Context Framework Extracting Based on Short Text

Xu Xiaoqing and Wang Jingzhong

***Abstract*—Base on HNC theory, this paper considers the domain, situation, background of a short text sufficiently and extracts the characters which have more effects on semantic comprehend. Then it fills the frame. Experiments show that the algorithm has better comprehend on short text context and deals with the separation, degradation and different meanings of semantic block effectively.**

***Index Terms*—HNC theory, context framework, domain, situation.**

I. Introduction

Along with the development of the Internet, people are greatly depend on the Internet to inquires the needed information, and about 90% network information is text information, so text information classification and filtering will become the hot issues, the key problem is the text character extraction.

The two ways for processing text are: vector space model (the Vector Space Model) and latent semantic indexing (the Latent Semantic Indexing). In text categorization, some researcher used words and part-of-speech category feature of mutual information extraction. In text filtering field, Oracle use encyclopedia to concepts, by the use of expansion concept of frequency characteristics extracted. Some researcher used word frequency and tagging the main results of feature extraction for content. Some used the language analysis technology of extracting abstract sentence [1].

No matter that the above feature extraction methods have made use of the lexical, syntax and other information, do not take full account of the text, there are a lot of separation, ambiguous sentences and other factors can not be understand. Reference [1] given a framework based on context feature extraction algorithm of text, while it demands more semantic knowledge base, participles coding is relatively complex, syntactic classification rule is not efficiency. This essay is proposed based on the context frame extraction methods, fully considers the text semantic pieces of separation, sentence ambiguity etc. The experiment proved that this method can better understanding the context information, and the methods are efficiently.

Manuscript received June 8, 2013; revise August 16, 2013.

Xu Xiaoqing is with the Computer Department, Beijing Electronic Science and Technology Institute, CO 100070, Beijing China (e-mail: xxqwch@aliyun.com).

Wang Jingzhong is with the Computer Department, North China University of Technology, CO 100144 Beijing, China (e-mail: jingzhongwang@163.com, xixiw@ncut.edu.cn).

II. HNC Theory and Contextual Framework

A. HNC Theory

HNC theory [2] is a concept of Hierarchical Network Concepts, with conceptualization, hierarchic, networked semantic expression is the foundation. HNC theory considers that the semantic block is the sentence semantic composition unit in form, it may be a word, a phrase or a sentence[3]. Semantic block falls into primarily semantic and auxiliary semantic. The primarily semantic block is the indispensable ingredient, the auxiliary semantic block is an additional element. Sentence class represents the type by semantic block structure, it is the basic framework of sentence semantics.

The seven basic sentence HNC class is role, process, transfer, words, sentences, relationship and judge sentence. On the basis of sentence classes, we can form a certain number of mixed sentence types. Basic sentence type and mixed sentence class gives the structure of expression statements, and provides a simple and completed mathematical expressions. For example:

Function sentence: XJ=A+X+B
Zhang San hit anybody's head.
Process sentence: PJ=PB+P
Zhang San's injury is greatly improved.
Transfer sentence: TJ=TA+T+TB+TC
Zhang San's friend told Zhang San parents the news.
Effects of sentence: YJ=YB+Y+YC;YBC+Y
Zhang San' leg is heal.
Relationship sentence: RJ=RB1+R+RB2; RB+R
Zhang San lost his friend for years.
State sentence: SJ=SB+S+SC; SB+S; SB+SC
Zhang San wearing soldier coat./
Judge sentence: DJ=DA+D+DBC
Zhang San think that Li can't do this.

B. Context Framework

Context framework [4] is put forward under the HNC system to the formal structure, there are three aspects in the contextual information abstraction frame: field, situation and background. The three constitute context is the key elements, and is the description of 3d information space. Context framework description text content, based on static characteristics and semantic relations, it can reflect the inner relation of text and can get different granularity framework and different levels of the text features [5].

For example: the Peter uglify the leader.

This sentence context framework for:

1) the field of information: political activities
2) scene information: Peter+uglify+leaders
3) background information: unite

This essay can put the articles fractionation into a sentence series (S1, S2, ... , S*n*), on which each sentence formation

DOI: 10.7763/IJIMT.2013.V4.444

context framework, get context framework sequence: (SemFrame (S1), SemFrame (S2),..., SemFrame (Sn)). Without considering the relationship between sentences context, we can define Text semantic structure: SemFrame (sites) = SemFrame (S1) + SemFrame (S2) +... + SemFrame (Sn).

