# A Hybrid Chinese Language Model based on a Combination of Ontology with Statistical Method

**Dequan Zheng, Tiejun Zhao, Sheng Li and Hao Yu**
MOE-MS Key Laboratory of Natural Language Proceessing and Speech
Harbin Institute of Technology
Harbin, China, 150001
{dqzheng, tjzhao, lisheng, yu}@mtlab.hit.edu.cn

## Abstract

In this paper, we present a hybrid Chinese language model based on a combination of ontology with statistical method. In this study, we determined the structure of such a Chinese language model. This structure is firstly comprised of an ontology description framework for Chinese words and a representation of Chinese lingual ontology knowledge. Subsequently, a Chinese lingual ontology knowledge bank is automatically acquired by determining, for each word, its co-occurrence with semantic, pragmatics, and syntactic information from the training corpus and the usage of Chinese words will be gotten from lingual ontology knowledge bank for a actual document. To evaluate the performance of this language model, we completed two groups of experiments on texts reordering for Chinese information retrieval and texts similarity computing. Compared with previous works, the proposed method improved the precision of nature language processing.

## 1 Introduction

Language modeling is a description of natural language and a good language model can help to improve the performance of the natural language processing.

Traditional statistical language model (SLM) is fundamental to many natural language applications like automatic speech recognition[1], statistical machine translation[2], and information retrieval[3]. Different statistical models have been proposed in the past, but n-gram models (in particular, bi-gram and tri-gram models) still dominate SLM research. After that, other approaches were put forward, such as the combination of statistical-based approach and rule-based approach[4,5], self-adaptive language models[6], topic-based model[7] and cache-based model[8]. But when the models are applied, the crucial disadvantages are that they can't represent and process the semantic information of a natural language, so they can't adapt well to the environment with changeful topics.

Ontology was recognized as a conceptual modeling tool, which can descript an information system in the semantic level and knowledge level. After it was first introduced in the field of Artificial Intelligence[9], it was closed combined with natural language processing and are widely applied in many field such as knowledge engineering, digital library, information retrieval, semantic Web, and etc.

In this paper, combining with the characteristic of ontology and statistical method, we present a hybrid Chinese language model. In this study, we determined the structure of Chinese language model and evaluate its performance with two groups of experiments on texts reordering for Chinese information retrieval and texts similarity computing.

The rest of this paper is organized as follows. In section 2, we describe the Chinese language model. In section 3, we evaluate the language model by several experiments about natural language processing. In section 4, we present the conclusion and some future work.

## 2 The language model description

Traditional SLM is make use to estimate the likelihood (or probability) of a word string, in

this study, we determined the structure of Chinese language model, first, we gave the ontology description framework of Chinese word and the representation of Chinese lingual ontology knowledge, and then, automatically acquired the usage of a word with its co-occurrence of context in using semantic, pragmatics, syntactic, etc from the corpus to act as Chinese lingual ontology knowledge bank. In actual document, the usage of lingual knowledge will be gotten from lingual ontology knowledge bank.

### 2.1 Ontology description framework

Traditional ontology mainly emphasizes the interrelations between essential concept, domain ontology is a public concept set of this domain[10]. We make use of this to present Chinese lingual ontology knowledge bank.

In practical application, ontology can be figured in many ways[11], natural languages, frameworks, semantic webs, logical languages, etc. Presently, popular models, such as Ontolingua, CycL and Loom, are all based on logical language. Though logical language has a strong expression, its deduction is very difficult to lingual knowledge. Semantic web and natural language are non-formal, which have disadvantages in grammar and expression.

For a Chinese word, we provided a framework structure that can be understood by computer combined with WordNet, HowNet and Chinese Thesaurus. This framework includes a Chinese word in concept, part of speech (POS), semantic, synonyms, English translation. Figure1 shows the ontology description framework of a Chinese word.

