Semantic Similarity Calculation of Chinese Word

Liqiang Pan, Pu Zhang, Anping Xiong
College of computer science and technology
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*— This paper puts forward a two layers computing method to calculate semantic similarity of Chinese word. Firstly, using Latent Dirichlet Allocation (LDA) subject model to generate subject spatial domain. Then mapping word into topic space and forming topic distribution which is used to calculate semantic similarity of word(the first layer computing). Finally, using semantic dictionary"HowNet" to deeply excavate semantic similarity of word(the second layer computing). This method not only overcomes the problem that it's not specific enough merely using LDA to calculate semantic similarity of word, but also solves the problems such as new words(haven't been added in dictionary) and without considering specific context when calculating semantic similarity based on semantic dictionary "HowNet". By experimental comparison, this thesis proves feasibility,availability and advantages of the calculation method.**

Keywords*— *semantic similarity; LDA; subject model; HowNet

I. Introduction

The semantic similarity calculation methods of word have been widely used in question-answering system, information retrieval, machine translation, etc. Different application Background have different definition of semantic similarity. In question-answering system and information retrieval, semantic similarity of word mainly focuses on the approximate degree of synonymity or same-meaning.While in machine translation it focuses on the approximate degree of mutual substitution in different contexts. The application background of this paper is Chinese question-answering system.So the understanding of word semantic similarity is approximate degree of synonymity of two words without caring about contexts. Semantic similarity of two words is higher if they are more synonymity in different contexts, otherwise the similarity is lower.

There are mainly two semantic similarity computing methods of word[1]. One is counting word information in documents, the other is constructing knowledge of "world". The first method, using statistical information of word to calculate word semantic similarity, is based on aggregation phenomenon of the analogue. The method is objective and specific, so it can reflect similarity and difference of word in syntactic, semantic, pragmatic, etc. However, the method is dependent on training corpus and counting algorithm. In addition, this method is easily interfered by data sparsity and noise. Sometimes there are some obvious errors. For example, using LDA(Latent Dirichlet Allocation) subject model[2] to generate distribution of subject-word and document-subject. Words are aggregated according to topics, so words in the same topic have semantic similarity.The second method, using knowledge of "world", is based on the fact that everything is interrelated. Generally it describes the characteristic of word and relation of word using special description-language and building a structure like dictionary. For example semantic dictionary "HowNet" describes the connections of word through relationship of "sememe" and reflects synonymity of word through the approximate degree of similarity of sememe [1].The method accurately reflects semantic similarities and differences of word, but the result obtained by this method is greatly influenced by subjective consciousness. From the perspective of development of things, construction dictionary can't be completed and can't keep pace with the times, thus it can not accurately reflect objective facts.

Above all,The two kinds of semantic similarity computing methods both have advantages and disadvantages. The thesis puts forward a new semantic similarity computing method (two layers computing method)by combining the two methods and redefining similarity calculation method of word. Firstly, The method uses LDA subject model to excavate topic-word distribution.Using LDA topic model reflects the objective existence of word. Then thesis uses semantic dictionary "HowNet" to further excavate the semantic similarity of word which reflects the objective substantiality of word. The new method lays foundation for similarity judgment of question sentence in Chinese question-answering system.

II. The first layer semantic similarity calculation

A. Problem description

Sentence C1:What is the fastest search engine in search field?

Sentence C2: In Chinese retrieval,Baidu is more efficient than Google.

We can see that there are no common words between C1 and C2, but they are still similar. The reason is that Google and Baidu are two specific examples of Search engine. In fact, we often encounter those problems such as correlation and similarity of word and sentence in Search engine algorithm and question-answering system. In traditional information retrieval field, there have been a lot of methods to measure sentence similarity, such as the classical VSM model. However, those methods are often based on a assumption that the more repetition of words between sentences, the more similar they are. Through the example above, we can see that it does not conform to the reality. Most of the time, the approximate degree of synonymity of sentences depends on semantic relations behind words rather than repetition of words, especially suitable for short texts and questions with few words. Therefore, we need to adopt LDA topic model to find subject distribution behind words and judge semantic similarity of word.

B. Brief introduction of Latent Dirichlet Allocation subject model

LDA subject model, proposed by Blei and etc, is a three layers Bayesian generative model—text-topic-word [2]. The essence of LDA is to find topic structure of text using feature of words co-occurrence in text.In generation process, each text is represented as mixture distribution of subjects, and each subject is a probability distribution over words. Based on pLSA[4], leading a hyper-parameter α into the model's document-topic probability distribution, thus the new model obeys Dirichlet distribution. Then Griffiths and etc apply Dirichlet prior distribution to another parameter β , which makes the LDA subject model come into being a completed model. The model is represented by Fig. 1, with the meanings of symbols shown in table 1.

