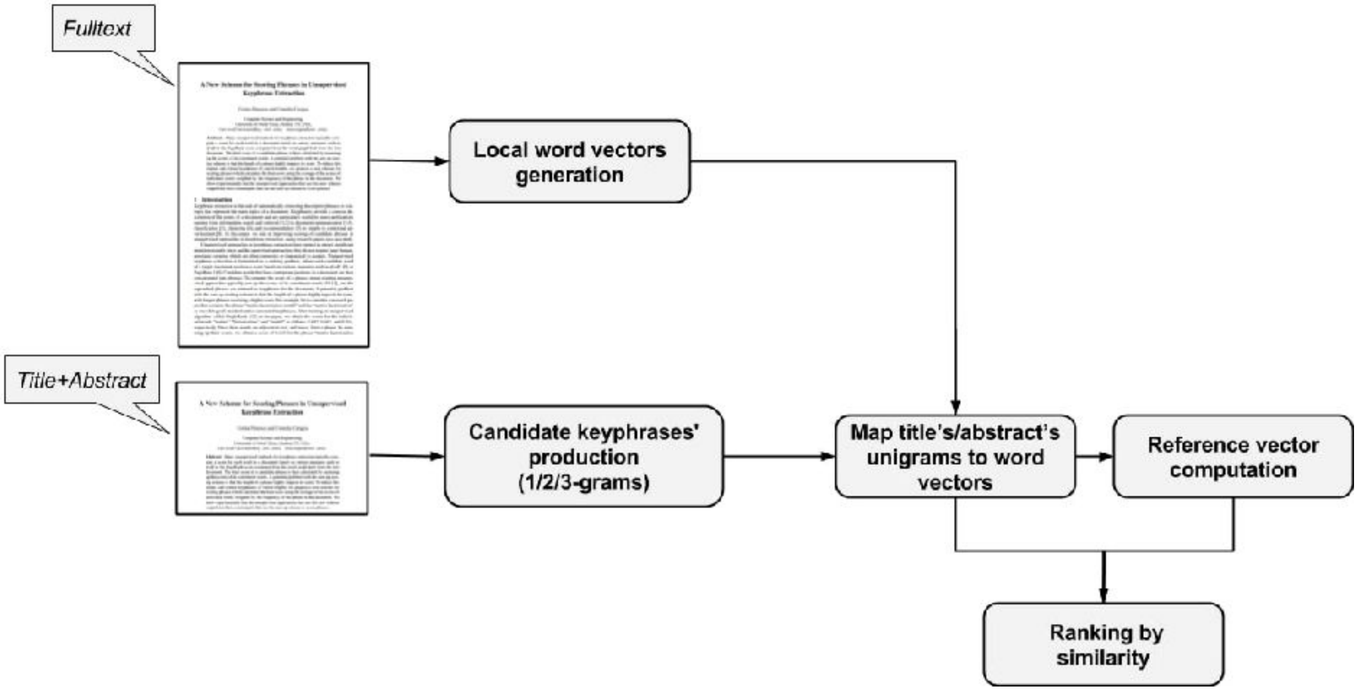
Humans liked diversity



- User study among 20 documents from Inspec and 20 documents from DUC2001. Users were asked to choose their preferred set of keyphrases between the one extracted with EmbedRank++ (λ = 0.5) and the one extracted with EmbedRank (λ = 1).
- EmbedRank++ did not perform better on the automated evaluation using F1@5, F1@10 and F1@15 performance metrics

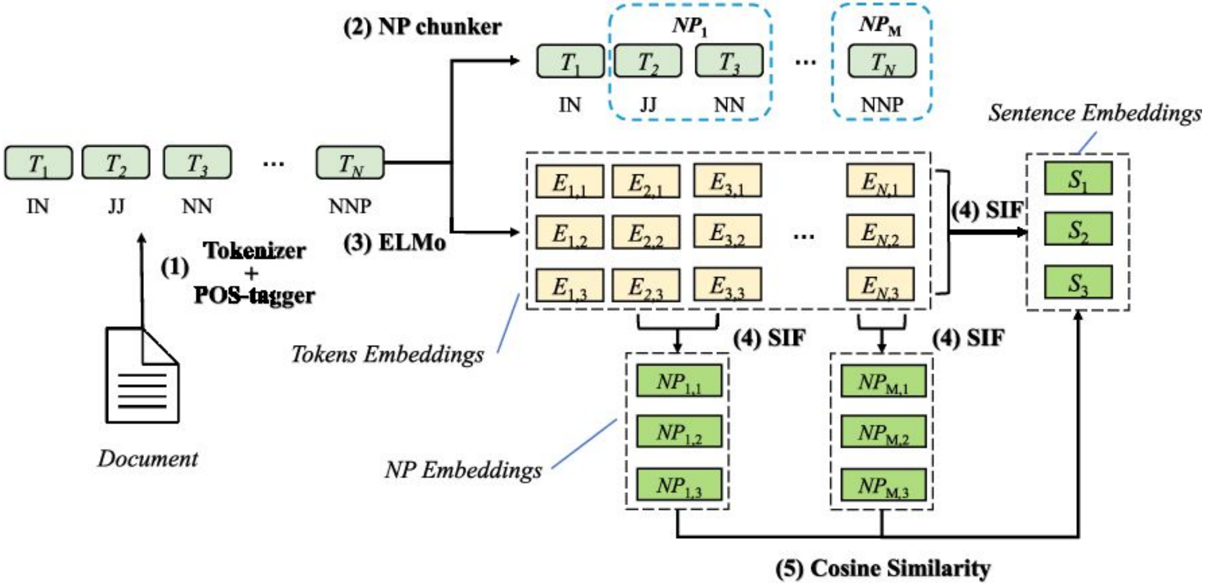
Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:1801.04470*.

Reference Vector Algorithm (RVA)



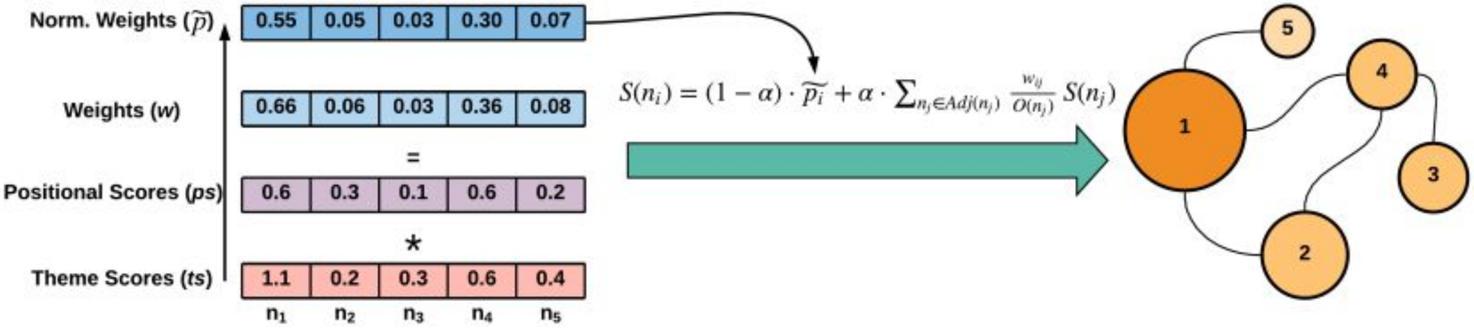
Papagiannopoulou, E., & Tsoumakas, G. (2018). Local word vectors guiding keyphrase extraction. *Information Processing & Management*, *54*(6), 888-902.

