中国科学技术大学
University of Science and Technology of China

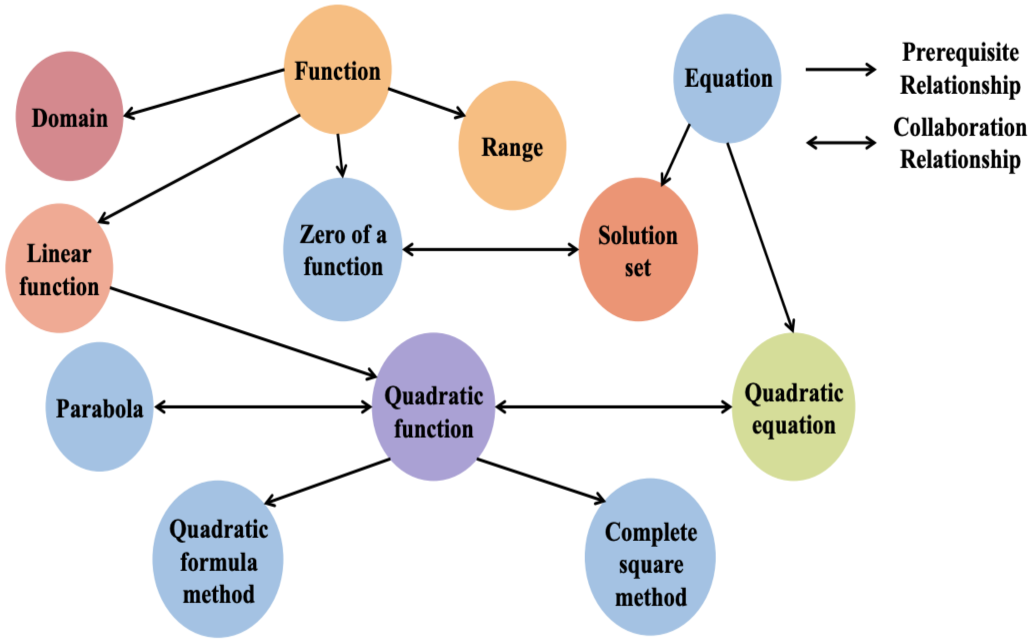
Guided Attention Network for Concept Extraction

IJCAI 2021

Songtao Fang[1] , **Zhenya Huang**[1*] , **Ming He**[2,3] , **Shiwei Tong**[1]
Xiaoqing Huang[1] , **Ye Liu**[1] , **Jie Huang**[1] , **Qi Liu**[1]

[1]Anhui Province Key Laboratory of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China
[2]Department of Electronic Engineering, Shanghai Jiao Tong University
[3]Didi Chuxing, Beijing, China;
songtao@mail.ustc.edu.cn, huangzhy@ustc.edu.cn, heming01@foxmail.com
{tongsw, xqhuang, liuyer, jiehuang}@mail.ustc.edu.cn, qiliuql@ustc.edu.cn

Concept Extraction



- **Definition**
  - Concept extraction aims at extracting words or phrases describing a concept (e.g., logistic regression, hash function,and infinite set) from a given corpus (e.g., research papers and textbooks)

- **Application**
  - Constructing knowledge bases
  - Transforming unstructured text into structured information
  - Producing meaningful representation of texts

Rule-based methods

- Pattern e.g. is called, define as
- Part of speech rules: noun phrase

The rule-based methods can be seen as hard matching methods and lacks generalization ability.

Statistics-based method

- Through statistical or part-of-speech information, phrases are extracted as candidate concept
- Sort candidate phrases based on evaluation indicators

Commonly used evaluation indicators：TF-IDF、C-value、NC-value、PMI

Based on statistical information, it is difficult to find low-frequency phrases when the distribution of words has a long-tailed distribution.

Deep learning-based methods

- Bi-LSTM+CRF
- Bert+Finetune
- Encoder-Decoder framework+domain knowledge

Deep learning-based methods：require many labeled data in a new domain.