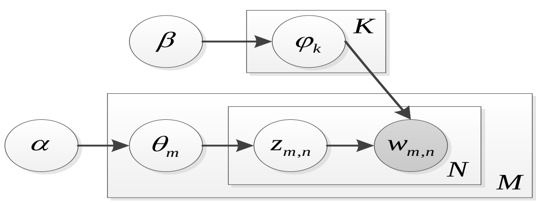

Fig.1. LDA probability graph model

TABLE I. SYMBOL IN LDA MODEL

Symbol	Meaning	Symbol	Meaning
α	Hyper-parameter of θ	$w_{m,n}$	word
β	Hyper-parameter of φ	M	Text No.
θ_m	Text-topic probability distribution	N	word No.
φ_k	Topic-word probability distribution	K	Topic No.
$z_{m,n}$	Distribution of words in a topic		

According to Fig. 1,the Joint probability distribution of LDA is:

$$p(w_m, z_m, \theta_m, \Phi \mid \alpha, \beta) = \prod_{n=1}^{N} p(w_{m,n} \mid \varphi_{z_{m,n}}) p(z_{m,n} \mid \theta_m) p(\theta_m \mid \alpha) \bullet p(\Phi \mid \beta) \quad (1)$$

We often set Hyper-parameter $\alpha = 50/K, \beta = 0.1$,K is number of topics. Seeing [4] for detailed information of choosing α and β values.

we can estimate the parameters using:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{N} n_k^{(t)} + \beta_t}$$
$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \quad (2)$$

Where $n_k^{(t)}$ denotes to the number of times that word t has been observed with topic k, $n_m^{(k)}$ denotes to the number of times that topic k has been observed with a word of document m. If you want more detailed information,you can see the paper of Blei [4].

C. Semantic similarity Calculation method of word in subject spatial domain

Running LDA topic model and doing Gibbs sampling on the document corpus D, we get K topics hidden in the documents and topic-word probability distribution Φ . The element $\varphi_{s_k w_t}$ of Φ shows the probability of word w_t belongs to topic s_k (1 ≤k≤ K).

K Topics build a feature space: $V = (s_1, s_2, s_3, \cdots, s_k)$ (3) So the word w_1 and w_2 distribution vector in K topics feature space is:

$$V_{w_1} = (\varphi_{s_1 w_1}, \varphi_{s_2 w_1}, \varphi_{s_3 w_1}, \cdots, \varphi_{s_k w_1})$$
$$V_{w_2} = (\varphi_{s_1 w_2}, \varphi_{s_2 w_2}, \varphi_{s_3 w_2}, \cdots, \varphi_{s_k w_2})$$

The semantic similarity calculation of two words w_1 and w_2 is:

$$Sim(w_1, w_2) = \cos ine(V_{w_1}, V_{w_2}) = \frac{V_{w_1} \bullet V_{w_2}}{|V_{w_1}||V_{w_2}|} \quad (4)$$

The value of (4) is higher, the similarity of two words w_1 , w_2 is more approximate , vice versa.

III. THE SECOND LAYER SEMANTIC SIMILARITY CALCULATION

A. Problem description

Sentence C1:What is the fastest search engine in search field?

Sentence C2:In Chinese retrieval, Baidu is more efficient than Google.

Sentence C3:The search result on Google is more accurate than on Baidu.

By constructing topic spacial, we find that Search engine, Google and Baidu have semantic similarity by calculating their subject distribution cosine (4). Concluding that C1 has similarity with C2 and C3. But after doing further analysis, we find that C1 describes search speed, C2 describes efficiency of retrieval, and C3 describes search accuracy. In other words, searching C1 on Search Engine, we expect that the feedback is more about performance information of search engine or not. So we need further judge synonymity of other words. As we all know, there have synonymity among speed, efficiency and accuracy, but the semantic similarity between speed and efficiency is higher than between speed and accuracy. Of course, we also see that the topic spatial domain created by LDA topic model can judge the correlation between words through calculating their topic distribution cosine (4), but for further specific semantic information of words can not be presented. In order to make up this shortcoming, we use the following method based on semantic dictionary "HowNet" to analyze specific semantic similarity between words.

B. *Brief introduction of "HowNet"*

"HowNet" is a common sense knowledge bases, of which description objects are concepts and semantic items, and can describe Chinese and English word using description objects [1]. Using the basic content of "HowNet" to compute the relationship of words or phrase. As the meaning of Chinese words are very complex, its semantic meanings are different in different contexts. So one word are described as the collection of several semantic items and concepts in "HowNet". "HowNet" use "sememe" to future describe semantic items. Special word "sememe" is the smallest unit of semantic meaning and does not vary with the contexts.