SIFRank



Sun, Y., Qiu, H., Zheng, Y., Wang, Z., & Zhang, C. (2020). SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, *8*, 10896-10906.

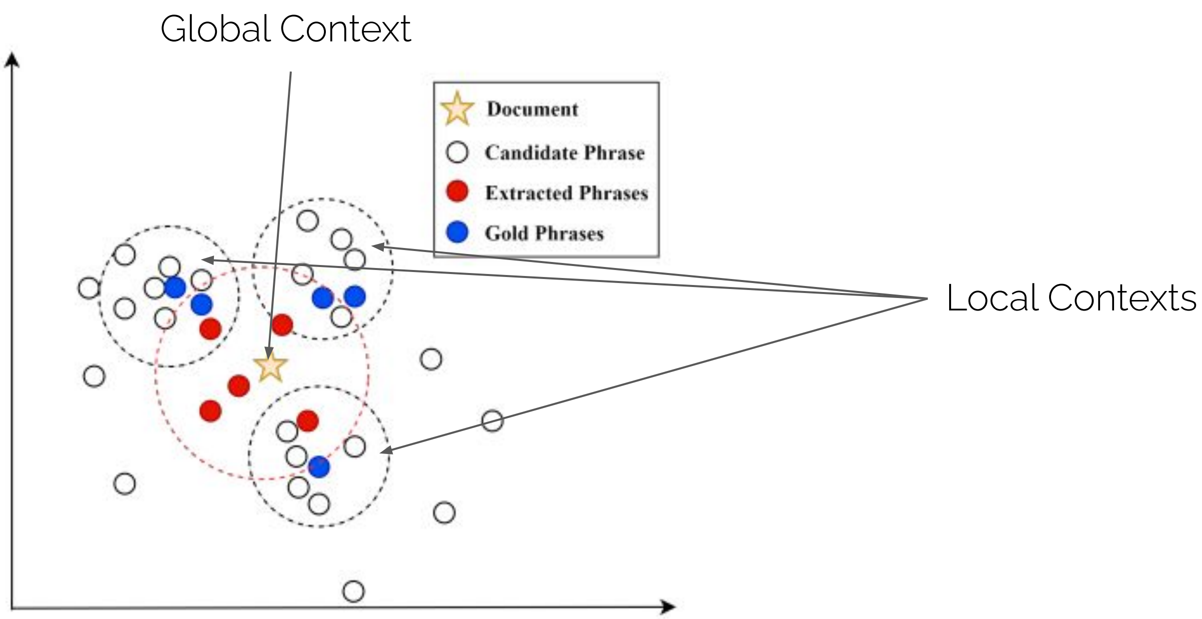
KPRank



- Motivated by Key2Vec
- Uses contextual word embeddings - SciBERT
- Also integrated positional information
- Ranks the phrases using biased PageRank

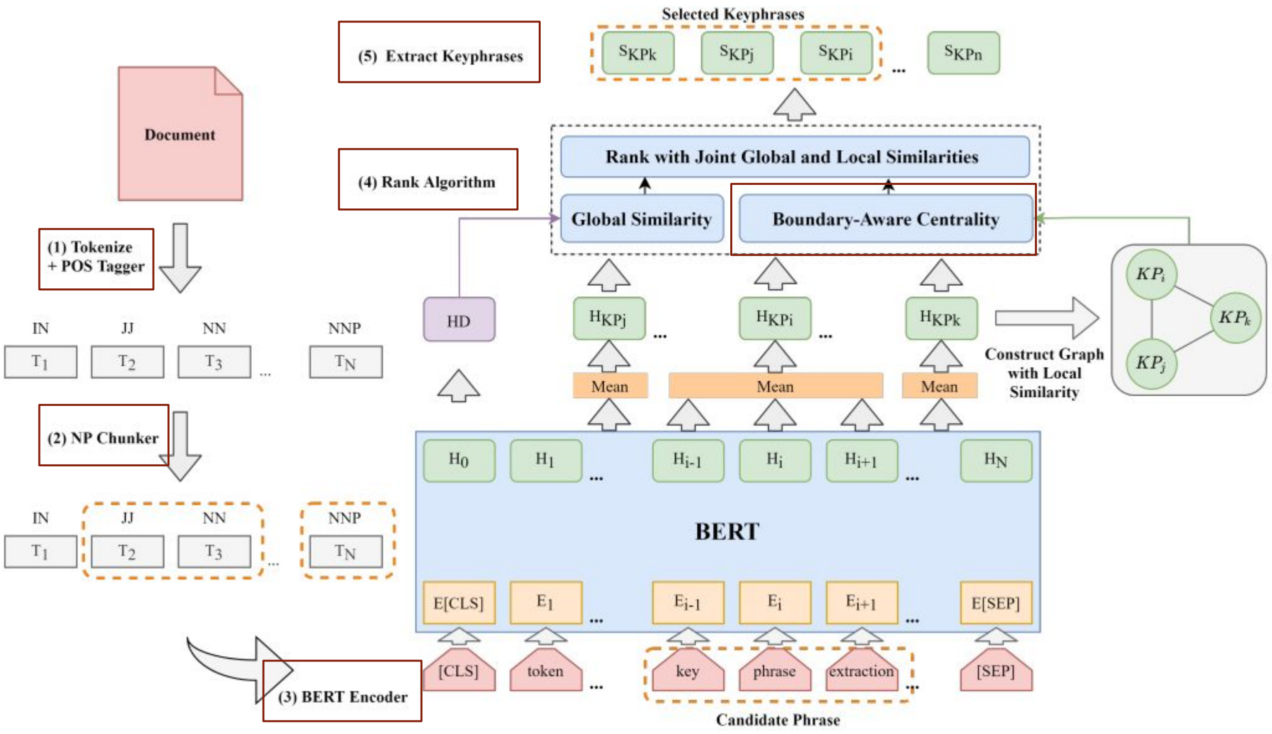
Patel, K., & Caragea, C. (2021, April). Exploiting Position and Contextual Word Embeddings for Keyphrase Extraction from Scientific Papers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1585-1591).

Jointly Modeling Local and Global Context



Liang, X., Wu, S., Li, M., & Li, Z. (2021). Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context. *arXiv preprint arXiv:2109.07293*.