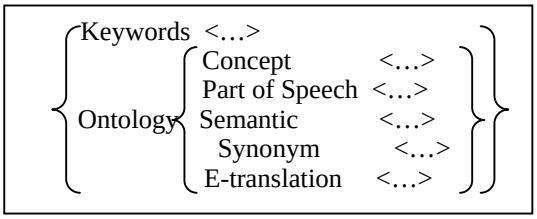


**Fig. 1. Ontology description framework**

### 2.2 Lingual ontology knowledge representation

A word is the basic factor that composes the natural language, to acquire lingual ontology knowledge, we need to know POS, means and semantic of a word in a sentence. For example, for a Chinese sentence, the POS, means and Semantic label of "打" in HowNet are shown in table 1. For the Chinese sentence "外国游客来北京游玩。", after words segmented, POS tagging and semantic tagging, we get a characteristic string. They are shown in table 2.

**Table 1. the usage of "打" in Chinese sentence**

| Chinese Sentence | POS | Means | Semantic Num |
|---|---|---|---|
| 打毛衣 | Verb | Weave | 525(weave\|编 辫) |
| 打酱 油 | Verb | Buy | 348(buy\|买) |

**Table 2. Segmentation, POS and Semantic tagging**

| Items | Results ("游客" acts as keyword) |
|---|---|
| Chinese sentence | 外国游客来北京游玩。 |
| Words segmentation | 外国 游客 来 北京 游玩 。 |
| POS tagging | 外国 nd/ 游客 Keyword/来 vg/ 北京 nd/ 游玩 vg/ 。 wj/ |
| Semantic label tagging | 外国 nd/021243 游客 Keyword/070366 来 vg/017545 北京 nd/021243 游玩 vg/092317 。 wj/-1 |
| Characteristic string | nd/021243 游客 Keyword/070366 vg/017545 nd/021243 vg/092317 |
| Explanation of Semantic label | 021243 represents "地方", 070366 represents "人", 092317 represents "消闲 ", "-1" represents not to be defined or exist this semantic in HowNet. |

In order to use and express easily, we gave a description for ontology knowledge of every Chinese word, which learned from corpus, to be shown as expression 1. All of them composed the Chinese lingual ontology knowledge bank.

$$\left(KeyWord(ontology),\bigcup_{l=1}^{m}\left(Sem_l,POS_l,L,\overline{C_l}\right)\bigcup_{r=1}^{n}\left(Sem_r,POS_r,L,\overline{C_r}\right)\right)$$

Where, *KeyWord*(ontology) is the ontology description of a Chinese word, $\left(Sem_i,POS_i,L,\overline{C_i}\right)$ is the left co-occurrence knowledge of a Chinese word got from its context and $\left(Sem_i,POS_i,L,\overline{C_i}\right)$ is the right co-occurrence knowledge. Symbol " ∪ " represents the aggregate of all the co-occurrence with the *KeyWord*.

$\left(Sem_i,POS_i,L,\overline{C_i}\right)$ denotes the multi-grams from context of a Chinese word, which is composed of semantic information $Sem_i$, part of speech $POS_i$, the position $L$ from the word *KeyWord* to its co-occurrence, the average distance $\overline{C_i}$ from the word to its left (or right) *i*-th word.

$(Keyword,(Sem_i,POS_i,L))$ denotes a semantic relation pair between the *keyword* and its co-occurrence in current context.

The multi-grams of a Chinese word in context, including the co-occurrence and their position will act as the composition of lingual ontology knowledge too. In figure 2, the characteristic string $W_1$, $W_2$, …, $W_i$ represents POS and semantic label, *Keyword* is keyword itself, *l* or *r*

is the position of word that is left or right co-occurrence with keyword.


**Fig. 2. Co-occurrence and the position information**

### 2.3 Lingual ontology knowledge acquisition

According to the course that human being acquires and accumulates knowledge, we propose a measurable description for Chinese lingual ontology knowledge through automatically learning typical corpus. In this approach, we will acquire the usage of a Chinese word in semantic, pragmatic and syntactic in all documents. We combine with the multi-grams in context including its co-occurrence, POS, semantic, synonym, position. In practical application, we will process every Chinese keyword that has the same grammar expression, semantic representation and syntactic structure with Chinese lingual ontology knowledge bank.

#### 2.3.1 Algorithm of automatic acquisition

Step 1: corpus pre-processing.

