STAC: Science Toolkit Based on Chinese Idiom Knowledge Graph

Meiling Wang[1], **Min Xiao**[2], **Changliang Li**[1*], **Yu Guo**[1], **Zhixin Zhao**[1] **and Xiaonan Liu**[1]
[1]AI Lab, KingSoft Corp, Beijing, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
{wangmeiling1, lichangliang, guoyu9}@kingsoft.com
{zhaozhixin, liuxiaonan1}@kingsoft.com
xiaomincloud@gmail.com

Abstract

Chinese idioms (Cheng Yu) have seen five thousand years' history and culture of China, meanwhile they contain large number of scientific achievement of ancient China. However, existing Chinese online idiom dictionaries have limited function for scientific exploration. In this paper, we first construct a Chinese idiom knowledge graph by extracting domains and dynasties and associating them with idioms, and based on the idiom knowledge graph, we propose a **S**cience **T**oolkit for **A**ncient **C**hina (STAC) aiming to support scientific exploration. In the STAC toolkit, idiom navigator helps users explore overall scientific progress from idiom perspective with visualization tools, and idiom card and idiom QA shorten action path and avoid thinking being interrupted while users are reading and writing. The current STAC toolkit is deployed at http://120.92.208.22:7476/demo/#/stac.

1 Introduction

Large scientific wealth has been accumulated during five thousand years' history of ancient China, and much knowledge passed down from ancients is still valuable for modern people, therefore lots of researchers are exploring ancient Chinese science and technology (Jia et al., 2004; Zhu et al., 1998b,a) continuously.

Chinese idioms (Cheng Yu) have seen the history and culture of China, meanwhile they contain large number of scientific achievement of ancient China (Dai, 2003). For the example in Table 1, "一寸光阴一寸金" (One inch of time, one inch of gold) mentions time measurement technique using sundial of ancient Astronomy domain in Han dynasty (汉朝). Therefore, Chinese idioms are regarded as an important source of ancient scientific achievement information. However, existing Chinese online idiom dictionaries, such as Baidu Chinese Channel[1], Han dictionary[2] and Cihai online dictionaries[3], have limited function for scientific exploration. Those online idiom dictionaries mainly store basic information of idioms, e.g., pronunciation, explanation, source, synonyms and antonyms, and they can be leveraged to search idioms by names or keywords and to get basic information of idioms, but it is difficult for researchers to get idioms by domain and dynasty information, and it is also impossible to obtain the trend of scientific progress from idiom perspective.

In this paper, we propose a **S**cience **T**oolkit for **A**ncient **C**hina (STAC) based on a Chinese idiom knowledge graph aiming to support scientific exploration. We first extract domains and dynasties from explanation and source of idioms, and then associate domains and dynasties with idioms to construct the idiom knowledge graph. Based on the knowledge graph, we design and implement idiom navigator, idiom card and idiom QA of STAC toolkit. Idiom navigator provides a visual presentation for relations among idioms, dynasties and domains, reflecting overall scientific progress from idiom perspective, and idiom card gives basic information of idioms contained in users' text, such as dynasty, domain, explanation and source, and idiom QA answers idioms to questions about dynasties and domains, such as "宋代的天文领域成语" (The idioms on Astronomy domain in Song dynasty). Both idiom card and idiom QA are designed for scenarios of text reading and writing to shorten the path of users' actions and avoid users' thinking being interrupted.

2 Dataset

We mainly collect idiom data from Han dictionary and Baidu Chinese Channel, and Han dictionary

[1]https://dict.baidu.com/
[2]www.zdic.net
[3]For example, http://www.cihai123.com/

Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, pages 57–61
Minneapolis, USA, June 6, 2019.

Name	一寸光阴一寸金 (One inch of time, one inch of gold)
Explanation	一寸光阴和一寸长的黄金一样昂贵，其中“一寸光阴”是指晷针的影子在晷盘上移动一寸距离所使用的时间。 (One inch of time is as expensive as one inch of gold, where “one inch of time” refers to time taking by shadow of gnomon to move one inch distance on sundial plate.)
Source	刘安所著《淮南子》 (“Huai Nan Zi” of Liu An)
Domain	天文 (Astronomy)
Dynasty	汉朝 (Han dynasty)

Table 1: An example of Chinese idioms.

is the most reliable and Baidu Chinese is much more comprehensive. Firstly, we get 31,605 idioms from Han dictionary and 30,923 idioms from Baidu Chinese Channel respectively, and properties of these idioms include pronunciation, explanation and source. Then we merge the two idiom sets by setting Han dictionary prior to Baidu dictionary for the duplicate idioms. The final dataset is stored in MySQL database, containing 31,632 idioms, whose average number of characters in explanation is 24 and average number of characters in source is 32.