Steps



Preprocess Input Text

Candidate Selection

Candidate and Document Representation - Global Context

Boundary-Aware Centrality - Local Context

Ranking Keyphrases Jointly with Local and Global Context

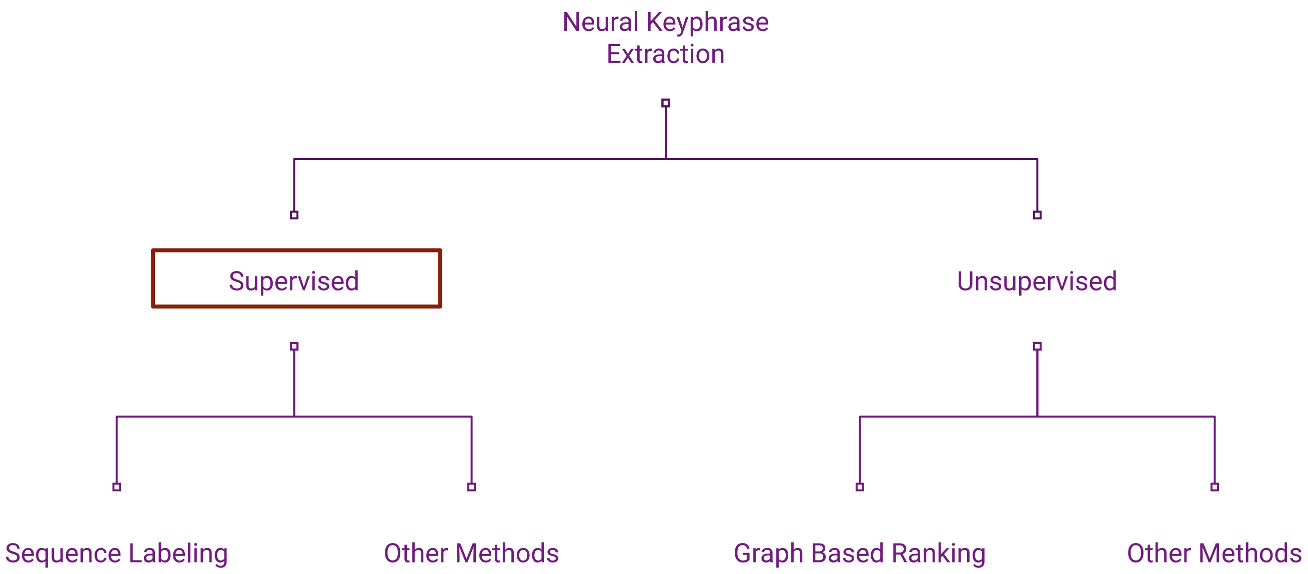
Liang, X., Wu, S., Li, M., & Li, Z. (2021). Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context. *arXiv preprint arXiv:2109.07293*.

Latest Paper at ECIR 2022

Venktesh, V., Mohania, M., & Goyal, V. (2022, April). Topic Aware Contextualized Embeddings for High Quality Phrase Extraction. In *European Conference on Information Retrieval* (pp. 457-471). Springer, Cham.

https://link.springer.com/chapter/10.1007/978-3-030-99736-6_31

Taxonomy of Extractive Methods



Keyphrase Extraction as Sequence Labeling

2016

- **CRF** (Gollapalli et al.)
- **Joint Layer RNN** (Zhang et al.)

2019

- **BLSTM-CRF** (Al-Zaidy et al.)
- **CRF with DOC features and embeddings** (Patel et al.)
- **BLING-KE** (Xiong et al.)

2020

- **BLSTM-CRF with Contextual Embeddings** (Sahrawat et al.)
- **DAKE** (Santosh et al.)
- **SaSAKE** (Santosh et al.)

2021

- **TNT-KID** (Martinc et al.)
- **Phraseformer** (Khasmakhi et al.)
- **Hypernetwork** (Duan et al.)

Fixed word embeddings

Transformer based contextual language models

Sequence Labeling

O O O O O O B-K I-K O O O O O

In this paper , we formulate keyphrase extraction from scholarly articles as a

B-K I-K I-K O O O B-K O O O O O O

sequence labeling task solved using a BiLSTM-CRF , where the words in the

O O O O O B-K I-K I-K

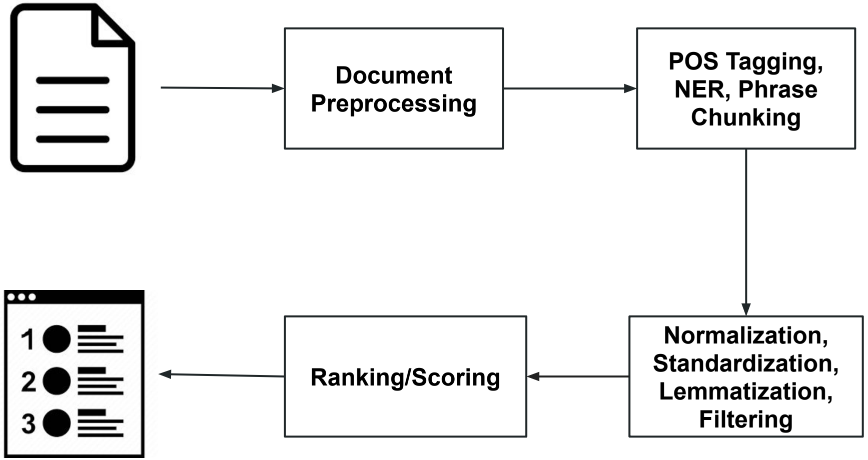
input text are represented using deep contextualized embeddings

Input - Sequence of words/sub-words with their embeddings

Output - Sequence of labels

1. **B-K** - Beginning of a keyphrase
2. **I-K** - Inside a keyphrase
3. **O** - Outside a keyphrase

Why Sequence Labeling

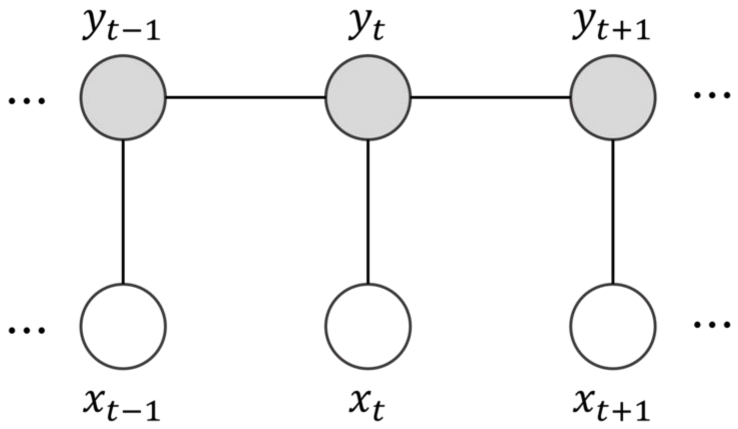


- ❖ Candidate Selection
    - Extracting named entities, noun phrases, POS Tags
    - Extracting n-grams, lexical patterns
    - Dependence on external gazetteers
    - Too many heuristics involved
- ❖ Hard to reproduce results
- ❖ Not a unified process

Sequence Labeling to The Rescue

- ❖ No heuristics
- ❖ Not much pre-processing
- ❖ Unified process
- ❖ Optimal assignment of keyphrases
- ❖ Leverages techniques of other sequence tagging tasks
- ❖ Captures long term dependencies

Sequence Labeling with CRF



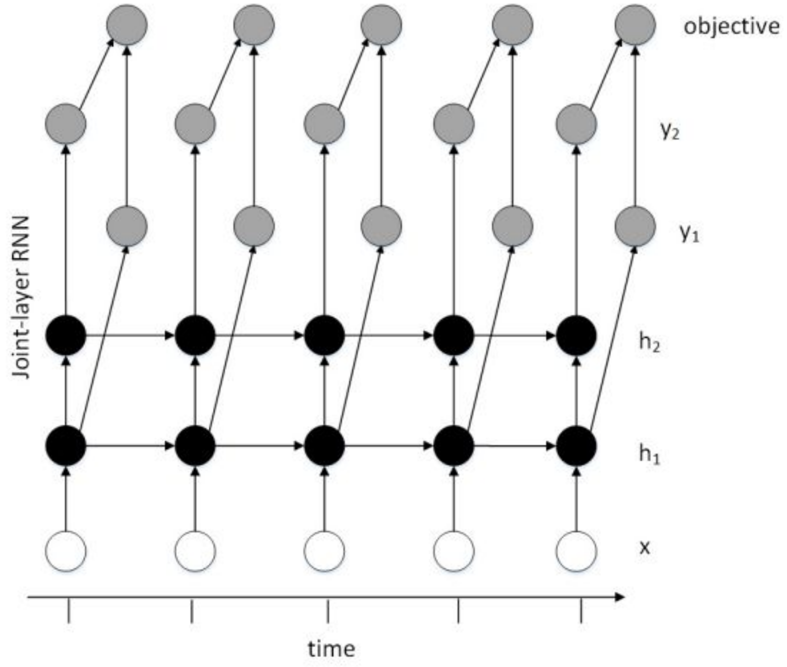
- Term, orthographic and stopword features
- Parse-tree features
- Title features

- How can we avoid pre-filtering of correct candidate phrases based on potentially erroneous POS tags during keyphrase extraction?
- Can we model the length of a keyphrase more naturally in our extraction methods?

