2019 IEEE 35th International Conference on Data Engineering (ICDE)

# CN-Probase: A Data-driven Approach for Large-scale Chinese Taxonomy Construction

Jindong Chen[1], Ao Wang[1], Jiangjie Chen[1], Yanghua Xiao[12*], Zhendong Chu[1], Jingping Liu[1], Jiaqing Liang[13], Wei Wang[1]
[1]*Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, China*
[2]*Shanghai Institute of Intelligent Electronics & Systems, Shanghai, China*
[3]*Shuyan Technology, Shanghai, China*
{*chenjd16, awang15, jiangjiechen14, shawyh, zdchu15, jpliu17*}*@fudan.edu.cn, l.j.q.light@gmail.com, weiwang1@fudan.edu.cn*

***Abstract*—Taxonomies play an important role in machine intelligence. However, most well-known taxonomies are in English, and non-English taxonomies, especially Chinese ones, are still very rare. In this paper, we focus on automatic Chinese taxonomy construction and propose an effective generation and verification framework to build a large-scale and high-quality Chinese taxonomy. In the generation module, we extract *isA* relations from multiple sources of Chinese encyclopedia, which ensures the coverage. To further improve the precision of taxonomy, we apply three heuristic approaches in verification module. As a result, we construct the largest Chinese taxonomy with high precision about 95% called CN-Probase. Our taxonomy has been deployed on Aliyun, with over 82 million API calls in six months.**

***Keywords*-Knowledge Base; Taxonomy Construction;**

## I. Introduction

Semantic networks and conceptual taxonomies are playing an increasingly important role in many applications. Conceptual taxonomies are composed of entities, concepts and hypernym-hyponym relations (a.k.a *isA* relations). For example, `apple` *isA* `fruit`, where `fruit` is the *hypernym* of `apple`. The opposite term for hypernym is *hyponym*, so `apple` is the hyponym of `fruit`. We use the expression *isA*(`A`, `B`) to express a hypernym-hyponym relationship, which means `A` is a hyponym of `B`.

Early taxonomies such as WordNet [1] and Cyc [2] are built by human experts. They are highly precise but limited in *coverage* and are expensive to construct. Therefore, most of succeeding research efforts are devoted to constructing taxonomies *automatically* from web corpus or online encyclopedia. These efforts have produced a lot of well-known English taxonomies such as WikiTaxonomy [3] and Probase [4]. However, non-English taxonomies, especially Chinese ones, are still very rare. The direct reason is the complexity of Chinese. Chinese is a lower-resourced language with flexible expressions and grammatical rules [5]. For example, the order of words in Chinese can be changed flexibly in the sentence. Besides, Chinese has no word spaces, no explicit tenses and voices, no distinct singular/plural forms.

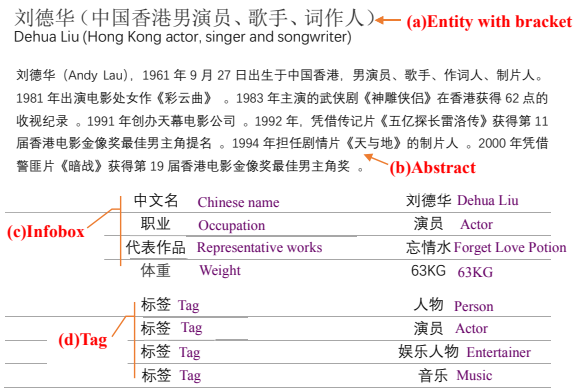


Figure 1: Page in Chinese encyclopedia.

In this paper, we aim at constructing a large-scale and high-quality Chinese taxonomy automatically from a Chinese encyclopedia website. Our observation is that there are multiple sources including bracket, abstract, infobox, tag in Chinese encyclopedia marked as (a), (b), (c), and (d) respectively, as shown in Figure 1. We highlight that these information, yet to be fully leveraged from previous work, is the key to achieve our objective. For example, the bracket in Figure 1 allows us to extract *isA*(`Dehua Liu`, `singer`). The triples in infobox, such as <`Dehua Liu, occupation, actor`>, allows us to extract *isA*(`Dehua Liu`, `actor`). And some tags directly tell us that *isA*(`Dehua Liu`, `person`). Thus, the full usage of these information allows us to find a significant number of *isA* relations. However, the tags still contain noise, and the inference of hypernyms from triples and text are still error-prone. For example, *isA*(`Dehua Liu`, `music`) is a wrong *isA* pair extracted from tag.