Motivation

Title:Properties of Sets
Main Body: A **set** is a well-defined collection of items. Each item is called an **element**. A **set** is usually named with a capital letter and may be defined in three ways. ... whose elements cannot be counted or listed is called an **infinite set**. If all of the elements can be counted or listed, the **set** is called a **finite set**.
concepts: set, element, infinite set, finite set

Table 1: An example of concept extraction from textbook. Red words are clue words, bold words are concepts.

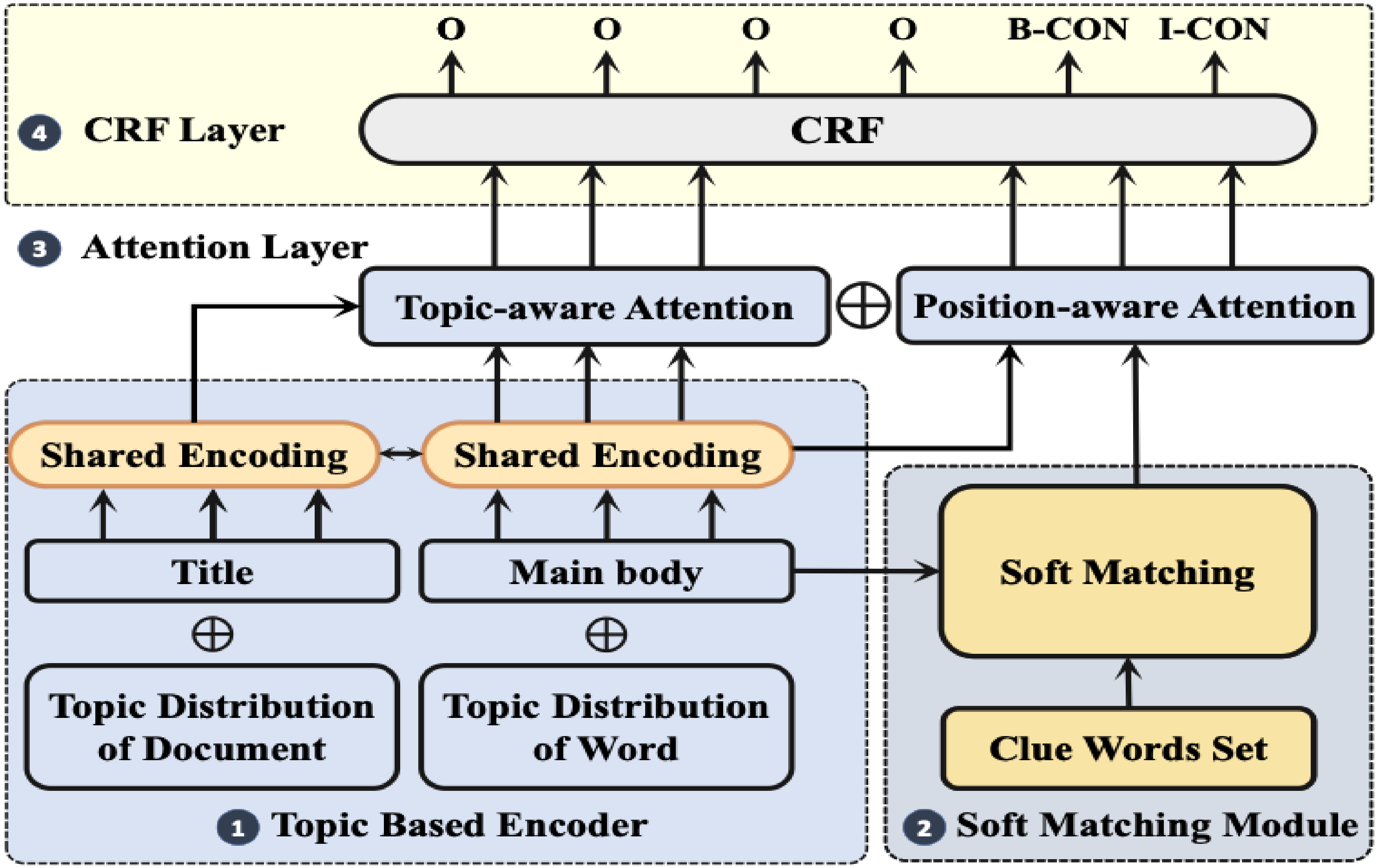
- Title or topic as global information play the leading role in concept extraction.
- the clue words "is called an",which suggests there should be a concept following the word "an".