For any Chinese document $D_i$ in the document set $\{D\}$, we treat the sentence that includes keyword as a processing unit. First, we have a Chinese word segmentation, POS tagging, Semantic label tagging based on HowNet, and then, confirm a word to act as the keyword for acquiring its co-occurrence knowledge. We wipe off the word that can do little contribution to the lingual ontology knowledge, such as preposition, conjunction, auxiliary word and etc.

Step 2: Unify the keyword.

Making use of the ontology description of Chinese word, we make the synonym into uniform one.

Step 3: Calculate the co-occurrence distance.

In our proposal, first, we treat the sentence that includes keyword as a processing unit and make POS tagging, semantic label tagging, then, we get Characteristic string. We take the keyword as the center, define the left and right distance factor $B_l$ and $B_r$ to be shown at formula 1.

$$B_l = \frac{1-\frac{1}{2}}{1-\left(\frac{1}{2}\right)^m} \qquad B_r = \frac{1-\frac{1}{2}}{1-\left(\frac{1}{2}\right)^n} \qquad (1)$$

Where, $m$ and $n$ represent the left and right number of word that centered with the keyword. In this way, we try to get the language intuition, in a word, if the co-occurrence is nearer to the keyword, we will get more the co-occurrence distant. Final, we respectively get the left-side and right-side co-occurrence distant from keyword to its co-occurrence to be shown as formula 2.

$$C_{li} = \left(\frac{1}{2}\right)^{i-1} B_l \ (\text{i=1,}\cdots\text{,m})$$
$$C_{rj} = \left(\frac{1}{2}\right)^{j-1} B_r \ (\text{j=1,}\cdots\text{,n}) \qquad (2)$$

Step4: Calculate the average co-occurrence distance.

For a keyword, in the current sentence of document $D_i$, we regard the keyword and its co-occurrence ($Sem_i$, $POS_i$, $L$) as semantic relation pair, and $C_j$ is their co-occurrence distance. We calculate the average of $C_j$ that appear in corpus and act as the average co-occurrence distance $\overline{C_l}$ between the keyword and its co-occurrence ($Sem_i$, $POS_i$, $L$).

When all of documents are learned, all of keyword and their co-occurrence information $\left(Sem_i, POS_i, L, \overline{C_i}\right)$ compose the Chinese lingual ontology knowledge bank.

Step 5: Rebuild the index.

In order to improve the processing speed, for acquired lingual ontology knowledge bank, we first build an index according to Chinese word, and then, we respectively make a sorting according to the semantic label $Sem_i$ for every Chinese word.

#### 2.3.2 Lingual ontology knowledge application

In practical application, we will respectively get different evaluation of a document from the lingual ontology knowledge bank. For the natural language processing, e.g. documents similarity computing, text re-ranking for information retrieval, information filtering, the general processing is as follow.

Step 1: Pre-processing and unify the keyword.

The processing is the same as Step 1 and Step 2 in section 2.3.1.

Step 2: Fetch the average co-occurrence distance from lingual ontology knowledge bank.

We regard a sentence including keyword in document $D$ as a processing unit. First, we make POS tagging, semantic label tagging and get Characteristic string, and then, for every keyword, if it has the same semantic relation pair as lingual ontology knowledge bank, i.e. the keyword and its co-occurrence ($Sem_i$, $POS_i$, $L$) in practical document is the same one as lingual

ontology knowledge bank, we add up all the average co-occurrence distance $\overline{C_l}$ from Chinese lingual ontology knowledge bank acquired in section 2.3.1.

Step 3: Get the evaluation value of a document.

Repeat Step 2 until all keywords be processed and the accumulation of the average co-occurrence distance $\overline{C_l}$ will act as the evaluation value of current document.

## 3 Evaluation of language model

We completed two groups of experiments on text re-ranking for information retrieval, text similarity computing to verify the performance of lingual ontology knowledge.

### 3.1 Texts reordering

Information retrieval is used to retrieve relevant documents from a large document set for a user query, where the user query can be a simple description by natural. As a general rule, users hope more to acquire relevant information from the top ranking documents, so they concern more on the precision of top ranking documents than the recall.