Gollapalli, S. D., & Li, X. L. (2016). Keyphrase extraction using sequential labeling. *arXiv preprint arXiv:1608.00329*.

Image Source - https://commons.wikimedia.org/wiki/File:Graphical_model_for_CRF.PNG

Sequence Labeling using Joint Layer RNN



- ❖ Extension of a stacked RNN with two hidden layers
- ❖ At time t, the training input is the concatenation of features from a mixture within a window
- ❖ Two output layers are combined into a objective layer

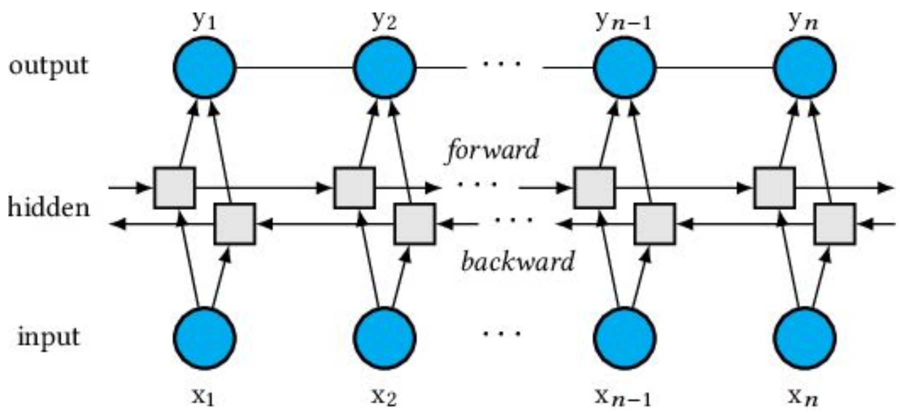
Input at every step - word embedding

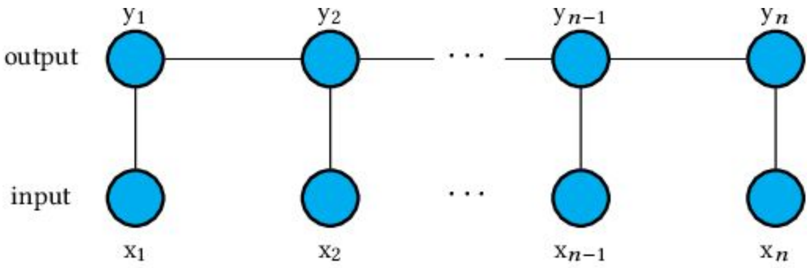
Output at every step -

1. Whether the current word is a keyword (**True/False**)
2. Current word
   a. **Single** - single keyword
   b. **Begin** - beginning of a keyphrase
   c. **Middle** - middle of a keyphrase
   d. **End** - end of a keyphrase
   e. **Not** - not part of a keyphrase

Zhang, Q., Wang, Y., Gong, Y., & Huang, X. J. (2016, November). Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 836-845).

Sequence Labeling using CRF and BLSTM-CRF with word embeddings as features





Word2Vec - Google News
Word2Vec - ACM Corpus

Alzaidy, R., Caragea, C., & Giles, C. L. (2019, May). Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In *The world wide web conference* (pp. 2551-2557).

Patel, K., & Caragea, C. (2019, September). Exploring word embeddings in crf-based keyphrase extraction from research papers. In *Proceedings of the 10th International Conference on Knowledge Capture* (pp. 37-44).

Few Questions and Answers

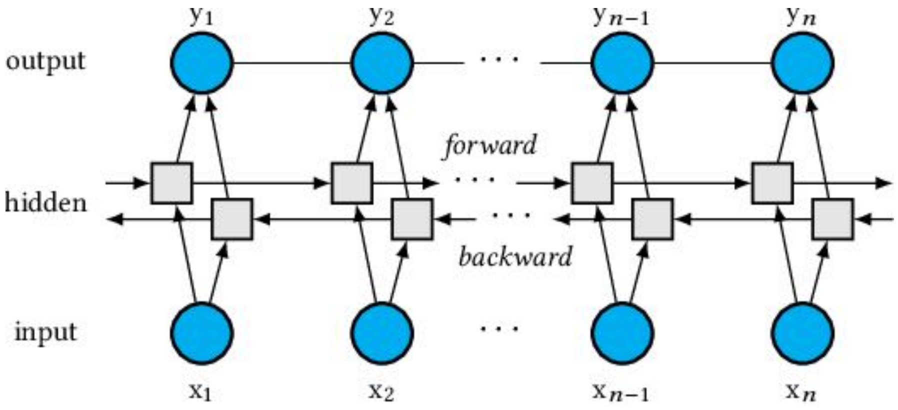
Q. *Why word embeddings with CRF?*

Ans: Only CRF with document level features do not capture the semantics of the words in context that are often hidden in text.

Q. *Why use BLSTM with CRF?*

Ans:

a. BLSTM - in order to capture long term dependencies
b. CRF - dependencies among the labels of neighboring words



Few Key Observations

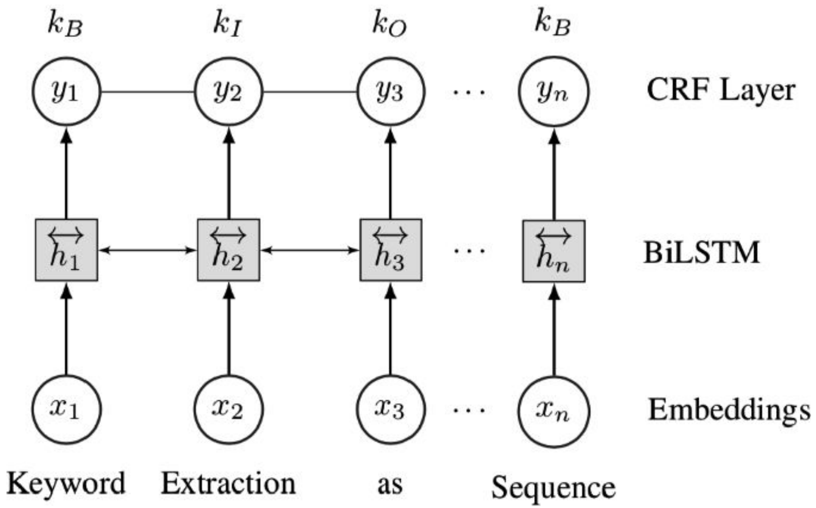
- CRF with linguistic and statistical features performed better than CRF with word embeddings

- CRF with word embeddings + linguistic and statistical features outperformed both CRF + linguistic and statistical features, and CRF + word embeddings

- BLSTM-CRF helps on a large training data

- BLSTM-CRF predicts long keyphrases well

Sequence Labeling with Contextual LMs



Fixed Embedding Models

- Word2Vec
- fastText
- GloVe

Contextual Embedding Models

- BERT
- SciBERT
- GPT
- GPT-2
- ELMO
- RoBERTa
- Transformer XL

Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., ... & Zimmermann, R. (2020, April). Keyphrase extraction as sequence labeling using contextualized embeddings. In *European Conference on Information Retrieval* (pp. 328-335). Springer, Cham.

