Description of "HIT Information Retrieval Research Laboratory Synonym Dictionary Expanded Edition"

1. English name

HIT IR-New Lab (Extended)

2. Vocabulary Construction

The first and second editions of Tongyi Cilin have the same vocabulary, with 53,859 entries .

Many words are no longer commonly used and have become so-called rare words.

With reference to multiple electronic dictionary resources and according to the frequency of occurrence of words in the People's Daily corpus, only

The frequency of some words is not less than 3 (the statistical result of a small-scale corpus), and 14,706 rare words and non-common words can be eliminated.

After this processing, there are still 39,099 entries left in the Synonymous Cilin .

For the needs of language processing, a dictionary of this size is obviously insufficient and can be said to be far from enough.

In order to expand the Synonymous Dictionary, our laboratory has used many word-related resources and invested a lot of

The Synonym Thesaurus of Information Retrieval Laboratory of Harbin Institute of Technology was completed with a large Chinese vocabulary.

The final vocabulary contains 77,343 words.

2. Word Classification

The Synonymous Cilin organizes all the entries into a tree-like hierarchy.

The collection is divided into three categories: large, medium and small. There are 12 large categories, 97 medium categories and 1,400 small categories .

There are many words in the category, and these words are divided into several word groups (paragraphs) according to the distance and relevance of the meaning.

The words in each paragraph are further divided into several lines, and the words in the same line either have the same meaning

(Some words have very similar meanings), or the meanings are strongly related. For example, "soybean", "edamame" and "yellow bean"

"Beans" and "tomato" are in the same row; "tomato" and "tomato" are in the same row; "everyone", "everyone", "everyone"

In the same line. In addition, "general", "colonel", "lieutenant" are in the same line, while "hired farmers", "poor farmers", "lower middle

"peasants", "middle peasants", "upper middle peasants", and "rich peasants" are in the same row, and "foreign businessmen", "official businessmen", "sedentary businessmen", and "private

"Shang" is also on the same line. These words have different meanings but are very related. In order to distinguish rows with related meanings from rows with synonyms,

Separate, the dictionary "Tong Yi Ci Lin" adds "* *" as a mark at the left end of the line.

The paragraphs within a subcategory can be considered the fourth level of classification, and the lines within a paragraph can be considered the fifth level of classification.

In this way, the dictionary "Tongyi Cilin" has a five- layer structure, as shown in Figure 1. As the level increases, the meaning of the word

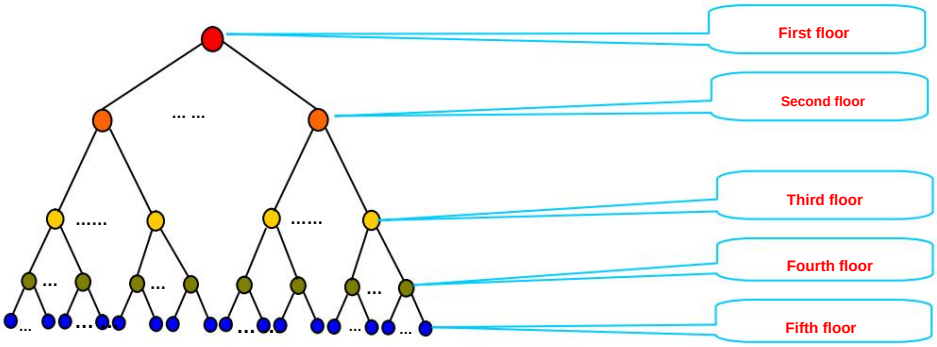
The description becomes more and more detailed. At the fifth level, the number of words in each category is not large, and many categories only have one word.

It is indivisible and can be called atomic word group, atomic class or atomic node.

To provide different services for natural language processing, such as the fourth-level classification and the fifth-level classification in information retrieval

It has been applied in research fields such as search, text classification, and automatic question answering. Studies have shown that the effective expansion of word meanings

The expansion or replacement of keywords with synonyms can significantly improve information retrieval, text classification and automatic question answering systems.



system performance.