We use the Chinese document set CIRB011 (132,173 documents) and CIRB020 (249,508 documents) from NTCIR3 CLIR dataset and select 36 topics from 50 search topics (see http://research.nii.ac.jp/ntcir-ws3/work-en.html for more information) to evaluate our method. We use the same method to retrieve documents mentioned by Yang Lingpeng[12], i.e. we use vector space model to retrieve documents, use cosine to calculate the similarity between document and user query. We respectively use bi-grams and words as indexing units[13,14], the average precision of top N ranking documents acts as the normal results. In this paper, we used a Chinese dictionary that contains about 85,000 items to segment Chinese document and query.

To measure the effectiveness of information retrieval, we use the same two kinds of relevant measures: relax-relevant and rigid-relevant[14,15]. A document is rigid-relevant if it's highly relevant or relevant with user query, and a document is relax-relevant if it is high relevant or relevant or partially relevant with user query. We also use PreAt10 and PreAt100 to represent the precision of top 10 ranking documents and top 100 ranking documents.

#### 3.1.1 Strategy of texts reordering

First, we get some keywords to every topic by query description. For example,

Title: 克隆之诞生 (The birth of a cloned calf)

Description: 查询与使用被称为体细胞移植的技术创造克隆牛相关的文章 (Find Articles relating to the birth of cloned calves using the technique called somatic cell nuclear transfer)

We extract "克隆, 体细胞, 移植, 无性繁殖" as feature word in this topic.

Second, acquire lingual ontology knowledge every topic by their feature words. In this proposal, we arrange 300 Chinese texts of this topic as learning corpus to get lingual ontology knowledge bank.

Third, get the evaluation value of every text about this topic, i.e. respectively add up all the average co-occurrence distance $\overline{C_l}$ to the same semantic relation pairs in every text from lingual ontology knowledge bank.

If a text has several keywords, repeat step3 to acquire every evaluation value to these keywords, and then, add up each evaluation value to act as the text evaluation value.

Final, we reorder the initial retrieval texts according to the every text evaluation value of every topic.

#### 3.1.2 Experimental results and analysis

We calculate the evaluation value of every text in each topic to reorder the initial relevant documents.

Table 3 lists the normal results and our results based on bi-gram indexing, our results are acquired based on Chinese lingual ontology knowledge to enhance the effectiveness. PreAt10 is the average precision of 36 topics in precision of top 10 ranking documents, while PreAt100 is top 100 ranking documents.

Table 4 lists the normal results and our results based on word indexing. Ratio displays an increase ratio of our result compared with normal result.

**Table 3. Precision (bi-gram as indexing unit)**

| Items | Normal | Our method | Ratio |
|---|---|---|---|
| PreAt10 (Relax) | 0.3704 | 0.4389 | 18.49% |
| PreAt100 (Relax) | 0.1941 | 0.2239 | 15.35% |
| PreA10 (Rigid) | 0.2625 | 0.3083 | 17.45% |
| PreAt100 (Rigid) | 0.1312 | 0.1478 | 12.65% |

**Table 4. Precision (word as indexing unit)**

| Items | Normal | Our method | Ratio |
|---|---|---|---|
| PreAt10 (Relax) | 0.3829 | 0.4481 | 17.03% |
| PreAt100 (Relax) | 0.2022 | 0.2306 | 14.05% |
| PreAt10 (Rigid) | 0.2745 | 0.3169 | 15.45% |
| PreAt100 (Rigid) | 0.1405 | 0.1573 | 11.96% |

In table 3, it is shown that compared with bi-grams as indexing units, our method respectively increases 18.49% in relax relevant measure and 17.45% in rigid in PreAt10. In PreAt100 level, our method respectively increases 15.35% in relax relevant and 12.65% in rigid relevant measure. Figure 3 displays the PreAt10 values of each topic in relax relevant measure based on bi-gram indexing where one denotes the precision enhanced with our method, another denotes the normal precision. It is shown the precision of each topic is all improved by using our method.