Sememes are the most basic unit of describing the meaning item and exiting complicated relations[1]. In "HowNet", there are eight relations of sememe: hyponymy, synonymy, relative, antonymy, part-whole, attribute-host, event-role, materials-production. Hyponymy is the most important sememe relation. It is a kind of hierarchy system, which is described through tree structure which is easy to operate by computer. The top describe abstract concepts and the bottom describe specific concepts. As follows, we will use the hyponymy relation of sememe to compute semantic similarity of words. If you want more concreteness calculation, you can take other relations of sememe into account .

C. *Similarity computing method of word based on "HowNet"*

There are two Chinese words: w_1 and w_2 . Assume w_1 has n semantic items, w_2 has m semantic items. And the similarity of w_1 and w_2 is the biggest similarity of their semantic items.

Thus, the similarity between two words is transformed into the similarity of two semantic items. Of course, the specific context of two words is not considered here. Actually it is best to use sentence context to disambiguate words first.In other words,designating the word for a particular semantic item.then computing similarity of corresponding semantic items, which is more accurate and will be further researched in future.

By observing semantic dictionary "HowNet", Finding that semantic items are divided into function semantic items and notional semantic items. So the description of semantic items is different with different classes in "HowNet". Function semantic item is described in {relation sememe} or {syntactic sememe}. So, function semantic item only needs to compute the similarity of corresponding relation sememe or syntactic sememe. However, descriptions of notional semantic item are more complex and are divided into four parts:

1) The first independent sememe Description: The first sememe of independent sememes (without special symbols or relation symbol in front of sememe).

2) Other independent sememes Description: Specific words and Independent sememes except the first sememe.

3) The relation sememe Description: Sememe Described in relation symbol.

4) The symbol sememe Description: Sememe Described in special symbol.

So, Notional semantic items S_1 , S_2 similarity calculation are divided into four parts and each parts similarity marked as $Sim_i(1<=i<=4)$. Different parts have different weight β_i . The first part present the main semantic of word, so it have the highest weight. In order to lower the weight of other parts. The calculation formula is as follows :

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \qquad (5)$$

In(5),the $\beta_1+\beta_2+\beta_3+\beta_4=1$ and $\beta_1\geq\beta_2\geq\beta_3\geq\beta_4$ .reflecting the latter parts have lower significance to the overall similarity.You can adjustable the parameters β_i .

In computing similarity between function word and notional word, we know that the possibility of same semantic they both express is very small in actual application. So we think the similarity of function word and notional word is always zero in the thesis.

Finally, all of similarity calculation of semantic items are ultimately attributed to similarity calculation of sememe.We use the hyponymy relation of sememe to compute semantic similarity of sememe. Obtained by experimental analysis:

$$Sim(p_1,p_2)=\frac{\alpha}{Distance(p_1,p_2)+\alpha} \qquad (6)$$

p_i present sememe, $Distance(p_1,p_2)$ is the path length of p_1 , p_2 in hierarchy tree. α is a parameter can be adjusted according to the practical application. know more information about "HowNet" [1].

IV. THE TWO LAYERS SEMANTIC SIMILARITY CALCULATION METHOD

The similarity calculation method of word based on LDA subject model Sim_1 embodies characteristic of words co-occurrence. The similarity calculation method of word based on semantic dictionary "HowNet" Sim_2 reflects the semantic connection of words.We combine the two algorithms to acquire a two layers semantic similarity calculation method Sim . If the words have similar subject distribution and semantic connection, the similarity of words should be high, Vice versa.

Computing similarity of words w_1 and w_2 use:

$$Sim(w_1,w_2)=\gamma_1 Sim_1(w_1,w_2)+\gamma_2 Sim_2(w_1,w_2) \qquad (7)$$

The γ_1 and γ_2 can be adjusted according to actual application.

V. EXPERIMENTS AND RESULTS

A. Preparations of Latent Dirichlet Allocation subject model

- Experimental data

document number M Using the complete version of Chinese text classified corpus of Sougou laboratory(107M) , The text sets have 10 categories, including automobile, finance, IT, health, sports, tourism, education, employment, culture , military(Each category has 8000 pieces, 80000 pieces of document in total).You can get this data sets from [8].