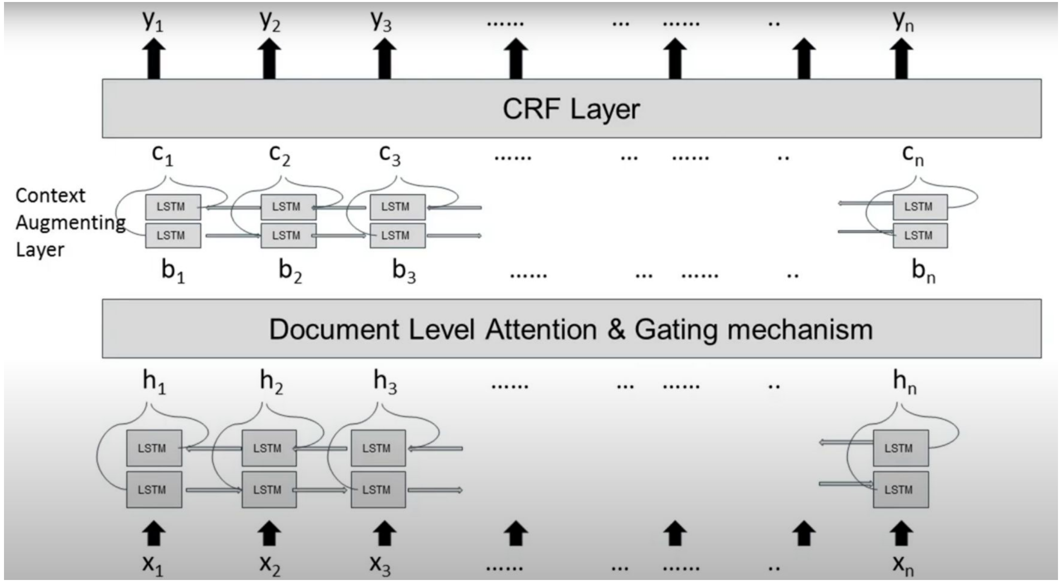
Sequence Labeling with Contextual LMs

F1 Scores

Embeddings	Inspec	SemEval 2010	SemEval 2017
SciBERT	0.593	**0.357**	0.521
BERT	0.591	0.330	**0.522**
ELMO	0.568	0.225	0.504
Transformer-XL	0.521	0.222	0.445
GPT	0.523	0.235	0.439
GPT-2	0.531	0.240	0.439
RoBERTa	**0.595**	0.278	0.508
Glove	0.457	0.111	0.345
Fasttext	0.524	0.225	0.426
Word2Vec	0.473	0.208	0.292

Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., ... & Zimmermann, R. (2020, April). Keyphrase extraction as sequence labeling using contextualized embeddings. In *European Conference on Information Retrieval* (pp. 328-335). Springer, Cham.

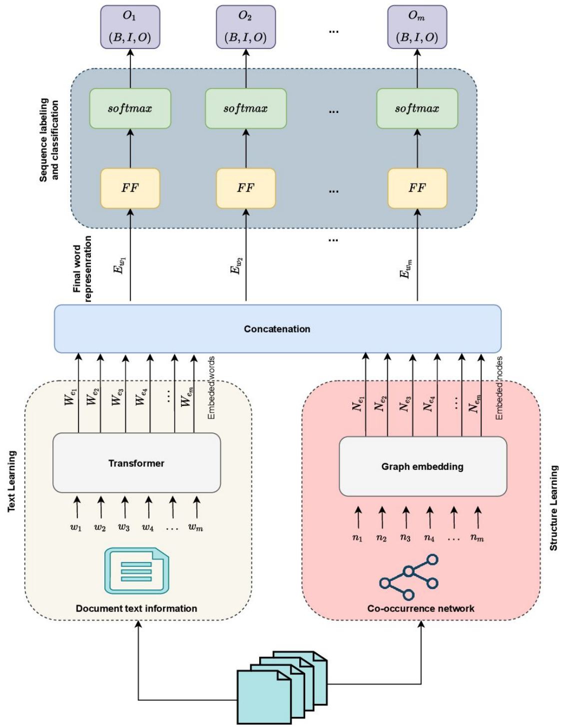
DAKE - Document Level Attention for Keyphrase Extraction



- Sequence labeling with CRF
- Local context from the sentence using BLSTM
- Document level attention for incorporating relevant information from the supporting information with respect to the local context
- Gating mechanism to filter out the irrelevant information

Santosh, Tokala Yaswanth Sri Sai, et al. "DAKE: document-level attention for keyphrase extraction." *European Conference on Information Retrieval*. Springer, Cham, 2020.

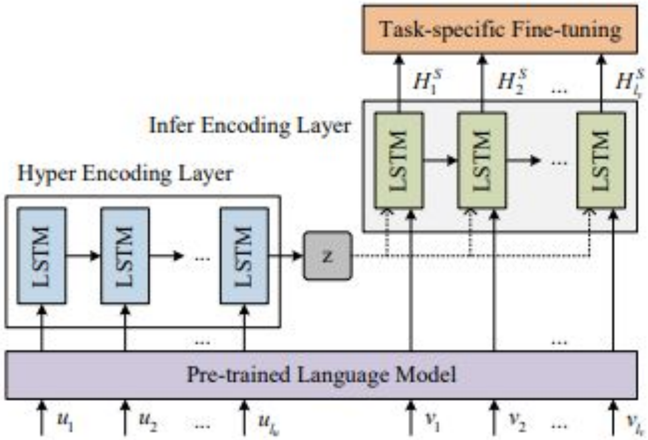
Phraseformer



- ❖ Multi-modal keyphrase extraction model
- ❖ Word representations
  - ➢ BERT
  - ➢ Graph embeddings learnt from word co-occurrence graph
  - ➢ Final embedding obtained by concatenating the BERT embeddings with graph embeddings
- ❖ Keyphrase extraction as a sequence labeling task

Nikzad-Khasmakhi, N., Feizi-Derakhshi, M. R., Asgari-Chenaghlu, M., Balafar, M. A., Feizi-Derakhshi, A. R., Rahkar-Farshi, T., ... & Ranjbar-Khadivi, M. (2021). Phraseformer: Multimodal key-phrase extraction using transformer and graph embedding. *arXiv preprint arXiv:2106.04939*.