Challenge

- How to combine the global information to pay attention to the topic-related words?
- Clue words have large numbers of linguistic variants, which is difficult to collect complete clue words.
- How to model the relationship between clue words and concept?

Architecture of GACEN

- Topic-based encoder
- Soft Matching module
- Attention Layer
- CRF Layer



B-CON: Beginning of a concept
I-CON: Middle part of the concept
E-CON: End of the concept
O: Out of a concept

Topic-based encoder & Topic-aware Attention module

We consider the title and main body of the document separately.

- Topic Distribution & Word Embedding

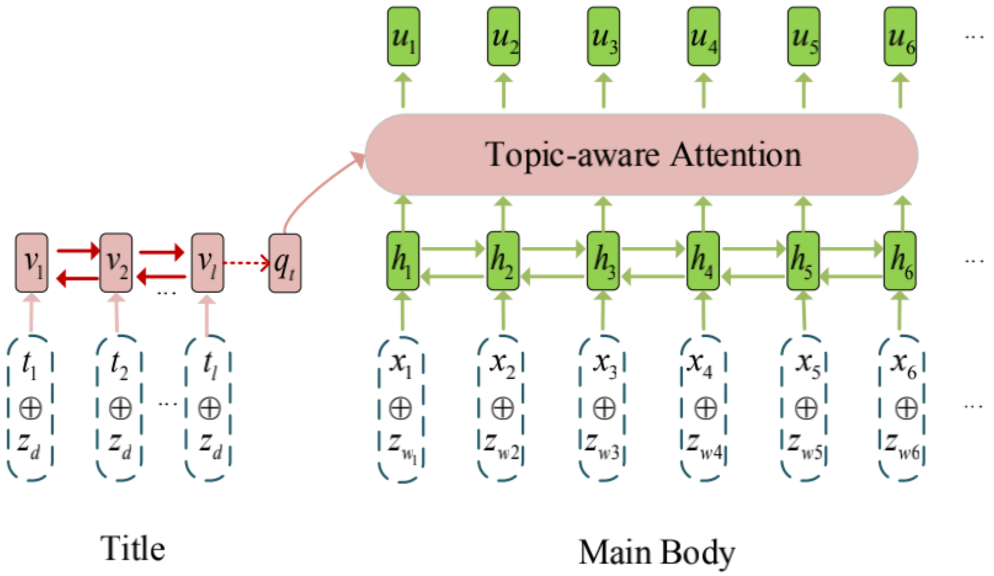
  :Topic Distribution of Document
  :Topic Distribution of word

- Shared Encoder & Topic-aware Attention

$$\overrightarrow{\mathbf{h}}_i = \mathrm{LSTM}\left((\mathbf{e}_{\mathbf{x}_i} \oplus \mathbf{z}_{\mathbf{w}_i}), \overrightarrow{\mathbf{h}}_{i-1}\right)$$

$$\overleftarrow{\mathbf{h}}_i = \mathrm{LSTM}\left((\mathbf{e}_{\mathbf{x}_i} \oplus \mathbf{z}_{\mathbf{w}_i}), \overleftarrow{\mathbf{h}}_{i+1}\right)$$

$$\mathbf{h}_i = \left[\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i\right]$$



topic-aware token representations u_i

$$\mathbf{u}_i = \alpha_i \mathbf{h}_i,$$

$$\alpha_i = \mathrm{Soft\,Max}\left(\boldsymbol{v}_1^\top \tanh\left(W_1 \mathbf{h}_i + W_2 \mathbf{q}_t\right)\right),$$

Soft Matching Module

The soft matching module is used to match the corresponding clue words for the unseen sentences and locate where the clue words appear.