- Experimental setup

Preprocess Do preprocessing, word segmentation, erasing stop-word to original documents.Algorithm of Chinese word segmentation adopts ICTCLAS segmentation system of Chinese Academy of Sciences. Algorithm of delete stop-word adopts conventional removal method at the beginning and then repeatedly observing generating data, writing regular expressions to remove some words(for example name entities and no specific meaning words such as time, place) again. Erasing stop-word can lower the spatial dimension of word which is useful for computing semantic similarity of words. The final word dimension is 207499(N word number). As we know, Chinese word is a combination structure with single characters and the combination method is very complex. It leads to very high word dimension. Reducing word dimension should be further study features of Chinese words formation.

Topic number K Abstract 20000 documents from M (each categories have 2000 documents) to acquire the most suitable topic number. By observing perplexity-index to determine number of topic . The perplexity-index represents uncertainty when forecasting data. The lower value, the better performance. The calculation formula is as follows[10]:

$$Perplexity\,(D) = \exp\left\{ - \frac{\sum_{m=1}^{M} \log p(w_m)}{\sum_{m=1}^{M} N_m} \right\} \qquad (8)$$

In (8), N_m denotes the length of document m , M denotes the documents sets. $p(w_m)$ denotes the possibility of word w in document m creating by LDA and It's calculation method as follows:

$$p(w_m) = \sum_{d} \prod_{n=1}^{N} \sum_{j=1}^{K} p(w_i \mid z_i = j) p(z_i = j \mid w_m) p(d)$$

(9)

Three experiments are made to set subject number. Each experiment as 10-100 (interval 10 add). The Fig.2 shows that topic number and perplexity-index present inverse relation. When topic number is about 97, Decline trend of perplexity-index is not obvious.Bigger topic number is, Calculation of LDA subject model's parameters estimating is more complicated, so setting K=100.

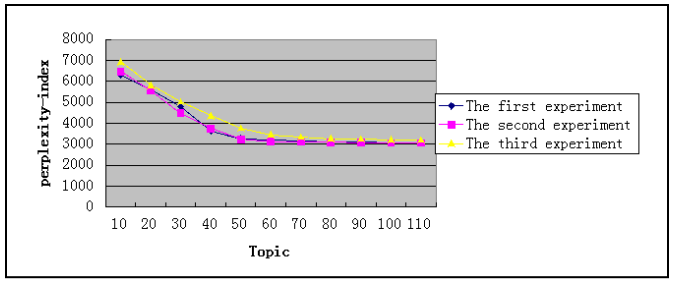

Fig.2. The relation of topic number and perplexity-index

Other parameters $\alpha = 50 / K, \beta = 0.1$.

B. Preparations of semantic dictionary "HowNet"

Data sets Quoting two data sets sorted out by Liuqun (gloss.dat 、 whole.dat). Gloss.dat stores description of semantic item of words(66142 records in total). Whole.dat stores hierarchy relation of sememe(1618 records in total). As gloss.dat data is massive and access frequency is high, gloss.dat is stored into mysql database. It makes search more faster.

Parameter settings

$$\alpha = 1.6, \beta_1 = 0.5, \beta_2 = 0.2, \beta_3 = 0.17, \beta_4 = 0.13$$

C. Preparation of computing method on two layers

Parameter setting Set γ_1 and γ_2 as 0.5, you can change the value according application.

D. Experimental

Table2 shows result of three semantic similarity computing methods of Chinese word. We choose seven groups word, detail information seeing experimental result.

Method 1: Based on LDA subject model described in the thesis

Method 2:Word's semantic similarity computation Based on the HowNet by Qun.Liu[1].

Method 3:The two layers Semantic similarity calculation method

TABLE II. THREE SEMANTIC SIMILARITY COMPUTING METHODS OF CHINESE WORD

No.	Phrases 1	Phrases 2	Method 1	Method 2	Method 3
1	Search engine	Google	0.999994	0.000000	0.499997
	Search engine	Baidu	0.999999	0.000000	0.499995
	Google	Baidu	0.999986	0.000000	0.499993
2	Speed	Efficiency	0.304053	0.557143	0.430598
	Speed	Accuracy	0.132498	0.557143	0.344821
	Efficiency	Accuracy	0.183966	0.588889	0.386428
	Patient	sick person	0.989468	0.500000	0.744734

3	Patient	Doctor	0.760043	0.588889	0.674466
	Patient	Disease	0.818214	0.093023	0.455619
4	Red	Pink	0.935126	0.700000	0.817563
	Red	Light red	0.942354	0.700000	0.821177
	Red	Blood red	0.039634	0.700000	0.369847
5	Like	Love	0.640461	0.500000	0.570321
	Like	Hobby	0.627409	0.500000	0.563704
	Like	Hate	0.469384	0.142870	0.306121
6	Strike	Attack	0.615560	0.500000	0.557780
	Strike	Assault	0.574370	0.500000	0.537185
	Strike	Fondle	0.027581	0.222222	0.111249
7	Apple	Computer	0.984101	0.093023	0.538562
	Apple	Jobs	0.988315	0.000000	0.494157
	Compute	HP	0.990974	0.000000	0.495487
	Compute	Google	0.039012	0.000000	0.019506
	Compute	Keyboard	0.949653	0.083333	0.516493
	Compute	Main engine	0.762508	0.222222	0.492365

E. Analysis experimental results

Words in group 1 and 2 are keywords extracted from previous examples.