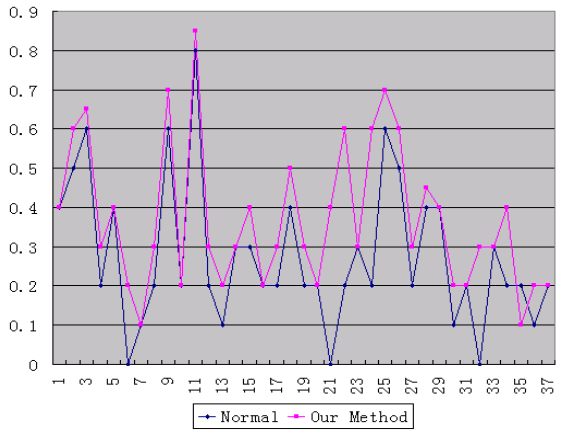


**Fig. 3. PreAt10 of all topics in relax judgment**

In table 4, using words as indexing units, our method respectively increases 17.03% in relax relevant measure and 15.45% in rigid in PreAt10. In PreAt100 level, our method respectively increases 14.05% in relax relevant measure and 11.96% in rigid.

In our experiments, compared with the two Chinese indexing units: bi-gram and words, our method increases the average precision of all queries in top 10 and top 100 measure levels for about 17.1% and 13.5%. What lies behind our method is that for each topic, we manually select some Chinese corpus to acquire the lingual ontology knowledge, and can help us to focus on relevant documents. Our experiment also shows improper extract and corpus may decrease the precision of top documents. So our method depends on right keywords in texts, queries and the corpus.

## 3.2 Text similarity computing

Text similarity is a measure for the matching degree between two or more texts, the more high the similarity degree is, the more the meaning of text expressing is closer, vice versa. Some proposal methods include Vector Space Model[16], Ontology-based[17], Distributional Semantics model[18].

### 3.2.1 Strategy of similarity computation

First, for two Chinese texts $D_i$ and $D_j$, we respectively extract $k$ same feature words, if the same feature words in the two texts is less than k, we don't compare their similarity.

Second, acquire lingual ontology knowledge every text by their feature words.

Third, get the evaluation value of every text, i.e. respectively add up all the average co-occurrence distance $\overline{C_l}$ to the same semantic relation pairs in two texts.

Final, compute the similarity ratio of every two text $D_i$ and $D_j$. The similarity ratio equals to the ratio of the similarity evaluation value of text $D_i$ and $D_j$, if the ratio is in the threshold $\alpha$, then we think that text $D_i$ is similar to text $D_j$.

### 3.2.2 Experimental results and analysis

We download four classes of text for testing from Sina, Yahoo, Sohu and Tom, which include 71 current affairs news, 68 sports news, 69 IT news, 74 education news.

For the test of current affairs texts, according to the strategy of similarity computation, we choose five words as feature word. They are "贸易, 协议, 谈判, 中国, 美国". In the texts, the word "经贸, 商贸" are all replaced by word "贸易" and other classes are similar. The testing result is shown in table 5.

**Table 5. Testing results for text similarity**

| Items | $0.95<\alpha<1.05$ | | | $0.85<\alpha<1.15$ | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$-measure | Precision | Recall | $F_1$-measure |
| Current affairs news | 97.14% | 97.14% | 97.14% | 94.60% | 100% | 97.23% |
| Sports News | 88.57% | 91.18% | 89.86% | 84.62% | 97.06% | 90.41% |
| IT news | 93.75% | 96.77% | 95.24% | 91.18% | 100% | 95.39% |
| Education news | 94.74% | 97.30% | 96.00% | 90.24 | 100% | 94.87% |
| General results | 93.57% | 95.62% | 94.58% | 90.07% | 99.27% | 94.42% |

We analyzed all the experimental results to find that the results for current affairs texts are the best, while the sports texts are lower than others. We think it is mainly because some sports terms are unprofessional for the lower sports texts recognition, such as "汉家军, 救主, 郝董". Other feature words are more fixed and more concentrated.