- Similarity scores

$$S_{ij}(X,q) = \mathrm{Score}\left(z_{c_{ij}}, z_q\right) = \cos\left(z_{c_{ij}} D, z_q D\right)$$

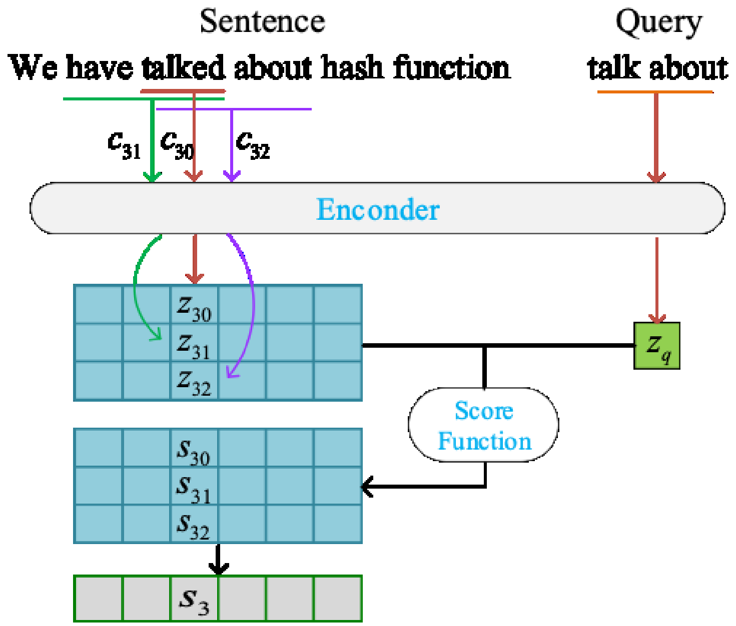
$$f_s(X,q) = S(X,q)v$$

S_ j indicates how likely a semantically similar phrase of query occurs at position j.

- Loss Function

$$L_{find} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|l_i|}(l_i \log f_s(X_i,q_i) + (1-l_i)\log(1-f_s(X_i,q_i))).$$

①

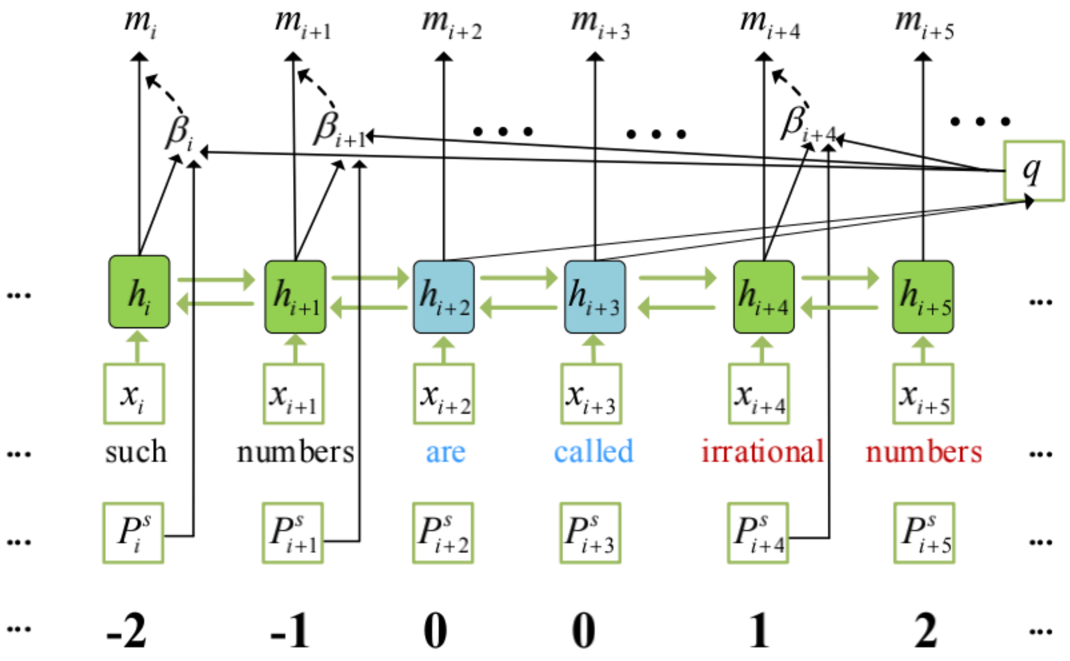


Position-aware Attention module

- Define a position sequence relative to the clue words:

$$p_i = \begin{cases} i - s_1, & i < s_1 \\ 0, & s_1 \leq i \leq s_2 \\ i - s_2, & i > s_2 \end{cases} \quad ①$$

[redacted] are the starting and ending indices of the clue words, respectively.



- Attention Scores

$$\mathbf{m}_i = \beta_i \mathbf{h}_i,$$

$$\beta_i = \mathrm{Soft\,Max}(\mathbf{v}_2^\top \tanh(\mathbf{W}_3\mathbf{h}_i + \mathbf{W}_4\mathbf{q} + \mathbf{W}_5\mathbf{p}_i^s)),$$

Representation-enhanced Sequence Tagging

$$\mathbf{h}_i' = \lambda \mathbf{u}_i + (1 - \lambda)\mathbf{m}_i, \quad ②$$

Finally, we concatenate the original token representation [redacted] with [redacted] as the input to CRF tagger

Datasets

We use three datasets to evaluate our model. The statistics of the datasets are as follows:

Datasets	#titles	#tokens	#labeled	#clue words
CSEN	690	1,242,156	4,096	36
KP-20K	20,000	4,040,212	50,768	95
MTB	284	691,534	1,092	24

Table 2: Dataset Statistics

CSEN: this dataset contains 690 video captions in Massive Open Online Courses (MOOCs) for Computer Science courses.

KP-20K: KP20K consists of 567,830 high-quality scientific publications from various computer science domains. We randomly select 20,000 articles from KP20K to form the KP-20K. We have collected concept phrases related to the computer field and automatically annotated the concept phrases in each article.

MTB: this dataset consists of mathematics textbooks for elementary, middle, and high schools.

Overall Performance & Ablation Study

	CSEN			KP-20K			MTB		
Method	Pr%	Re%	F1%	Pr%	Re%	F1%	Pr%	Re%	F1%
TextRank	23.46	27.82	25.45	15.29	23.01	18.37	24.78	30.65	27.40
TPR	31.46	29.21	30.29	14.83	25.12	18.65	25.19	32.75	28.48
Positionrank	31.80	30.37	31.07	18.92	25.47	21.71	28.37	39.04	32.86
CopyRNN	28.12	41.08	33.39	27.71	36.79	31.61	37.46	39.12	38.27
Joint-layer RNN	61.31	46.23	52.71	57.83	31.85	41.08	60.37	55.71	59.86
BERT-CRF	58.73	52.17	55.26	54.19	33.93	41.73	63.80	56.98	60.20
GACEN-topic	68.21	57.94	62.66	57.67	34.90	43.48	65.83	62.63	64.19
GACEN-position	64.13	**61.08**	62.57	52.78	**37.74**	44.01	60.55	64.71	62.56
GACEN*REs	**70.12**	57.12	62.96	**59.23**	34.71	43.77	**66.97**	62.98	64.91
GACEN	69.70	60.21	**64.60**	58.10	37.65	**45.69**	66.43	**64.72**	**65.56**

GACEN-topic, GACEN-position represent removing topic-aware attention module, position-aware attention module in GACEN, respectively. The GACEN*REs is a model, which replaces the soft matching module in GACEN with Regular Expression matching.