To solve the above problems, we propose a *generation and verification* framework, which is shown in Figure 2. The input of our framework is Chinese encyclopedia. In the *generation step*, we leverage different algorithms to extract

*Yanghua Xiao is the corresponding author. This paper is the was supported by National Key R&D Program of China (No. 2017YFC0803700), by National NSFC (No.61732004) and by Shanghai Municipal Science and Technology project (No.16JC1420400).

2375-026X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDE.2019.00178

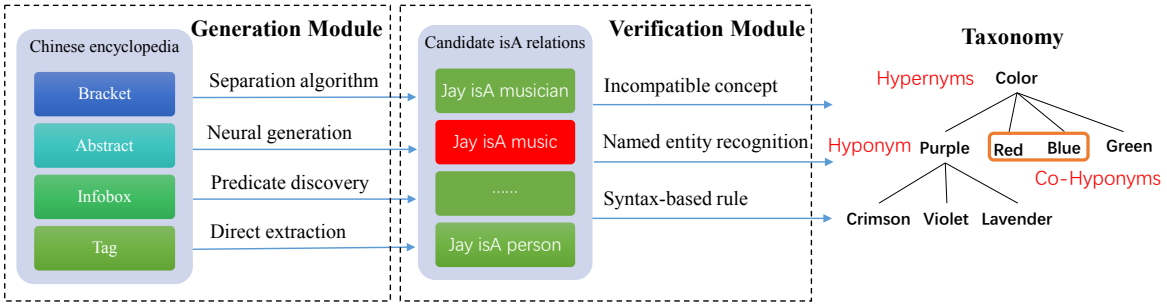


Figure 2: Framework of CN-Probase.

*isA* relations from multiple sources of Chinese encyclopedia, which ensures the *coverage*. The candidate *isA* relations are produced by merging all *isA* relations generated from different sources of Chinese encyclopedia. In the *verification step*, we employ three heuristic strategies to remove noise, which ensures the *precision* of *isA* relations. A candidate *isA* relation will be filtered if any of the strategies makes the judgment that it is a wrong case. Our contributions in this paper can be summarized as follows.

- We design an effective generation and verification framework for Chinese taxonomy construction.
- We build the large-scale Chinese conceptual taxonomy CN-Probase with high precision (95%), including 15 million disambiguated entities, 270 thousand distinct concepts and 33 million *isA* relations.
- In experiments, we demonstrate the size, precision and coverage of CN-Probase.

## II. Generation Module

In this section, we acquire *isA* relations from four sources of Chinese encyclopedia (i.e., *bracket, abstract, infobox* and *tag*) by four corresponding algorithms.

*Separation algorithm* is proposed to acquire the hypernyms of the entity from the noun compound in the bracket. The input of the algorithm is a disambiguated entity denoted as $e(x)$, where $e$ is the entity name and $x$ is the noun compound. Let $(x_1, x_2, ..., x_n)$ be the word sequence of length $n$ by conducting word segmentation on $x$. An example is shown in Figure 3, `ANT FINANCIAL chief strategy officer` is segmented into {`ANT`, `FINANCIAL`, `chief`, `strategy officer`}. The output of the algorithm is the hypernyms of the input entity. Let $\oplus$ denote the operation of string concatenation. The algorithm begins with the rightmost three elements of the word sequence. For simplicity, we will use $(x_{i-1}, x_i, x_{i+1})$ to explain the algorithm:

Step 1: Given $(x_{i-1}, x_i, x_{i+1})$, if PMI$(x_{i-1}, x_i)$<PMI$(x_i, x_{i+1})$ holds, the algorithm goes to *step 2*, otherwise goes to *step 3*.