Sequence Labeling with Hypernetworks

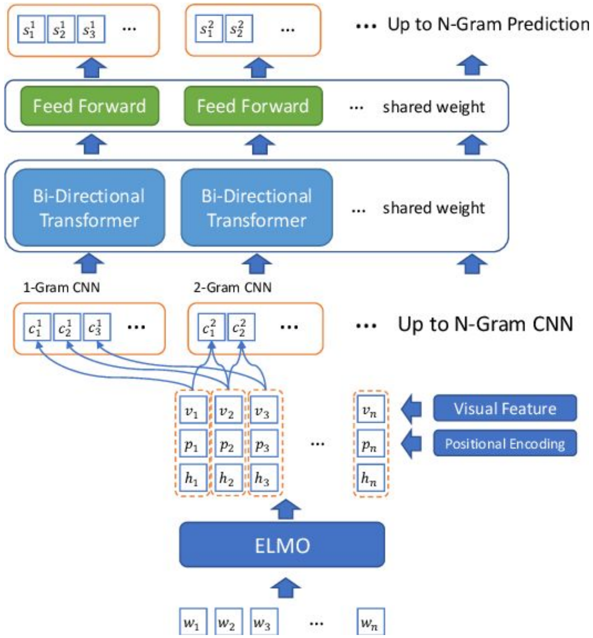


- ❖ Current models only exploit the plain text features
- ❖ Treats input text equally and ignore the inherent context such as *title*, *sections*, *main body* or semantic roles - *who*, *what* and *how*
- ❖ Models pay over-weighted attention to insignificant words

- ❖ Models descriptive meta-information via hypernetworks
- ❖ Extracts the descriptive meta-information in the meta-text for more effective processing of the main text
  - ➢ *Title* - neta-text
  - ➢ *Body* - main text
- ❖ Meta and main text are first encoded by a pre-trained language model
- ❖ The embedding of the meta-text is then fed into a hyper encoding layer to generate the weights of the infer encoding layer
- ❖ Infer encoding layer converts the embedding of the main text into representations for task-specific finetuning

Duan, W., He, X., Zhou, Z., Rao, H., & Thiele, L. (2021). Injecting descriptive meta-information into pre-trained language models with hypernetworks.

BLING-KPE - Beyond Language Understanding Keyphrase Extraction



- ❖ Keyphrase extraction in the wild - Open Domain
- ❖ Convolutional transformer models the language properties
  - ➢ N-grams and their interaction
- ❖ Uses visual presentations of text pieces integrated with word embeddings
  - ➢ Location
  - ➢ Size
  - ➢ Font
  - ➢ HTML structure
- ❖ Weak supervision from search queries
  - ➢ Query prediction as a pre-training task
- ❖ Zero-shot evaluation on DUC-2001 outperforms model trained on scientific domain

Data

- ❖ https://github.com/microsoft/OpenKP
- ❖ https://huggingface.co/datasets/midas/openkp

Xiong, L., Hu, C., Xiong, C., Campos, D., & Overwijk, A. (2019). Open domain web keyphrase extraction beyond language modeling. *arXiv preprint arXiv:1911.02671*.

Interesting Works using Sequence Labeling

- **SaSAKE** - Santosh, T., Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2020, December). Sasake: syntax and semantics aware keyphrase extraction from research papers. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 5372-5383).

- **TNT-KID** - Martinc, M., Škrlj, B., & Pollak, S. (2020). TNT-KID: Transformer-based neural tagger for keyword identification. Natural Language Engineering, 1-40.

B I O Tagged Data

Dataset Preview   Go to dataset viewer

Subset: extraction

Split: train

id (int)	document (json)	doc_bio_tags (json)
1,001	["A", "conflict", "between", "language", "and", "atomistic", "information", "Fred",…	["O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O",…
1,002	["Selective", "representing", "and", "world-making", "We", "discuss", "the", "thesis",…	["B", "I", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O",…
1,000	["Does", "classicism", "explain", "universality", "?", "Arguments", "against",…	["O", "B", "O", "B", "O", "O", "O", "O", "O", "B", "I", "I", "I", "O", "O", "O", "O",…
100	["Separate", "accounts", "go", "mainstream", "-LSB-", "investment", "-RSB-", "New",…	["O", "O", "O", "O", "O", "B", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O",…
1,012	["Evolving", "receptive-field", "controllers", "for", "mobile", "robots",…	["O", "O", "O", "O", "B", "I", "O", "O", "O", "B", "I", "O", "O", "O", "O", "O", "O",…
1,016	["A", "scalable", "model", "of", "cerebellar", "adaptive", "timing", "and",…	["O", "B", "I", "O", "B", "I", "I", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O",…
1,046	["A", "suggestion", "of", "fractional-order", "controller", "for", "flexible", "spacecraft"…	["O", "O", "O", "O", "O", "O", "B", "I", "I", "I", "O", "O", "O", "O", "O", "B", "I",…

Datasets

- Inspec
- KP20K
- WWW
- SemEval 2010
- SemEval 2017
- KDD
- CSTR
- NUS
- PubMed
- ACM
- CiteULike
- KPTimes
- DUC-2001
- KPCrowd
- OpenKP
- LDKP3K
- LDKP10K

Outline of Part II

Part I - Neural Keyphrase Extraction

Part II - Neural Keyphrase Generation (Rui)

Part III - Hands-on Practice with OpenNMT-kpg and DLKP

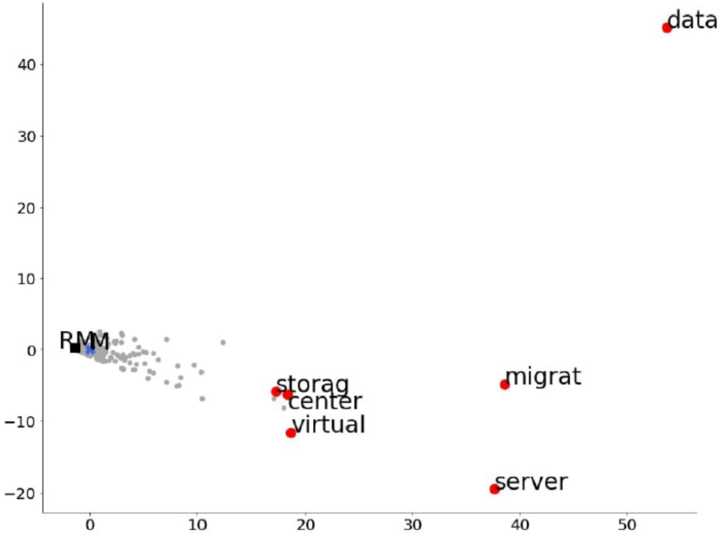
Break

Taxonomy of Extractive Methods

Supervised		Unsupervised/ Distantly supervised
	Embedding	Key2Vec, EmbedRank, LV. (Mahata et al, 2018, Bennani-Smires et al, 2017) SIFRank, UKER (Sun et al, 2020, Liang et al, 2021) KeyGames (Saxena et al, 2020)
JL-RNN (Zhang et al, 2016) BiLSTM-CRF, BiLSTM-CRF-BERT, TNT-KID etc. (Alzaidy et al, 2019, Sahrawat et al. 2019, Martinc et al. 2021)	Sequence Labeling	UCPhrase (Gu et al, 2021)
BLING-KPE, JointKPE (Xiong et al, 2019, Sun et al, 2021)	Ranking	
PtrNet, DivGraphPointer (Sun et al, 2019)	Pointer	