From group one, we find that those new specific words(Search engine, Baidu and Google) are not included into the semantic dictionary "HowNet". So we cannot use the semantic dictionary "HowNet" to calculate their semantic similarity. The result from group one embodies the "limitations" of application scope of "HowNet". However, LDA topic model uses statistical approach (training from large scale corpus, then generating potential theme and assembling words according to their subject distribution) property to break through the limitations of new word. Therefore, as long as training corpus is wide enough and updated, the application field of LDA subject model can be extended without limit. The extensibility of LDA subject model can make up the limitation of "HowNet " very well.

In group two, our purpose is to find the most similar word to "speed" from "efficiency" and "accuracy". We know that speed reveal the degree of fast or slow, and accuracy refers to the degree of precision or recall rate in search field, while efficiency is a comprehensive noun which can express both speed and accuracy. Through experimental data, we can see that if we only use the semantic dictionary "HowNet", speed, efficiency, accuracy are consistent in similarity, without any differences. However, the calculation method of semantic similarity on two layers can reflect the differences between words very well.

Analysis the third group of phrases,we want to find the highest similarity with "patient" from "sick person", "doctors", "diseases".From experiment result, Only using LDA to compute phrases similarity, we will find that doctors,sick person and disease both have high similarity with patient. This method has some certain distinction, but can not reach the aim of our application. Because Our application is ultimately used in Chinese question answering system, and the feature of our phrase similarity is synonymity. LDA topic model guarantee phrases similarity difference by sampling and complicated calculation, but is also a probabilistic model which reflect word co-occurrence. As we know, word patient frequently appear in a document,which is very likely to be have sick person, doctor, diseases and etc. That is the reason sick person, doctor, diseases have high similarity with patient when only using LDA to compute similarity. Therefore, in order to further distinguish difference of phrases semantic, we use the semantic dictionary to further mining phrase semantic. As the table shows that the two layer of the semantic similarity calculation method can reflect the greatest similarity(patient and sick person). At the same time, distinguishing doctor and disease with patient. Introducing the semantic dictionary to refine similarity of phrases.

In group four, the keywords are about colors. In semantic dictionary "HowNet", semantic items do not reflect approximation degree of color attribute value (red and pink are similar, while red and white look much different). It's very difficult to describe the approximation degree of colors using objective language, but it is able to tell differences to some extent by the method in this thesis. But the effect will be not consistency using different training corpus.

Words in group five are about affection tendency. Words in group six are about the degree of action. Words in group seven are computer vocabulary.Through analysis of the experimental data, the method in this thesis is also able to distinguish similarity between phrases to some extent, showing the most intuitive feeling and proving the feasibility of the method.

VI. CONCLUSION

The paper presents a two layers semantic similarity calculation method to excavate semantic similarity of Chinese words. Through lots of experiments, this method is feasible and applicable.

ACKNOWLEDGMENT

This work is supported by National Social Science Foundation Project of P.R. China (No.14BFX156).

REFERENCES

[1] Q.Liu and S.J.li, "Word's semantic similarity computation Based on the HowNet", The 3rd Chinese lexical and semantic proseminar, Taipei, China, 2002.
[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research. 3:993–1022, January 2003.
[3] B.Ge,F.F.Li,S.L.Guo, "Word's semantic similarity computation method based on HowNet", Application Reserarch of Computers, Vol.27, No.9, pp.3329-3333, Sep.2010.
[4] T. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation", Technical Report, 2003.
[5] T.Griffiths and M.Steyvers, "Finding scientific topics", Proc of the National Academy of Sciences, 2004.
[6] Hu, Feng Song, Guo, Yong, "An improved algorithm of word similarity computation based on HowNet", Computer Science and Automation Engineering, IEEE International Conference, Vol.3, May 2012.
[7] Z.Dong and Q.Dong,HowNet,http://www.keenage.com.
[8] Text categorization corpus of sougou. http://www.sogou.com/labs/dl