3 Idiom Knowledge Graph (IKG) Construction

We construct an idiom knowledge graph based on the dataset collected in Section 2. Hereinafter the idiom knowledge graph is referred to as IKG. The ontology definition of IKG contains:

(1) three types of entities, which are idiom entity denoted as *IDIOM*, dynasty entity denoted as *DYNASTY*, and domain entity denoted as *DOMAIN*;

(2) three types of properties, which are explanation of idiom denoted as $explanation\_of$, source of idiom denoted as $source\_of$, and pronunciation of idiom denoted as $pronunciation\_of$;

(3) two types of relations, which are relation between dynasties and idioms denoted as $dynasty\_of$, and relation between domains and idioms denoted as $domain\_of$.

Instances of *IDIOM* are selected from 31,632 idioms of the dataset in relation extraction process, and instances of $explanation\_of$, $source\_of$ and $pronunciation\_of$ are queried directly from the dataset. There are 14 *DYNASTY* instances, which are defined according to the main dynasties of ancient Chinese history, such as “战国” (Warring), “汉” (Han) and “宋” (Song), and there are 11 *DOMAIN* instances, which almost cover all the domains in ancient China, such as “天文” (Astronomy), “手工业 ”(Handicraft) and “医药” (Medicine).

The relation extraction process of $domain\_of$ and $dynasty\_of$ is divided into following steps as shown in Figure 1:

(1) For each idiom in the dataset, concat its explanation string and source string and tokenize the result string into a word bag with jieba tool[4], and then for each word in the word bag, add its hypernym and hyponym words from semantic dictionaries (e.g., HowNet[5]) into the word bag, until the word bag is no longer changing in its size, and the result word bag is used as a feature of the idiom.

(2) Load a Chinese word vectors corpus pretrained on Chinese Wikipedia and Baidu Encyclopedia (Li et al., 2018), and then embeddings of 31,632 idioms, 14 *DYNASTY* instances and 11 *DOMAIN* instances can be looked up from it.

(3) Compute correlation based on WMD (Word Mover’s Distance) algorithm (Kusner et al., 2015) that can achieve better results for short texts, and confirm final relations by human reviewers:

- for each *DOMAIN* instance, compute its correlation with all the idioms, and send top 100 idioms for human review to confirm final instances of $domain\_of$ relation;

[4]https://github.com/fxsjy/jieba
[5]http://www.keenage.com/zhiwang/c_zhiwang.html

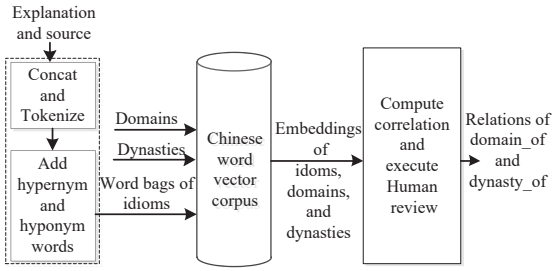


Figure 1: Relation extraction framework of IKG.

- for each idiom, compute its correlation with all the *DYNASTY* instances and send top 3 *DYNASTY* entities for human review to confirm final instances of $dynasty\_of$ relation, and human reviewers could make decision with information of books and authors contained in source text, e.g., "Huai Nan Zi" (《淮南子》) and "Liu An" (刘安) in Table 1.

Finally, 542 instances of $domain\_of$ relation are extracted and 532 *IDIOM* instances are selected from the 31,632 idioms, and for the 532 *IDIOM* instances, 541 instances of $dynasty\_of$ relation are extracted. The whole knowledge graph is stored in Neo4j[6] graph database.

Figure 2 describes some statistics about IKG. From Figure 2(a), we can see that the scientific progress in "战国" (Warring), "汉" (Han) and "宋" (Song) is more significant than in other dynasties, and from Figure 2(b), we can see that the scientific progress in "医药" (Medicine), "手工业"(Handicraft) and "物理" (Physics) is more significant than in other domains.

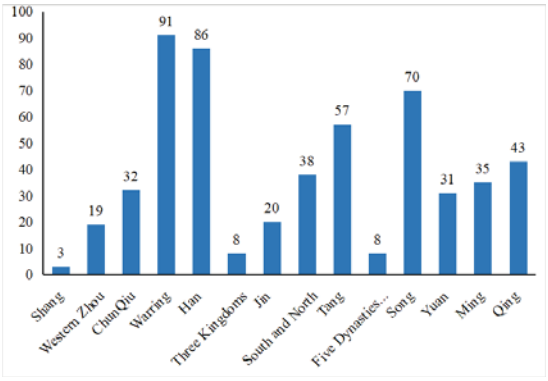
4 STAC Toolkit

Based on IKG, we design and implement STAC toolkit for scientific exploration of ancient China, and the toolkit contains functions of idiom navigator, idiom card and idiom QA.

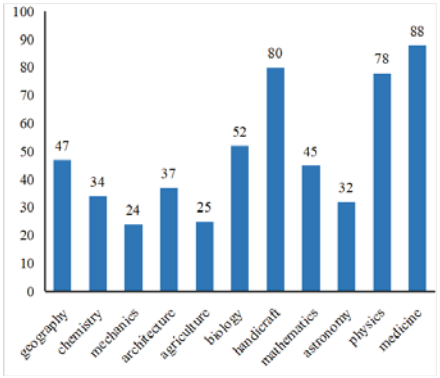
4.1 Idiom navigator

Idiom navigator is an idiom visualization tool, and it gets all the idioms, dynasties and domains from IKG and organizes them in tree structures based on the relations contained in IKG. With idiom navigator, users can browse idioms starting from dynasties or domains. For example, after selecting each dynasty, users can get expanded all the domains that were developed in the dynasty, and after selecting one of these domains, they can also get expanded all the idioms related with both the domain and the dynasty. Then users could gain information on scientific progress level from idiom perspective.