III. Extraction Algorithm

A. Data Structure

This paper described the contextual frame rules, ignore state or background of excessive description on semantic, retain field, situational behavior corpus, situational behavior receptor, situational behavior, background, time, space, so concise background frame data structure:

```
Class SemFrame
{ public:
   Class  domain；// Subordinate field
   Class  situation；// scene
   Class  background；// background
   ……
}
Class Domain
{ public:
  CString domain；//Subordinate field
  …
}
Class Situation
{ public:
   CString subject;// Scene behavior subject
   CString object；//Scene behavior receptor
   CString action;//Scene behavior, center of verbs
  …
}
Class Background
{ public:
  CString time; // Background of time
  CString scene; //Background of space
  …
}
```

B. Database Design

In order to realize the context of extraction, the framework of each step analysis result will save to database. The data structure is list as follow:

Table Word_character: used to save participle and the part-of-speech tagging of result.

```
Table Word_character
{
   text_id  int；
   sentence_id  int；
   word_id  int；
    word_context  char；
   word_first_character  char；
   word_second_character  char；
   word_third_character  char；
   remark char；
}
```

Table Sentence_type, used to deposit the judgment result.

```
Table Sentence_sort
{
    text_id int；
    sentence_id int；
    sentence_type  int；
    remark  char；
}
```

Table Sem_frame, used to store information, including the framework data items.

```
Table Sem_frame
{
   text_id int；
   sentence_id int；
   domain_first_type  int；
   domain_second_type int;
   domain_context char；
   situation_subject char；
   situation_object char；
   situation_action char；
   background_time  char；
   background_scene  char；
   remark  char；
}
```

C. Extract Context Framework

Context framework extraction is mapping the natural language to the semantic symbols by semantics analyzing, and then the semantic role between semantic relation and the role of sentence group contribution is analyzed, finally context framework is extracted. Specific steps are shown as follows:

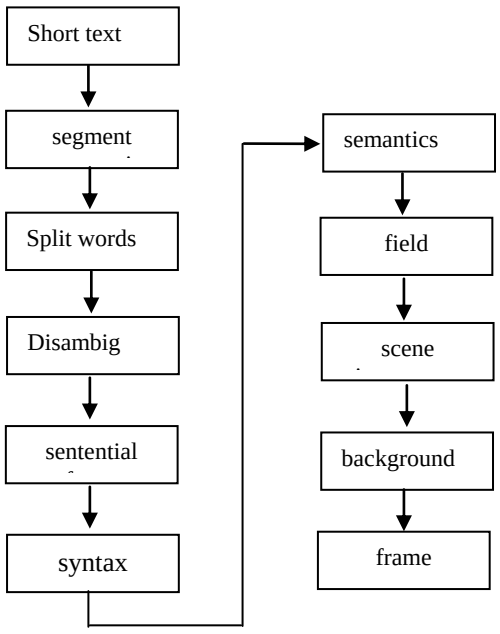

Fig. 1. Technological process.

Step 1 Using punctuation to participle, and then, according to the standard of lexicon and part-of-speech tags, to participle and tag, and the results will save to the database in Word_character.

Step 2 Using corpora trained data dictionary, to process the result of word segmentation, and eliminating the word ambiguity.

Step 3 According to the participle and the HNC rules, analysis each syntactic and sentence type, the result will kept to the database.

Step 4 According to the analysis results, dealing with semantic blocks of separation, sentence exuviate and ambiguity effectively. By analysis the sentence in all, to extract situation and background information, the result will kept to the database, and generate context framework.

D. Special Contextual Situation

1) The semantic pieces of two components in the sentence is usually connected, while there is an exception, this kind of situation is called semantic block of separation. For example: Zhang San interrupted the leg of Li Si, this sentence of "Li Si" and "leg" is the same semantic pieces of two parts, and is separated. In passive sentences, the syntactic can make special treatment and can correct the extraction of the subject and predicate in the sentences.
2) Sentence exuviate [6].The sentence exuviate refers to the sentence reduced to semantic block, also is the semantic block contained in the sentences. For example: governmental and non-governmental economy research institute of all believe that information technology can promote economic growth.

In the syntax classification, the mixing sentence is made a special marking, and it will break down as a basic sentence for syntactic parsing, the sentence exuviate processing is effectively.

3) For the different areas of the essay, constructing the different ambiguity data dictionary, use this dictionary, make the second treatment for the segmentation results, the processing of participle ambiguity is effectively, and the syntactic ambiguity resolution method[7] is used when syntactic analyzing, thus the ambiguity probability is reduced.

