

HowNet and Its Computation of Meaning

Zhendong Dong
Research Center of Computer
& Language Engineering, CAS
dzd@keenage.com

Qiang Dong
Canada Keentime Inc.
dongqiang@keenage.com

Changling Hao
Canada Keentime Inc.
support@keenage.com

Abstract

The presentation will mainly cover (1) What is HowNet? HowNet is an on-line common-sense knowledgebase unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents. (2) How it functions in the computation of meaning and as a NLP platform? The presentation will show 9 HowNet-based application tools. All of them are not merely demonstration of some methodology or algorithm, but are real application tools that can be tested by users themselves. Apart from the tools that are specially designed to deal with Chinese, most of the tools are bilingual, even the WSD tool.

1 What is HowNet

HowNet is an on-line common-sense knowledgebase unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents. To put it simply, relationship is the soul of HowNet, as well as the world knowledge. The relationships that represent knowledge can be divided into two categories: Concept Relationship (CR) and Attribute Relationship (AR).

It is believed that concept relationships fall into a net, which is called Concept Relation Net (CRN) and attribute relationships fall into a net too, called Attribute Relation Net (ARN). Different individual has different CRN, even of the same concept. This reflects different levels of knowledge among people. CRN is elastic or extendable as it varies with individual persons. The

more knowledge one has, the more concepts he will master, and what is more, the larger or more complicated CRN of the concepts he will know. It can be imagined that a 6-year child may know “doctor” but his CRN of “doctor” would be far from that as shown in Fig. 1, which is believed to be mastered by an ordinary adult. The same case goes with mankind as a whole. Mankind increases his knowledge with each passing year when he enlarges his volume of concepts and at the same time, the CRN of the concepts.

Careful observations find that the meaning of concepts is displayed not only by its CRN but also by the relationships among attributes of the concepts, as called Attribute Relation Net. In many cases it is the attributes of a concept that act in the role of meaning representation. Fig. 2 reveals that it is not “paper” as a whole that is related to “write”, but only one of its attributes, say “color”, is related to “write” with “contrast” as the condition. Therefore in a strict sense, “paper” is not necessarily related to “write”. We can sometimes even write on the sand with a twig or on the table with our wet finger. On the contrary, we cannot write on a piece of white paper with a chalk or on the blackboard in black ink. Therefore, for writing, what affects may not be the whole lot of the concept like “paper”, but some attributes of the concept. Besides, we can use “paper” to wrap up something because of its attributes of the material, which are almost the same as cloth or plastic. HowNet is unique in its four peculiarities: (1) Use of sememes: HowNet uses sememes to interpret concepts. Sememes are regarded as the basic unit of the meaning. (2) Definition in a structuralized language: Each concept in HowNet lexicon is defined in a language, called Knowledge Database Markup Language (KDML). The KDML is mainly composed of sememes and semantic roles. The

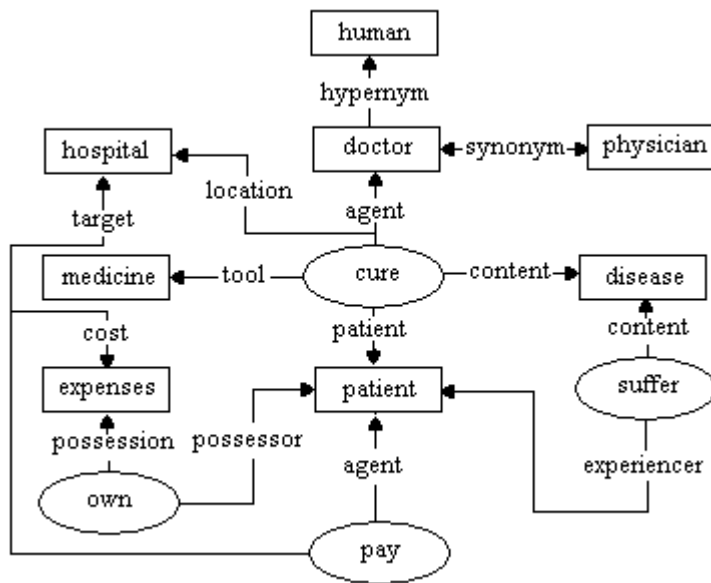


Figure 1 Concept Relation Net (CRN) of “doctor”