Step 2: Separate the sequence as $(x_{i-1}, x_i \oplus x_{i+1})$. Then move the sliding window to left by one unit (word) and acquire $(x_{i-2}, x_{i-1}, x_i \oplus x_{i+1})$, go to *step 1*.

Step 3: Move the sliding window to left by one unit and acquire $(x_{i-2}, x_{i-1}, x_i)$, then go to *step 1*.

Step 4: When the leftmost element $x_1$ locates in sliding window and the sequence $(x_1, x_2, x_3)$ satisfies PMI$(x_1, x_2)$>PMI$(x_2, x_3)$, we separate it as $(x_1 \oplus x_2, x_3)$. Move the sliding window to right by one unit and acquire $(x_1 \oplus x_2, x_3, x_4)$, then go to *step 1*.

The output of the algorithm is a binary tree. We extract all the leaf nodes along with the rightmost path of the binary tree as the hypernyms. To the best of our knowledge, we are the first to extract *isA* relations from entity brackets, and we obtain nearly 2 million *isA* relation with a precision of 96.2% from this data source.

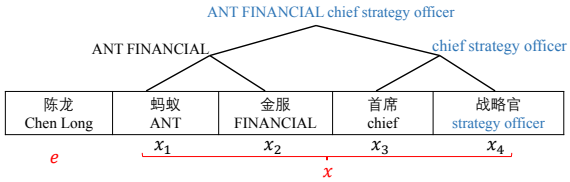


Figure 3: Example of hypernym acquisition. The blue phrases are the hypernyms of the entity.

*Neural generation* is used to obtain the hypernym (concept) of an entity from the abstract of the entity. We first utilize *distant supervision* [6] to construct dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where $n$ is the number of samples. We have acquired numerous *isA* relations with a precision over 96% from bracket . For $i$-th *isA* relation, we regard the abstract of its hyponym as $x_i$ and the hypernym as $y_i$. In this way, we build the dataset consisting of more than 300,000 samples. Then, we employ an encoder-decoder model to generate concepts from the abstract. But merely using this basic model suffers from out-of-vocabulary (OOV) problem. Hence we use copynet [7] to perform this task.

*Predicate discovery* is proposed to acquire *isA* relations

from the infobox. First, we apply the idea of *distant supervision* [6] to discover the predicates such as `occupation` which are the implicit *isA* relationships. Specifically, we employ the *isA* relations that have been extracted from bracket as prior knowledge, since they have a precision over 96%. Then, we use these *isA* relations (e.g., *isA*(`Jay Chou`, `singer`)) to align the SPO triples (e.g., <`Jay Chou`, `occupation`, `singer`>) and discover 341 candidate predicates (e.g., `occupation`) in total. However, there are noises in these candidates. To further purify these candidates, we manually select 12 predicates as the implicit *isA* relationships to acquire *isA* relations from its corresponding SPO triples.

*Direct extraction* is used to obtain *isA* relations from tag. A tag is a word or phrase which is used to describe the entities in Chinese encyclopedia. A majority of tags are the hypernyms of the entities. We directly regard the tags as the hypernyms of an entity.

## III. Verification Module

In this section, we propose three effective heuristic strategies to filter the wrong *isA* relations produced in the generation module and improve the precision.

### A. Incompatible concepts

Two concepts such as `singer` and `actor` are *compatible* since they have some common entities. In some case, two concepts such as `person` and `book` sharing no entities are *incompatible*, which motivates us to filter wrong *isA* relations by detecting incompatible concept pairs. Our approach is composed of two steps: incompatible concept pairs construction and wrong *isA* relations detection. In the first step, we construct the incompatible concept pairs based on the Jaccard similarity between the hyponyms set of two concepts and the cosine simiarity between the distribution of concept attribute. In the second step, given an entity $e$ and its two incompatible concepts $c_1$ and $c_2$, we detect the wrong one by the KL divergence:

$$D_{KL}(v_{att(e)}||v_{att(c)}) = -\sum_{x} v_{att(e)} log \frac{v_{att(c)}}{v_{att(e)}} \quad (1)$$

$v_{att(e)}$ and $v_{att(c)}$ are the attribute distribution of the entity $e$ and concept $c$. Then, we filter the concept with a larger KL score.