Embedding-based KPE

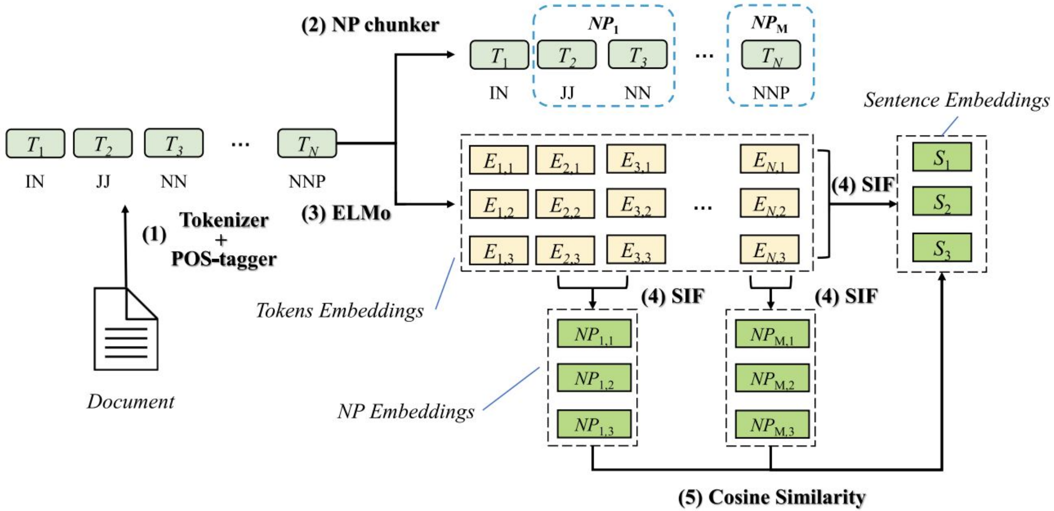
- Local vectors (LV)
  - Fine-tune GloVe vectors with local text (words in a target doc)
    - Keywords lie far from the mean of all words in local vector space
    - Main bulk of the words that determine the document embedding (mean pooling) are not important



(Papagiannopoulou et al. 2020). Keywords lie far from the mean of all words in local vector space. arXiv

Embedding-based KPE (w/ Pretrained LM)

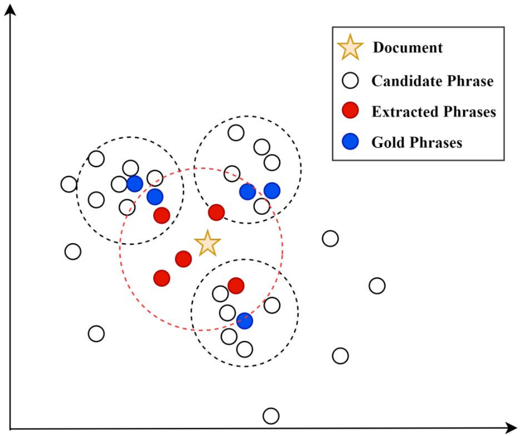
- SIFRank
  - Represent documents with Pretrained Language Models (ELMo)
  - Rank noun phrases by their semantic similarity to document embedding

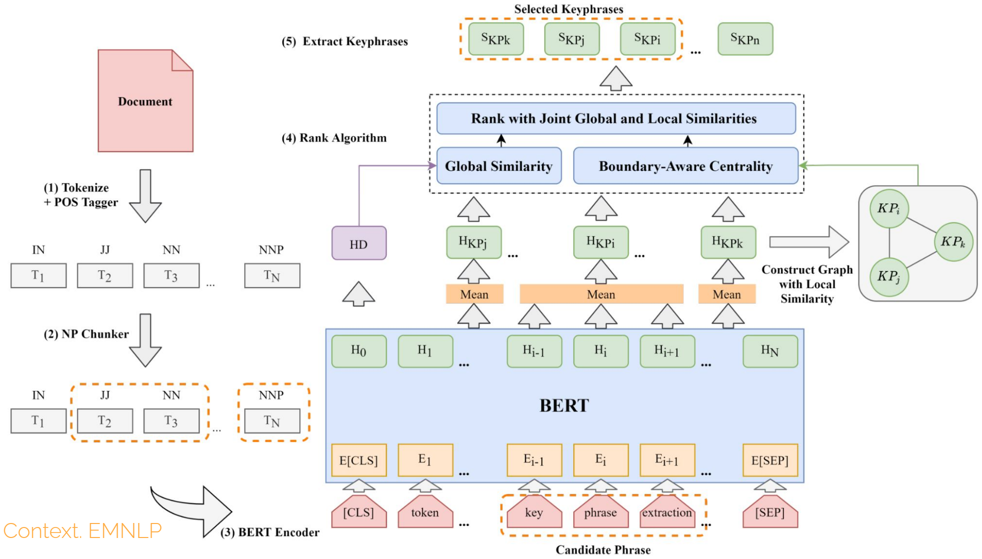


(Sun et al. 2020). SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. IEEE.

Embedding-based KPE (w/ Pretrained LM)

- UKER
  - Score each candidate phrase by global and local context
    - Global relevance: phrase-document similarity using BERT
    - Local salience
      - Measured by the degree of nodes (candidate phrases) in the local graph





(Liang et al. 2021). Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context. EMNLP

Key2Vec

ACL Anthology

Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings

Debanjan Mahata, John Kuriakose, Rajiv Ratn Shah, Roger Zimmermann

Abstract

Keyphrase extraction is a fundamental task in natural language processing that facilitates mapping of documents to a set of representative phrases. In this paper, we present an unsupervised technique (Key2Vec) that leverages phrase embeddings for ranking keyphrases extracted from scientific articles. Specifically, we propose an effective way of processing text documents for training multi-word phrase embeddings that are used for thematic representation of scientific articles and ranking of keyphrases extracted from them using theme-weighted PageRank. Evaluations are performed on benchmark datasets producing state-of-the-art results.

PDF

Cite

Search

Note

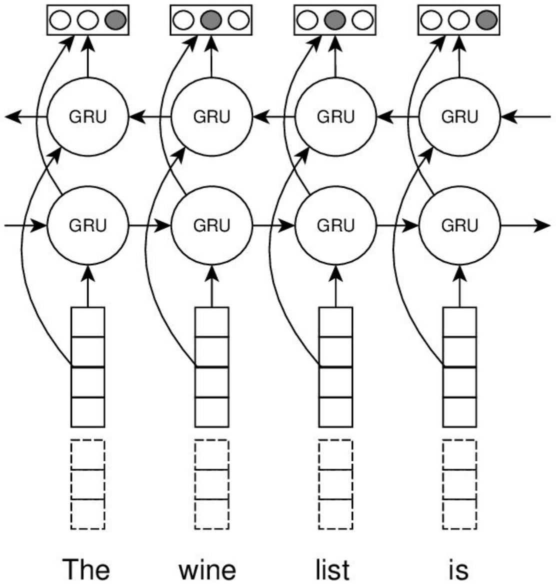
Mahata, D., Kuriakose, J., Shah, R., & Zimmermann, R. (2018, June). Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 634-639).

Taxonomy of Extractive Methods

Supervised		Unsupervised/ Distantly supervised
	Embedding	Key2Vec, EmbedRank, LV. (Mahata et al, 2018, Bennani-Smires et al, 2017); SIFRank, UKER (Sun et al, 2020, Liang et al, 2021); KeyGames (Saxena et al, 2020)
JL-RNN (Zhang et al, 2016); BiLSTM-CRF, BiLSTM-CRF-BERT, TNT-KID etc. (Alzaidy et al, 2019, Sahrawat et al. 2019, Martinc et al. 2021)	Sequence Labeling	UCPhrase (Gu et al, 2021)
BLING-KPE, JointKPE (Xiong et al, 2019, Sun et al, 2021)	Ranking	
PtrNet, DivGraphPointer (Sun et al, 2019)	Pointer	

Sequence Labeling based KPE

- Predict B/I/O tags for each input token
    -



- (Zhang et al, 2016),Keyphrase extraction using deep recurrent neural networks on Twitter. EMNLP.
- (Alzaidy et al., 2019). "Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. WWW.
- (Sahrawat, et al. 2019), Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings." *arXiv*
- (Martinc et al. 2020). "TNT-KID: Transformer-based neural tagger for keyword identification." *Natural Language Engineering*.

