Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2014, *8*, 739-744

Open Access

Uyghur-Chinese Translation Disambiguation Method Research Based on Knowledge Automatic-Acquisition

Ren Ge[1], Yang Yong[1,*] and Xu Chun[2]

[1]*College of Computer Science, Xinjiang Normal University, Urumqi, 830054, China;* [2]*College of Computer Science and Engineering, Xinjiang University of Finance and Economics, Urumqi, 830012, China*

Abstract: This thesis studies the disambiguation method in Uyghur-Chinese translation, and proposes the design philosophy of automatic-acquisition in translation label library aiming at the deficiency of disambiguation corpus in Uyghur. It refers to the existing Uyghur-Chinese bilingual dictionary, Chinese corpus and the Internet, and acquires the corresponding Chinese translation label examples to Uyghur automatically. On this basis, this thesis designs a Uyghur-Chinese disambiguation model, and refines the meaning probability definition based on HMM model. It increases the model's backwards-dependence, and the model's ability to obtain contextual information. To use the model to learn disambiguation knowledge from translation label corpus can eliminate the ambiguity in new Uyghur ambiguous words. From the experiment result, it can prove that the Uyghur-Chinese translation disambiguation framework based on automatically acquired corpus is effective, and increases with the increasing of the scale of translation label corpus, degree of accuracy of the disambiguation result also increases accordingly.

Keywords: Knowledge automatic-acquisition, uyghur-chinese translation disambiguation, word sense disambiguation.

1. INTRODUCTION

Guided translation disambiguation needs large quantities of high-quality translation label corpus. Since it needs a large amount of work to manually establish translation label corpus, and the establishment period is long, there has always been the bottleneck problem of knowledge acquisition for guided translation disambiguation research. In order to solve the establishment difficulty of large-scale label corpus, it has become a heated research method to adopt automatic translation label case and design translation disambiguation algorithm on the basis.

Integrating the previous researches [1-10], a lot of literatures have reported label corpus automatic achieving study aiming at word sense disambiguation in a monolingual. Different from tasks of word sense disambiguation, Uyghur translation disambiguation is to solve the correct translation of Uyghur ambiguous terms in target language in nature, *i.e.*, to find the most suitable Chinese translation for Uyghur ambiguous terms according to the context. If establishing a Chinese translation ambiguous label corpus that includes all meanings of Uyghur ambiguous terms, by means of the existing machine learning algorithm, translation disambiguation would turn to the classic classification problem. In order to overcome the knowledge-achieving bottleneck problem in translation disambiguation research, this thesis will study the automatic-acquisition of translation label corpus. It proposes automatic establishment algorithm of Uyghur polysemy's corresponding Chinese translation label corpus. On this basis, it uses the improved HMM model to establish translation disambiguation framework, and effectively completes disambiguation work of Uyghur translation.

Section Two of this Chapter explains the translation label library's automatic-acquisition design philosophy, and designs specific realization algorithm. Section Three puts forward the translation disambiguation framework of the improved HMM model, including the improved training method, improved HMM model expression, and improved Viterbi algorithm and so on. Section Four describes the experiment process, and description of the result. Section Five is the conclusion.

2. AUTOMATIC CONSTRUCTION ALGORITHM OF CHINESE TRANSLATION LIBRARY

Ambiguous terms include several meanings. For Uyghur-Chinese translation disambiguation, we can use Uyghur-Chinese bilingual dictionary to achieve the corresponding Chinese translation of each meaning of Uyghur ambiguous terms. Using these Chinese translations and searching for the existing Chinese monolingual corpus or Web can obtain examples of Chinese translation labels. Integrating all achieved Chinese translation label examples, translation label corpus is established. On this basis, guided translation disambiguation model can be designed and we can learn translation disambiguation knowledge from label corpus and conduct translation disambiguation on new Uyghur ambiguous terms, and this design philosophy is shown as Fig. (**1**). Since this method conduct disambiguation by different meaning division in different languages, if ambiguity in the source language is retained in the target language, *i.e.*, different meanings of ambiguous words in the source language is translated in the same way in the target language, we cannot conduct