The 39,099 words **retained in the dictionary "Tongyi Cilin" also retain the original hierarchical structure.**

The 36,267 **newly added words do not have such a structure.**

The workload of classifying structural systems is huge. Some aspects of classification can be completed automatically by machines.

However, the results of automatic completion are not very ideal, and each link still mainly relies on manual completion.

3. Coding

The Synonymous Cilin only provides three levels of coding, namely, large categories are represented by capital letters, medium categories are represented by small letters, and

Write in English letters, and sub-categories are represented by two-digit decimal integers. For example: "Ae 07 Farmers, herdsmen, fishermen",

"Ae 07" is the code, and "farmers, herders, fishermen" is the title of the category. The title is composed of one or more

The title word is divided into four layers of "paragraph headers (i.e. the first word of each paragraph)".

See Table 1 for the number of fourth-level classes.

Table 1 Dictionary structure example

Ae07 **Farmers, herdsmen, fishermen**
Farmer farmer farmer farmer farmer farmer farmer farmer fellow
Smallholder farmers
tenant farmer
Upper middle peasants and wealthy middle peasants
*** * Vegetable farmers, cotton farmers, tea farmers, tobacco farmers, sugarcane farmers, flower farmers, pesticide farmers, forestry farmers**
Hired peasants Poor peasants Lower-middle peasants Middle peasants Upper-middle peasants Rich peasants
Self-employed farmers Semi-self-employed farmers Collective farmers Members of people's communes

ØFourth level
Ø can be divided into
Do the fifth
Class

Pastoralists, herders, herders

Fisherman Fisherman Fisherman Fisherman

For ease of use, the fourth and fifth level classifications also need to be coded.

The fifth level code and the original three-level code form a complete code, which is the only representative in the dictionary.

The words that appear. For example:

Ba01A02 = **Material quality**

Cb02A01 = **East, South, West, North and South**

Ba01A03@ **Everything**

Cb06E09@ **civil**

Ba01B08# **Solid Liquid Gas Fluid Semi-fluid**

Ba01B10# **ConductorSemiconductorSuperconductor**

The encoding method is described as follows:

The fourth level is represented by capital letters, and the fifth level is represented by a two-digit decimal integer.

The classification results need special explanation, for example, some rows are synonyms, some rows are related words, and some rows have only

One word can be divided into three specific situations. Sometimes it is necessary to distinguish these three situations in use.

Therefore, it is necessary to add more marks to represent the different situations. See Table 2 for the specific marks.

Table 2 Word coding table

Code bit 1		2	3	4	5	6	7	8
Symbol Example 0		a 1 5			B	0 2 = \	# \ @	
Symbolic nature	Major category	Middle category	Subcategory		Word group	Atomic word group		
Level Level 1	Level 2	Level 3	Level 4	Level 5				

The code bits in the table are arranged from left to right. There are three **types of marks for the eighth bit, which are**

"=", "#", "@", "=" means "equal", "synonymous". The "#" at the end means "not equal", "same"

The "@" at the end represents "self-enclosed" or "independent", which is not found in the dictionary.

There are synonyms and no related words.

IV. Improvement of Dictionary

At present, the 1.0 version **of "HIT Information Retrieval Laboratory Synonym Dictionary Extended Edition" has been launched.**

It can meet the application needs in many research fields.

The laboratory will continue to organize manpower to make necessary improvements to the dictionary functions and modify the dictionary.

Errors in classification.

Version 1.0 **adheres to the compilation style of "Tongyi Cilin" and adopts a five-level coding system to provide practical**

The Chinese vocabulary is used to meet the needs of various research fields of natural language.

This lab intends to add more word information, such as part of speech, pronunciation, word frequency, syntactic relationship and linguistic structure.

The addition of this information will greatly change the structure and function of the dictionary, and will also be used in natural language

Processing areas play a greater role.