## 4 Conclusion

In this paper, we presented a hybrid Chinese language model based on a combination of ontology with statistical method. We discuss the modeling and evaluate its performance. In the test about texts reordering, our experiences show that our method can increase the performance of Chinese information retrieval about 17.1% and 13.5% at top 10 and top 100 documents measure level. In another test about texts similarity computing, F1-measure is above 95%.

On the other hand, in the current disposal of our information processing, we only make use of some characteristics ontology and use some co-occurrence information, such as semantics, POS, context, position, distance, and etc. For the further research and experiment, we will be on the following: (1) Research on the characteristics of relations between semantics and combine with some mature natural language processing techniques. (2) Research traditional ontology representation to keep up with international stand. (3) Apply our key techniques to English information retrieval and cross-lingual information retrieval systems and study a general approach.

## References


1. Jelinek, F. 1990. Self-organized language modeling for speech recognition. In Readings in Speech Recognition,A. Waibel and K. F. Lee, eds. Morgan-Kaufmann, San Mateo, CA,1990, 450-506.
2. Brown, P., Pietra, S. D., Pietra, V. D., and Mercer, R. 1993. The mathematics of statistical machine-translation: Parameter estimation. Computational Linguistics 19, 2 (1993), 269-311.
3. Croft, W. B. and Lafferty, J. (EDS.) 2003. Language Modeling for Information Retrieval. Kluwer Academic,Amsterdam.
4. Wang Xiaolong, Wang Kaizhu. 1994. Speech input by sentence, Chinese Journal of Computers, 17(2): 96-103
5. Zhou Ming, Huang Changning, Zhang Min, Bai Shuanhu, and Wu Sheng. 1994. A Chinese parsing model based on corpus, rules and statistics, Computer research and development, 31(2):40-49
6. R DeMori, M Federico. 1999. Language model adaptation. In: Keith Pointing ed. Computational Models of Speech Pattern Processing. NATO ASI Series. Berlin: Springer Verlag, 102-111
7. R Kuhn , R D Mori. 1990. A cache-based natural language model for speech reproduction. IEEE Trans on Pattern Analysis and Machine Intelligence, PAM2-12(6), 570-583
8. Daniel Gildea, Thomas Hofmannl. 1999. Topic-based language models using EM1. In : Proceeding of the 6th European Conf on Speech Communication and Technology, Budapest, Hungary: ESCA, 2167-2170
9. Neches R., Fikes R., Finin T., Gruber T., Patil R., Senator T., and Swartout W. R.. 1991. Enabling Technology for Knowledge Sharing. AI Magazine, 12(3) :16~36
10. Gruber, T. R. 1993. Toward principles for the design of ontologies used for knowledge sharing. International Workshop on Formal Ontology, Padova, Italy
11. Uschold M. 1996. Building Ontologies-Towards A Unified Methodology. In expert systems 96
12. Yang Lingpeng, Ji Donghong, TangLi. 2004. Document Re-ranking Based on Automatically Acquired Key Terms in Chinese Information Retrieval. In Proceedings of the COLING'2004, pp. 480-486
13. Kwok, K.L. 1997. Comparing Representation in Chinese Information Retrieval. In Proceeding of the ACM SIGIR-97, pp. 34-4
14. Nie, J.Y., Gao, J., Zhang, J., Zhou, M. 2000. On the Use of Words and N-grams for Chinese Information Retrieval. In Proceedings of the IRAL-2000, pp. 141-148
15. Robertson, S.E. and Walker, S. 2001. Microsoft Cambridge at TREC-9: Filtering track: In Proceeding of the TREC 2000, pages 361-369
16. Salton, G., Buckley, C. Term weighting approaches in automatic text retrieval. Information Processing and Management, 1988, 24(5), pp.513–523
17. Vladimir Oleshchuk, Asle Pedersen. Ontology Based Semantic Similarity Comparison of Documents, 14th International Workshop on Database and Expert Systems Applications, September, 2003, pp.735-738
18. Besancon, R., Rajman, M., Chappelier, J. C. Textual similarities based on a distributional approach, Tenth International Workshop on Database and Expert Systems Applications, 1-3 Sept. 1999, pp.180-184