IV. Tendency Text Filtering Based on Semantic Framework

A. The Distance Function and the Frame Matching Weight

Because the author writes the text in the different way, one sentence is not put up all of the characteristics, when the feature item vacancy occurs in the process of filling of the sentence context framework, it is need to remote matching, which uses the feature [8] that reflect certain context to fill the framework in the same paragraph even farther within the scope. The distance function is show as:

$$f_{dist}(s)=\begin{cases}1, & \text{In the same sentence}\\ 0.5, & \text{In the sentence near to}\\ 0.25, & \text{In the same paragraph}\\ 0.1, & \text{Other}\end{cases}$$

In order to show the difference between the distance matching and non distance matching in the field, situation and background, the frame matching weight can be determine according to the following formula. The set of feature of the framework is {S1, S2, ... , S*n*, weight}, while Si is used to describe the feature items, weight description the matching weight of the framework. To reflected importance of the key feature matching, according to the experience, Si with different weights:

$$w(S)=\begin{cases}2, & \text{S= subject, receptor of the situational behavior}\\ 3, & \text{S= Situational behavior, background comments}\\ 1, & \text{S= Field, background time, background space}\end{cases}$$

The weight calculation formula of the field situation, background of the framework of *F* is show as follow:

$$weight(F)=\frac{\sum_{i=1}^{n} w(s_i)\, f_{dist}(s_i)}{\sum_{i=1}^{n} w(s_i)}$$

After the characteristics of the frame of the field, situation, background is filled, value, the matching weight can be calculated according to the above formula, the value of weight is given.

For the weight values of the frame of the whole sentence and text, the calculate formula is:

$$weight(F)=\sum_{i=1}^{n} w(s_i)$$

The weight of global semantic framework is calculated, it can effectively eliminate miscarriage caused by referencing the contrary argument, because the weight can be positive and negative offset in some extent.

B. The Calculation of Semantic Similarity

The semantic similarity is the similarity degree between two semantic frames, defined as follow:

Sim (S1, S2) = a*SimDom (S1.Domain, S2.Domain) + b*SimSit (S1.Situation, S2.Situation) + c*SimBac (S1.Background, S2.Background). while, S1 is semantic framework,S1.Doma is information of the field, S1.Situation is the information of Situation, S1.Background is the information of background, SimDom (S1.Domain, S2.Domain) is the similarity of field, SimSit (S1.Situation, S2.Situation) is similarity of Situation, SimBac (S1.Background, S2.Background) is the similarity of background, a, b, c is the regulating factor.

1) The field similarity SimDom (S1.Domain, S2.Domain) is defined as

For highlighting the tendency of the application of the professional field, the definition of each similarity are integers that are greater than or equal to 0. Some paper gives the definition and algorithm of similarity, while these algorithms are not prominent to the tendency of text. In order

to highlight the text tendency, the definition and the algorithm of the similarity for the background, field and situation is improved.

The field similarity SimDom (S1.Domain, S2.Domain) is as follows:

SimDom（S1.D, S2.D）=0
for S2'field S2.Di
for S1'field S1.Dj
if （S2.Di = = S1.Dj） SubSim=2
if （S2.Di ∈ S1.Dj） SubSim=1
if （SubSim> SimDom）SimDom= SubSim

2) The situation similarity SimSit (S1.Situation, S2.Situation) is define as

SimSit (ST1, ST2) = SimSubject （ ST1.subject, ST2.subject）
SimObject (ST1.object, ST2.object) SimAction (ST1.action, ST2. action)

Only when the subject, receptor and behavior of the scene are meet, situational similarity can be greater than or equal to 0, if there is no contextual information (when the semantic analysis on failure), then SimSit=1. The Semantic is composed of HNC semantic Subject|Object|Action, only when the text Subject and Object are agree, the text Action can be computed, and the text similarity of scene is 0. For the characteristics C, the relationship between Subject, Receptors and Behavior is show as follow:

if C1≌C2 SimSit（C1,C2）=2
if C1∽C2 SimSit（C1,C2）=1
if C1∈C2 SimSit（C1,C2）=1

≌ expressed equality of the features, ∽approximate of the features, ∈is inclusion relation.

3) The background similarity SimBac (S1.Background, S2.Background) is define as

SimBac （BK1, BK2）=SimBac（BK1.time, BK2. time）+b*SimBac(BK1.scene,BK2. scene)+c* SimBac（BK1.object1. polarity，BK2. object1. polarity）* SimBac (BK1.object2. polarity, BK2. object2. polarity）*…*SimBac (BK1. objectn. polarity，BK2. objectn. polarity）)

The a, b, c is the regulator, used to adjust the tendency of the characteristics, under the condition that the concept of time and space of the background is not clear, to highlight the background tendency. If there is no background information, then SimBac =1.

1) When calculating the similarity of time and space, it is to be compared in their scope. If TL represents the time or space, then the algorithm is:

if TL1=TL2 SimBac（TL1,TL2）=1
if TL1∈TL2 SimBac（TL1,TL2）=1
otherwise SimBac（TL1,TL2）=0

2) For a particular object the similarity is defined as:

SimBac (positive, positive) = 1, SimBac (negative, negative) =1, SimBac (positive, negative) =0, SimBac (negative, positive）=0, Before treatment, the relationship of position must be figure out.