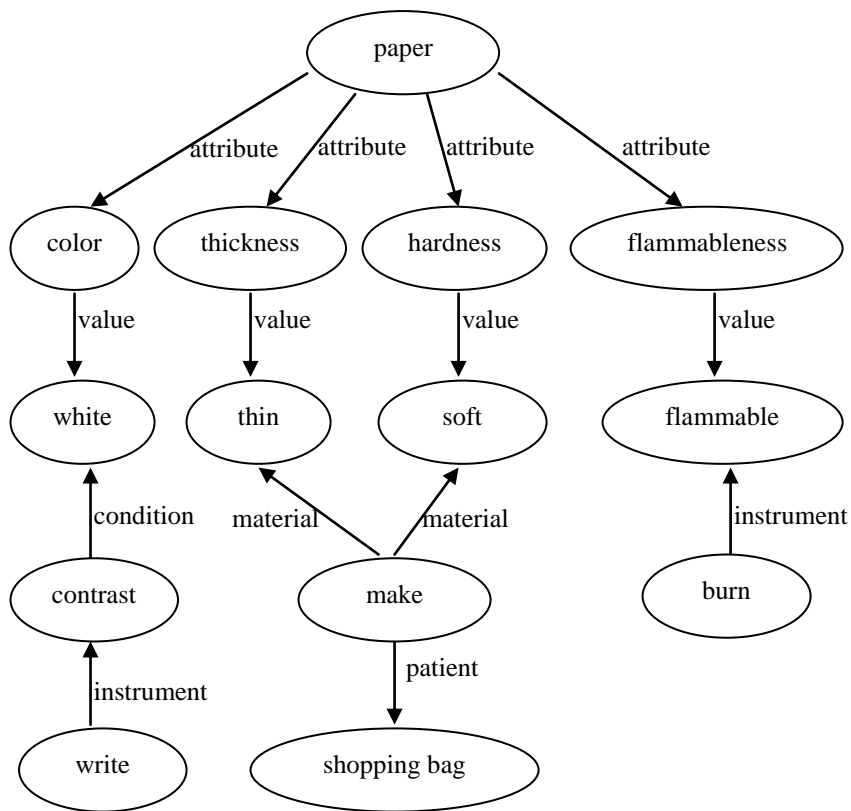


Figure 2 Attribute Relation Net (ARN) of “paper”

Knowledge Database Mark-up Language uses 2089 sememes, 128 secondary features and 94 semantic roles as its vocabulary and adopts an

extended BNF as its syntax. The concept of “doctor (medical)” is defined in HowNet as:
DEF={human|人:HostOf={Occupation|职位},

domain={medical|医},{doctor|医治:agent={~}}

All the computation of meaning in HowNet is based on the definitions of the concepts.

(3) Self-sufficiency: Systematic integration of hierarchical taxonomies, axiomatic inference, KDML-defined concepts.

(4) Language independence: In the final analysis, HowNet is not word-oriented as WordNet, but concept-oriented. Only with the HowNet's shared definitions can we achieve a shared ontology for all languages.

Table 1 shows the latest statistics of the basic data of HowNet.

Chinese Character	7182	
Chinese Word & Expression	100385	
English Word & Expression	96565	
Chinese Meaning	115278	
English Meaning	121262	
Definition	30014	
Record	192191	
Semantics	Chinese	English
Event	14554	12881
Attribute	4351	4879
AttributeValue	10160	10140
Things	72016	72016
Time	2683	2683
Space	1244	1244
Component	8577	8577

Table 1 statistics of the basic data of HowNet

2 HowNet functions as a NLP platform

HowNet is developing toward a NLP platform. HowNet is a powerful tool for the computation of meaning. To date, 9 HowNet-based application tools have been developed. They are:

1. HowNet_Browser (E/C bilingual)
2. HowNet_Relevance (E/C bilingual)
3. HowNet_Similarity (E/C bilingual)
4. HowNet_Inference_Pool (E/C bilingual)
5. HowNet_SenseColonyTester (E/C bilingual)
6. HowNet_Translate (E-to-C)
7. HowNet_Morpho_Processor (Chinese monolingual)
8. HowNet_VN – disambiguator for Chinese V-N structure (Chinese monolingual)
9. HowNet_VXY -- disambiguator for Chinese V-N-的-N structure (Chinese monolingual)