### B. Named entity recognition

The fact that whether a hypernym is a *named entity* (NE) plays an important role in detecting wrong *isA* relations, since NE usually cannot be a hypernym of an entity. For example, *isA*(`iPhone`, `America`) is a wrong *isA* relation due to the NE hypernym `America`. Inspired by the above observation, we detect wrong *isA* relations by recognizing the NE hypernyms. We use $s_1(H)$ and $s_2(H)$ to represent the support of a hypernym $H$ as a NE in the Chinese text corpus and our taxonomy respectively. Specifically, $s_1(H) = NE(H)/total(H)$ where $NE(H), total(H)$ represent the number of occurrence of $H$ as a NE and the total number of occurrence of $H$ in Chinese text corpus respectively. Similarly, we acquire $s_2(H)$ by replacing Chinese text corpus with our taxonomy. We further use a *noisy-or model* to combine the two scores.

$$s(H) = 1 - (1 - s_1(H)) \cdot (1 - s_2(H)) \quad (2)$$

The rationale of the noisy-or model is to amplify the support signal. We set the threshold emperically and filter the *isA* relations whose support $s(H)$ is greater than the threshold.

### C. syntax-based rule

We also use some syntax rules to further filter wrong *isA* relations. We describe the most typical rules as follows: (1) A good hypernym should not be a thematic word such as `politics`, `military`. We collect a Chinese lexicon from Li et al. [5] including 184 non-taxonomies, thematic words. Then we filter the *isA* relations whose hypernym in this lexicon; (2) The stem of the lexical head of hypernym should not occur in the non-head position of the hyponym. We filter the wrong candidate *isA* relations *isA*(`educational institution`, `education`) by this rule.

## IV. EXPERIMENTS

We apply the proposed framework on Chinese encyclopedia. As a result, we construct the large-scale and high-quality Chinese taxonomy including 270,026 distinct concepts, 15,066,667 disambiguated entities, 32,398,018 entity-concept relations and 527,288 subconcept-concept relations (32,925,306 *isA* relations in total). Readers can refer to *http://kw.fudan.edu.cn/cnprobase/search/* for complete experimental results.

### A. Experiment Settings

*Data Source:* CN-DBpedia [8] is one of the largest open-domain Chinese encyclopedia derived from Baidu Baike Hudong Baike and Chinese Wikipedia. The experimental dataset is from CN-DBpedia dump generated on May 20, 2017, which includes 15,990,349 entities, 8,096,835 pieces of abstract information, 132,435,632 SPO triples and 19,929,407 tags.

*Metrics:* There are five commonly used metrics in taxonomy evaluation: the number of entities, concepts and *isA* relations, precision and coverage. To estimate the precision of taxonomies, we randomly select 2000 *isA* relations in total from taxonomies and manually label whether a relation is correct or not.

*Baselines:* We compare CN-Probase with the following well-known Chinese taxonomies including Chinese WikiTaxonomy [5] , Bigcilin [9] and Probase-Tran proposed by us. We translate Probase from English to Chinese by utilizing Google Translator. Then we use three heuristic

Table I: Comparisons with other taxonomies. '-' represents results that are not provided.

| Taxonomy | # of entities | # of concepts | # of *isA* relations | precision |
|---|---|---|---|---|
| Chinese WikiTaxonomy | 581,616 | 79,470 | 1,317,956 | 97.6% |
| Bigcilin | 9,000,000 | 70,000 | 10,000,000 | 90.0% |
| Probase-Tran | 404,910 | 151,933 | 1,819,273 | 54.5% |
| **CN-Probase** | **15,066,667** | 270,025 | **32,925,306** | 95.0% |

methods from three aspects (meaning, transitivity, POS) to filter translation errors.