[6]https://neo4j.com/



(a) Distribution of *IDIOM* instances across *DYNASTY* instances.



(b) Distribution of *IDIOM* instances across *DOMAIN* instances.

Figure 2: Distribution of *IDIOM* instances across *DYNASTY* instances and *DOMAIN* instances.

4.2 Idiom card

Idiom card provides basic information for the idioms contained in users' text, and users do not need to switch to online idiom dictionaries, therefore the action path to get information of idioms is shortened and users' thinking is not interrupted.

Given a piece of text, we first extract all the idioms from the text by multi-pattern matching algorithm (e.g., Aho-Corasick string match algorithm (Aho and Corasick, 1975)), and then for each idiom extracted, we query its dynasty, domain, explanation, source and pronunciation from IKG. In detail, domain and dynasty are queried by relations, and explanation, source and pronunciation are queried by properties. Finally, queried results for all the idioms are presented to users.

Figure 3: Snapshot of STAC toolkit landing page.

4.3 Idiom QA

For questions about dynasties and domains, idiom QA gives corresponding accurate idioms as answer.

Given a question, we first extract all the dynasties and domains from the question by multi-pattern matching algorithm, and then we construct a query statement using the extracted dynasties and domains and execute the statement on IKG to get idioms. In detail, the query statement is constructed as "select all the idioms that are associated with all the dynasties and domains". Finally, queried idioms are presented to users.

4.4 Deployment

STAC toolkit is developed using Django[7] web framework as backend, Neo4j as graph database and Vue.js JavaScript library[8] for frontend page rendering, and we implement visualization of idiom navigator with Zoomcharts.js library[9]. STAC toolkit is currently accessible at http://120.92.208.22:7476/demo/#/stac with Google Chrome browser (Please note that the first visit may take about 60 seconds). Figure 3 shows a snapshot of STAC toolkit landing page.

We are continuously improving STAC toolkit, and currently, users can use it in the following way:

(1) Call out idiom navigator visualization tool by clicking button "Domains" or "Dynasties", and double-click nodes of type *DOMAIN* or *DYNASTY* to expand related nodes until reaching end. Please note that some nodes of *DOMAIN* or *DYNASTY* cannot be expanded because there are no related nodes under them.

(2) Input some text into the left "Input" area, and get card for idioms contained in text by clicking button "Card". For example, input "从'一寸光阴一寸金'可知..." (From the idiom "One inch of time, one inch of gold" we can see that...) and the idiom card of "一寸光阴一寸金" is displayed in the right "Idiom" area, containing its dynasty, domain, explanation, source and pronunciation.

(3) Call out QA dialog box by clicking button "QA", and then enter some question about dynasties and domains (e.g., the question example in Section 1), and finally click "OK" button to get idioms as answer. Meanwhile idioms in answer could be inserted into text by clicking button "Insert". Please note that dynasties and domains in questions are assumed to be correct, and similar words are not supported for questions.

(4) Call out glossary window by clicking button "Glossary", and then read the Chinese-English glossary of dynasties and domains.

5 Conclusion

In this paper, we first construct a Chinese idiom knowledge graph and then propose STAC toolkit that contains functions of idiom navigator, idiom card and idiom QA for scientific exploration. Currently, idiom navigator helps users explore overall scientific progress from idiom perspective, and idiom card and idiom QA shorten action path and avoid thinking being interrupted while users are reading and writing. In future, we plan to improve idiom QA by context understanding and conduct more evaluations on the idiom knowledge graph and STAC.

[7]https://www.djangoproject.com/
[8]https://vuejs.org/
[9]https://zoomcharts.com/

References

Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

Wusan Dai. 2003. *Ancient Technology in Idioms*. Baihua Literature and Art Publishing House, Tijin, China.

Wei Jia, Wenyuan Gao, Yongqing Yan, Jie Wang, Zhaohui Xu, Wenjie Zheng, and Peigen Xiao. 2004. The rediscovery of ancient chinese herbal formulas. *Phytotherapy Research: An International Journal Devoted to Pharmacological and Toxicological Evaluation of Natural Product Derivatives*, 18(8):681–686.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Jiashi Zhu, Georges M Halpern, and Kenneth Jones. 1998a. The scientific rediscovery of a precious ancient chinese herbal regimen: Cordyceps sinensis part ii. *The Journal of Alternative and Complementary Medicine*, 4(4):429–457.

Jiashi Zhu, Georges M Halpern, and Kenneth Jones. 1998b. The scientific rediscovery of an ancient chinese herbal medicine: Cordyceps sinensis part i. *The Journal of alternative and complementary medicine*, 4(3):289–303.