The purpose for developing these tools is (1) to check the HowNet's data and framework for its accuracy and coverage so as to test the soundness of its philosophy and design; (2) to push HowNet near to end applications so as to provide evidence of its value as knowledge resources;

Of all these tools, HowNet Browser is the key. The Browser contains all HowNet basic data and provides various kinds of elementary or shallow computation of meanings. The basic data in HowNet can be divided into two parts: firstly, the basic lexical data and secondly taxonomies. In the lexical database, each concept is described in a fixed structure, for example,

NO.=046048

W_C=富

G_C=adj [fu4]

S_C=PlusSentiment|正面评价

E_C=~人, ~婆, ~国, ~家子弟, ~得流油, 穷的穷~的~, 贫~差别, 先~起来, 农村~了

W_E=rich

G_E=adj

S_E=PlusSentiment|正面评价

E_E=

DEF={rich|富}

RMK=

With the browser the user can retrieve all kinds of basic relations between concepts, such as synonym, hypernym, hyponym, etc. It should be noticed that these kinds of relations in HowNet are not coded manually as the way as done in WordNet, but are computed on the basis of concept definitions. The browser can give all sorts of semantic roles for a given verb concept. To take “treat” as a given event, we retrieve all its “agents”, “locations”, “patients”, “instruments”. This is regarded as the shallow relations between verb concepts and their relevant noun concepts.

Particular attention should be given to our newly developed tool, HowNet Inference Pool (E/C bilingual). With the help of an activator of the tool we can build a senses pool for any concept in HowNet. The pool covers all sorts of relationships under the key concept, for instance, when the concept of “money” as the key, it has a pool with 2600 concepts, including “bank”, “deposit”, “borrow”, “buy”, “steal”, etc. Hence

suppose a question like “can we borrow money from a bank?” is raised to an inference machine, we are sure that the machine can give a correct answer with correct selection of meanings, like “bank” as “financial bank”. Moreover based on the inference machine we have developed a word sense disambiguation tool called HowNet SenseColony Tester (E/C bilingual). The tool is designed to be skilled in tackling the ambiguity of discourse type both in Chinese and English. The words “governor”, “state” in the following paragraph are so-called those of discourse-ambiguity type:

“We provided \$250 in relief to more than 5 million California seniors -- many whose life savings had taken a big hit in the financial crisis. And we provided emergency assistance to our *governors* to prevent teachers and police officers and firefighters from being laid off as a result of *state* budget shortfalls. At a time when California is facing a fiscal crisis, we know that this has saved the jobs of tens of thousands of educators and other needed public servants just in this *state*. And what was true in California was true all across the country.”

The tool is language independent; it employs the data resources and the algorithm of the same type.

HowNet English-Chinese MT system is a rule-based system. It uses HowNet basic data as its English-Chinese bilingual dictionary. It is powerful in its strongly semantic basis. The system will surely have a bright future in its application to PDA products and Chinese language learning aids.

All the HowNet tools are not merely a demo of certain methodology, but are real applications that can be tested by users themselves.

References

- Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, 2005, Extended-HowNet: A Representational Framework for concepts, Proceedings of Second International Joint Conference 2005
- Keh-Jiann Chen, 2009, E-HowNet- a Lexical Semantic Representation System and its Relation to Morphology, Syntax and Semantics, (keynote talk, at ROCLING XXI 2009)
- Zhendong Dong and Qiang Dong, 2006. *HowNet and the Computation of Meaning*, World Scientific Publishing Co. Pte. Ltd., Singapore

Fellbaum, 1998, WordNet: An Electronic Lexical Database. Ed. Cristiane Fellbaum, The MIT Press, Cambridge, London, England, 1998.

Nagao, Makoto, 1997 Machine Translation Through Language Understanding, MT Summit VI Proceedings

Yarowsky, D. (1993) One sense per collocation. In *Proceedings, ARPA Human Language Technology Workshop*, pp. 266-271.

董振东, 董强, 2001, 知网和汉语研究, 当代语言学, 第三卷, 第1期, pp.33-44

冯志伟, 2001, 计算语言学基础, 北京, 商务印书馆.