Unsupervised Phrase Mining

- UCPhrase (Gu et al, 2021)
  - Silver labels
    - Word spans that appear more than once in a document
  - Context-aware Lightweight Classifier
    - Use attention map of BERT as features
    - A light CNN-based sequence labelling to predict BIO tags

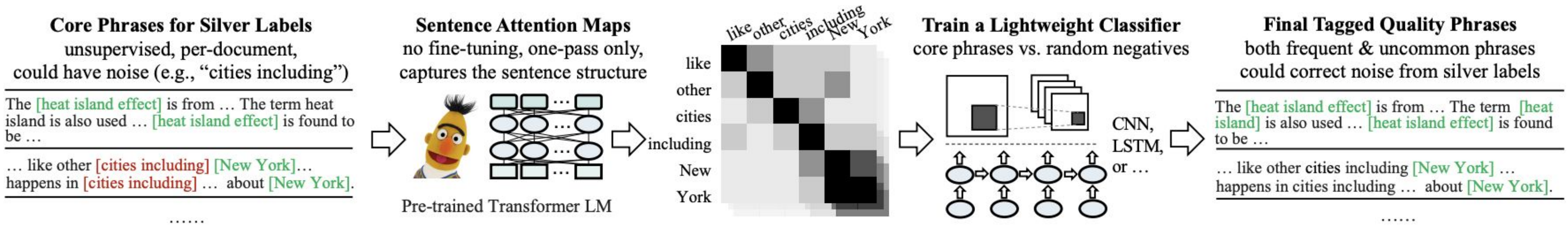


Figure 1: An overview of our UCPhrase: unsupervised context-aware quality phrase tagging.

(Gu et al. 2021) UCPhrase: Unsupervised Context-aware Quality Phrase Tagging. KDD

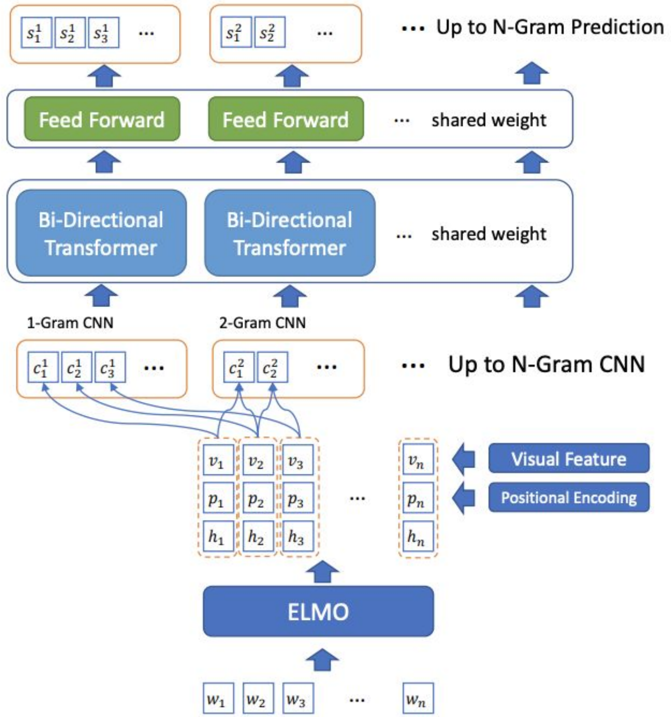
Taxonomy of Extractive Methods

Supervised		Unsupervised/ Distantly supervised
	Embedding	Key2Vec, EmbedRank, LV. (Mahata et al, 2018, Bennani-Smires et al, 2017) SIFRank, UKER (Sun et al, 2020, Liang et al, 2021) KeyGames (Saxena et al, 2020)
JL-RNN (Zhang et al, 2016) BiLSTM-CRF, BiLSTM-CRF-BERT, TNT-KID etc. (Alzaidy et al, 2019, Sahrawat et al. 2019, Martinc et al. 2021)	Sequence Labeling	UCPhrase (Gu et al, 2021)
BLING-KPE, JointKPE (Xiong et al, 2019, Sun et al, 2021)	Ranking	
PtrNet, DivGraphPointer (Sun et al, 2019)	Pointer	

KPE by Learning to Rank

- Scoring and Ranking all n-grams in text
    -
- Related work
    - BLING-KPE, Xiong et al. 2019
        - Incorporated vision features in KPE
        - Contributed OpenKP
    - JointKPE, Si et al. 2021
        - Maximizing global informativeness score between phrase p_k and document D

          $f_{info}(p_k, D)$
        - Utilized pretrained language models



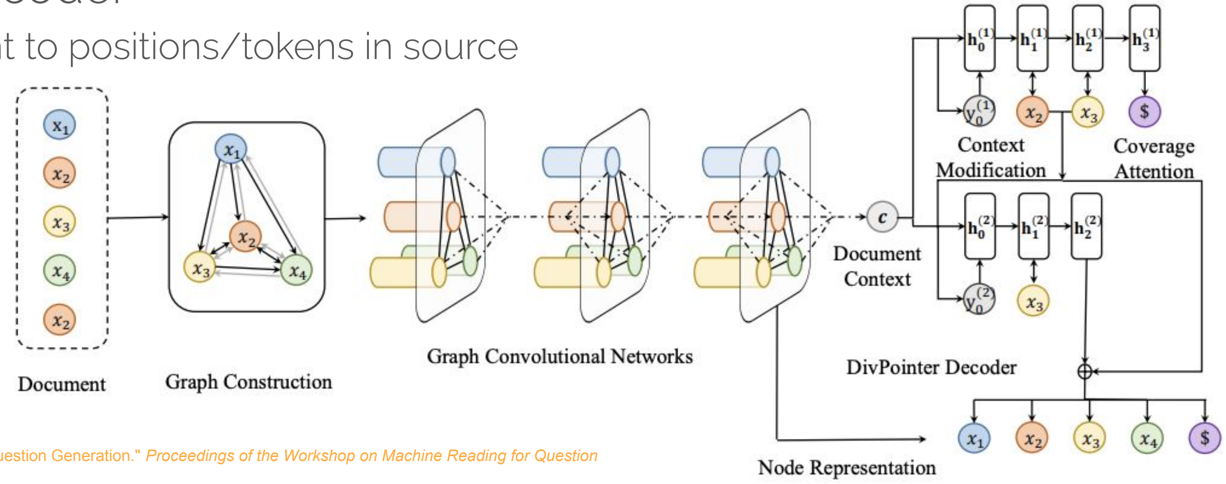
- Xiong, Lee, et al. "Open Domain Web Keyphrase Extraction Beyond Language Modeling." *EMNLP-IJCNLP*. 2019.
- Sun, Si, et al. "Capturing global informativeness in open domain keyphrase extraction." *CCF NLPCC*. Springer, Cham, 2021.

Pointer-based KPE

- Representing target keyphrases
  - as a sequence of start/end positions in text

$$Y = \{(p_1^{start}, p_1^{end}), (p_2^{start}, p_2^{end}), \ldots, (p_T^{start}, p_T^{end})\}$$

  - or as a sequence of tokens appearing in text
- Pointer Network as decoder
  - Utilize attention to point to positions/tokens in source



- Subramanian, Sandeep, et al. "Neural Models for Key Phrase Extraction and Question Generation." *Proceedings of the Workshop on Machine Reading for Question Answering*. 2018.
- Sun, Zhiqing, et al. "Divgraphpointer: A graph pointer network for extracting diverse keyphrases." *SIGIR* 2019.