Learning Efficiency & Case Study

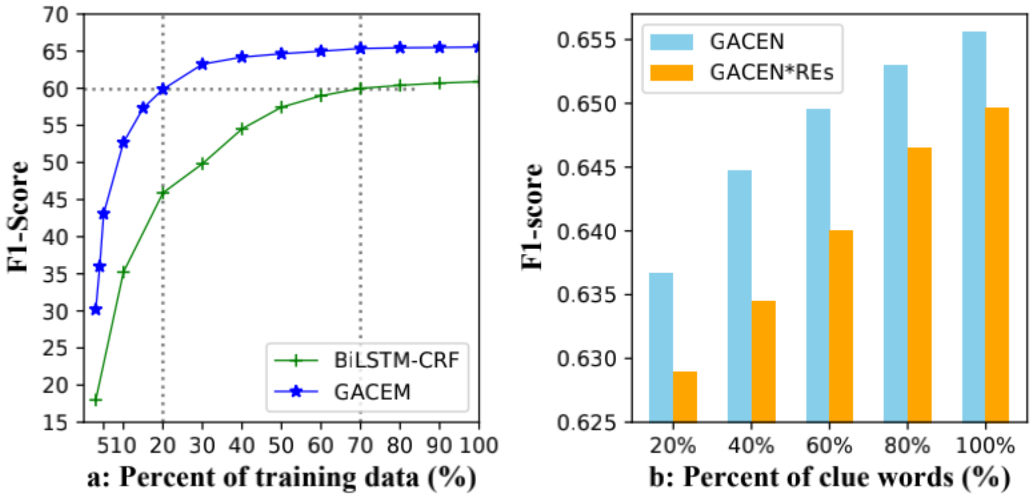


Figure 1:The experimental results on MTB

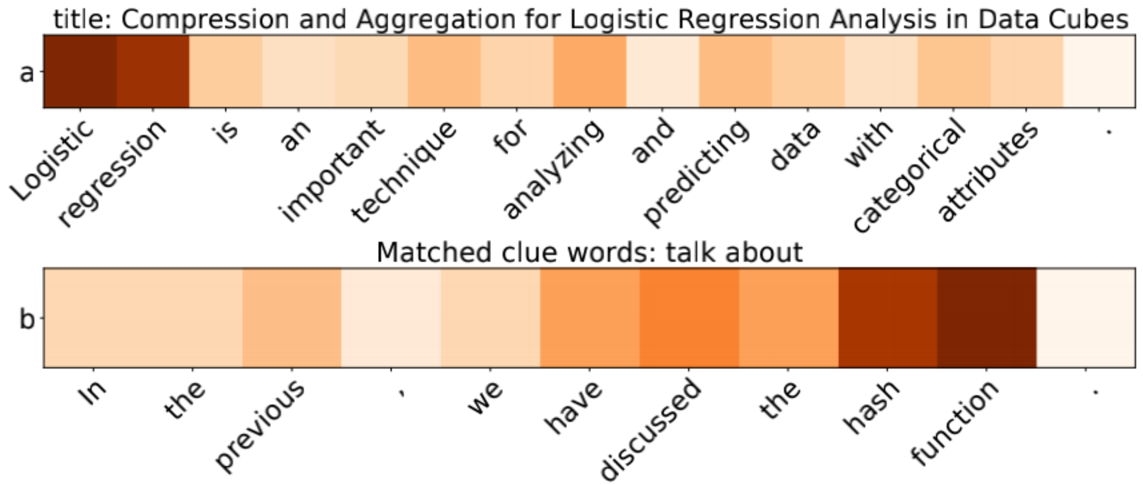


Figure 2: Two case studies of attention during inference.

Conclusion

In this paper, we proposed a novel model GACEN for the concept extraction task, which explicitly considered the structured information in the raw textual data.

- In GACEN, to incorporate the topic information into the feature representation, we first design a shared topic-based encoder to model the title and main body of the document with topic vectors at the document- and word-level separately.
- Then, To solve the problem of variants of clue words and improve the generalization ability, we pre-train a soft matching module with neural networks to capture semantically similar words.
- Finally, we design two attention modules, one of them is to gather the relevant global topic information for each context word according to the semantic relatedness based on topic enhanced representation, and the other aims to model the complex implicit relationship between clue words and concept with the semantic and position information of clue words.