V. Testing and Analyzing

Selecting five types of testing text on the Internet, this essay involved field, quantity. The article includes separating, sentence exuviate, ambiguity and other special context.

The text is: To build a socialist harmonious society is a great strategic task. After long-term efforts, we already have a harmonious socialist society of favorable conditions and see in the historical opportunity of harmonious society construction.

Participle and part-of-speech tagging: Constructing socialist harmonious society/n/n is/v a/m item/q major/a strategic task/n/n. / w after/p long-term/b/w/an effort, we/r has/d have a/v/v/u constructing socialist harmonious society/n/u various/r favorable/a condition/n, / w/v/u see in the history/n opportunity/n. /w of/u harmonious society/n / construction/n.

Syntactic analysis: First statement: role sentence: XJ = X + B, state sentence: SB + SC.Second sentence: state sentence: SJ = SB + S + SC

Context framework: First statement: field information: political + political activities. scene information: to construct a socialist harmonious society + is + task. background information: no. second sentence: field information: political + political activities. scene information : we +have + conditions, we + see in + opportunities.

For the testing, 5 fields, 20 training and 100 text sample for every fields are selected, show as Table I. The testing results show that the accuracy of frame extraction is 98%, and the semantic block separation, sentence exuviate and ambiguity is treated effectively.

TABLE I: The Testing Corpus

Class	Field	The Number of Training	The Number of Testing
A	The political system and policy	20	100
B	Diplomatic activities	20	100
C	banking business	20	100
D	Economic and technology	20	100
E	Economy and culture	20	100

VI. Conclusion

Based on the HNC, this essay realized the extraction algorithm for context frame, the algorithm use the context information, treat the semantic pieces of separation, sentence exuviate and ambiguity effectively. In some lengthy text, the dispersing feature caused the false match, while, it is processed effectively by semantic analysis. The framework in the sentence can distinguish various characteristics from others, and it can represent the text semantic. Experiments show that for some field the framework can be used to understand semantic, and it has a higher efficiency.

Acknowledgment

This work was supported in part by National Science and Technology to Support Key Projects (2009BA171B02), Natural Science Foundation of BeiJing (KZ2010009008）and PHR (IHLB).

REFERENCES

[1] Y. H. Jin and C. J. Miao, "A framework based on the text context feature extraction algorithm," *Computer Research and Development*, April 2004.

[2] C. J. Miao, "HNC sentence based on semantics of the sentence type system," *Applied Linguistics*, Feb. 2006.

[3] Z. Y. Liu, "A sentence-level semantic annotated corpus based on HNC theory," in *Proc. 2011 International Conference on Asian Language Processing*, 2011.

[4] Y. L. Chi, "A general shape context framework for object identification," *Computer Vision and Image Understanding*, vol. 112, Dec. 2008.

[5] J. Wang, "Chinese relation extraction using web features and HNC theory, *Journal of Information and Computational Science*, vol. 9, Nov. 2012.

[6] H. Markus, "A context aware information quality framework," in *Proc. 4th International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, 2009.

[7] Z. Y. Liu, "The research of sentence testing based on HNC analysis system of sentence category," in *Proc. 2010 International Conference on Asian Language Processing*, 2010.

[8] C. B. Wu, "Text clustering based on combined features of concepts and words," *Journal of Information and Computational Science*, vol. 9, Dec. 2012.

Xu Xiaoqing was born in 1961. She got bachelor's degree from Inner Mongolia University, P.R.China, being special in Computer Science, department of Computer Science, Adjunct Professor of Computer Science She is now the Adjunct Professor of Computer Science of the Beijing Electronic Science and Technology Institute Beijing, Main research field is operating system and security. Taking the Beijing Natural Science Foundation project "the evaluation and prediction of network behavior based on operation semantic rules" and other topics. Having published the books, such as The Technology of Computer Communicating Information Security Etc. Having published the thesis such as "The Application of Zerotree Wavelet on Image Watermarking" CiSE 2009.

Wang Jingzhong was born in 1962. He got master's degree from Inner Mongolia University, P.R.China, being special in Communication and Electronic System. He is now a professor of North China University of Technology. His main research field are computer network, computer information security, digital image processing. He has achieved many research works, such as The Failure Diagnosis of Equipment of Nonferrous Metals by the Technology of Infrared Imaging, The Study on Fire ware Based on Context Filter, and so on. He has published two books, which are Computer Network and The Technology of Computer Communicating Information Security. He has published thesis more than 80 pieces.