### B. Results

The main results are shown in Table I. Compared with other Chinese taxonomies, CN-Probase is the largest one when it comes to the number of entities, concepts and *isA* relations. The main reason is that we extract *isA* relations from multiple sources of Chinese encyclopedia. Besides, CN-Probase is a high-quality taxonomy with a precision 95% that outperforms Probase-Tran and Bigcilin. Although part of the noise has been reduced by three heuristic methods, the precision of Probase-Tran is still quite low due to various sources of noise. Hence simple cross-language translation cannot produce high-quality Chinese taxonomy. Bigcilin also extracts *isA* relations from multiple sources, but its precision is worse than ours since we use verification module to further improve the precision. Chinese WikiTaxonomy is built only from a single source (i.e. tag) of Chinese encyclopedia. As a result, it has a high precision but low coverage, and the number of *isA* relations in our taxonomy is 25x larger than Chinese WikiTaxonomy. We also evaluate the precision of each source for our taxonomy, and the precision of *isA* relations derived from the tag is 97.4% which is comparable to Chinese WikiTaxonomy.

Given that CN-Probase has more concepts and entities than other Chinese taxonomies such as Bigcilin and Probase-Tran, a reasonable question to ask is whether they are more effective in understanding text. We measure one aspect of the effectiveness here by examining CN-Probase's coverage on QA task. A question is said to be *covered* by a taxonomy if the question contains at least one concept or entity within the taxonomy. The dataset is from the QA task of NLPCC2016 that includes 23,472 questions. In all, CN-Probase covers 21,520 questions with a coverage of 91.68%. The covered entities have 2.14 concepts on average, which shows the significant effectiveness of CN-Probase in text understanding

## V. System and Application

By adopting the proposed methods, CN-Probase has already been deployed on Aliyun. We also publish three APIs on *http://kw.fudan.edu.cn/apis/cnprobase/* to make our taxonomy accessible from Web. By September 2018, these APIs have already been called 82 million times by research institutions and companies since published on March 2018. Table II shows the function of each API and their usage statistics. So far, CN-Probase has been used in many applications including short text classification [10], information extraction, etc.

Table II: APIs and their descriptions.

| API name | Given | Return | Count |
|---|---|---|---|
| men2ent | mention | entity | 43,896,044 |
| getConcept | entity | hypernym list | 13,815,076 |
| getEntity | concept | hyponym list | 25,793,372 |

## VI. Conclusion

In this paper, we propose an effective generation and verification framework for automatic taxonomy construction. We build the large-scale and high-quality Chinese taxonomy CN-Probase from multiple sources of our Chinese encyclopedia. So far, CN-Probase has been used in many real-world applications.

## References


[1] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the Acm*, vol. 38, no. 11, pp. 39–41, 1995.

[2] D. B. Lenat and R. V. Guha, *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Pub. Co, 1989.

[3] S. P. Ponzetto and M. Strube, "Wikitaxonomy: A large scale knowledge resource," in *Conference on ECAI 2008: European Conference on Artificial Intelligence*, 2008, pp. 751–752.

[4] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 481–492.

[5] J. Li, C. Wang, X. He, R. Zhang, and M. Gao, "User generated content oriented chinese taxonomy construction," in *Asia-Pacific Web Conference*. Springer, 2015, pp. 623–634.

[6] M. Mintz, S. Bills, R. Snow, and J. Dan, "Distant supervision for relation extraction without labeled data," in *Joint Conference of the Meeting of the ACL and the International Joint Conference on Natural Language Processing of the Afnlp: Volume*, 2009, pp. 1003–1011.

[7] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1631–1640.

[8] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, and Y. Xiao, *CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System*, 2017.

[9] R. Fu, B. Qin, and T. Liu, "Exploiting multiple sources for open-domain hypernym discovery." in *EMNLP*, 2013, pp. 1224–1234.

[10] J. Chen, Y. Hu, J. Liu, Y. Xiao, and H. Jiang, "Deep short text classification with knowledge powered attention," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2018.