1874-110X/14

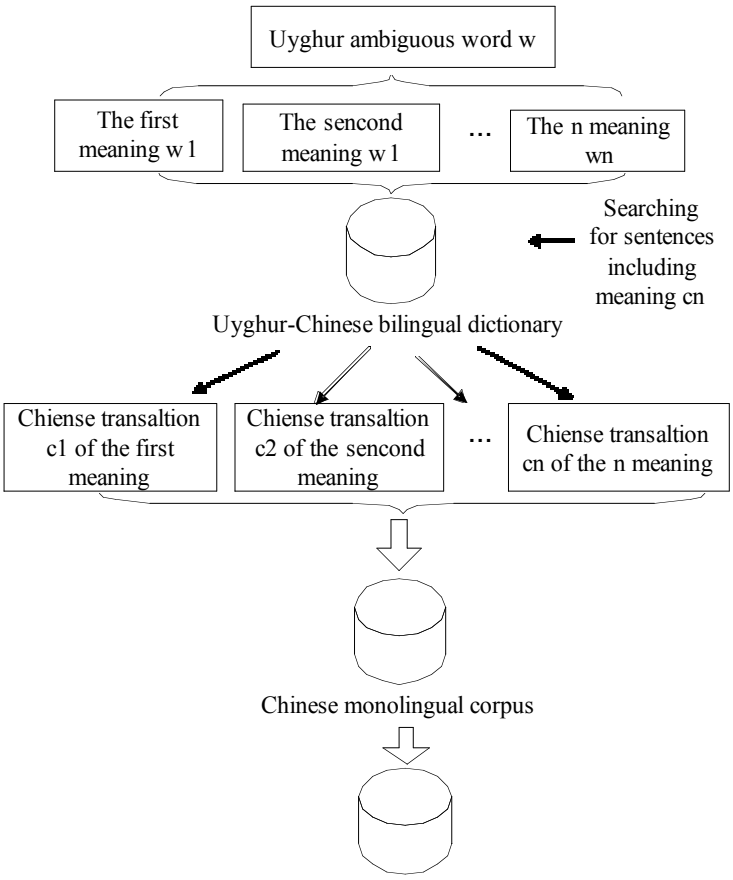

Fig. (1). Design philosophy of corpus automatic-acquisition.

disambiguation any more. This phenomenon is very general in English, French or German which belong to the same Latin language family. Since Uyghur belongs to Altay language family while Chinese belongs to Sino-Tibetan language family, the above situation will not appear.

In the following parts, we will take Uyghur ambiguous term w as an example, and describe the automatic construction algorithm in Chinese translation label library.

Algorithm 1 (input: Uyghur ambiguous term. Input: corresponding Chinese translation label library)

Step 1 Using Uyghur-Chinese dictionary to achieve the corresponding Chinese translation of all meanings of Uyghur ambiguous term w $H=\{h_1,h_2,\ldots,h_n\}$, n means the number of meanings of ambiguous term w.

Step 2 Establishing Chinese label set B={}.

Step 3 Using h_i as key words to search Chinese monolingual corpus. If Chinese sentence examples including h_i are found, put them in B.

Step 4 Repeating step3 until i=n.

Step 5 Conducting checksum and de-weight on Chinese examples in B.

Through the above algorithm, we can get the Chinese translation label corpus T of all meanings of any Uyghur ambiguous term w. The scale of the corpus is related to the size of the Chinese monolingual corpus. Now there are many existing Chinese monolingual corpuses, and to use Web extraction can enlarge the existing corpus on the Internet.

3. HMM TRANSLATION DISAMBIGUATION FRAMWORK BASED ON IMPROVEMENT

3.1. Design philosophy of disambiguation framework

Through algorithm 1, we can establish Chinese translation label corpus for disambiguation. Using the label corpus as training corpus, we can achieve the context of the ambiguous word, adopt guided translation disambiguation methods, and learn translation disambiguation knowledge from training corpus. Theoretically translation disambiguation methods can be realized by any machine learning algorithm. This thesis adopts improved Hidden Markov Model (Hmm) for translation disambiguation. The model's disambiguation process can be shown as Fig. (**2**) and explains the model's application process taking Uyghur ambiguous word e as example. Uyghur ambiguous word "3" can be translated to Chinese $c_1,c_2,...,c_n$. The task of Uyghur-Chinese translation disambiguation is to choose appropriate Chinese translation for the Uyghur ambiguous terms "e". Using the existing Chinese translation label corpus, we can achieve all statistical information of $c_1,c_2,...,c_n$, for example, co-occurrence of c_n and different context. Then we obtain the context information of c_n in Chinese translation waiting for disambiguation; using improved HMM model to calculate c_n with largest probability value in the current context, so as to achieve the correct translation.

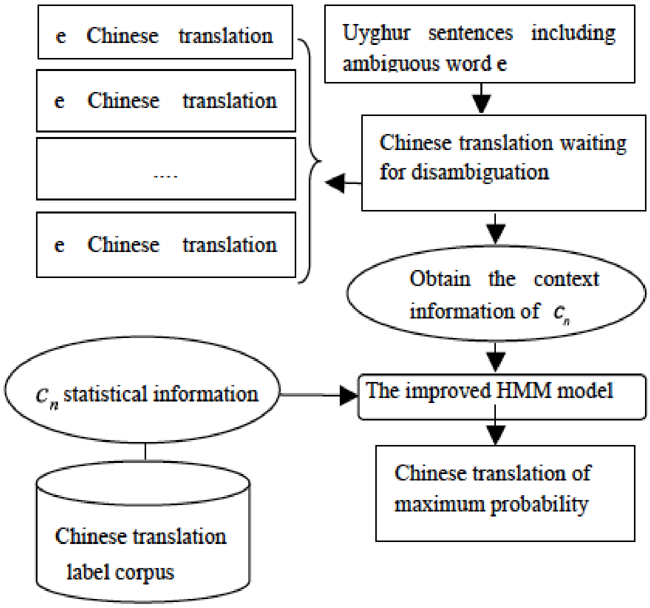

Fig. (2). Translation disambiguation framework based on corpus automatic-acquisition.

3.2. Improved HMM for Disambiguation

When traditional HMM is used for disambiguation, there will be two problems; one is only using the previous information of the ambiguous word while neglecting the following information. In HMM model, $P(w_i \mid c_i)$ means word w_i is labeled as probability of meaning c_i, while $P\,(c_i \mid c_{i-1})$ means under the situation when word w_{i-1} is labeled as meaning c_{i-1}, word w_i is labeled as the probability of c_i. From this we can see that the calculation of meaning label of word w_i only needs the meaning label probability of the previous word w_{i-1} of the word, which is called forward dependency. In real meaning label, when we are to judge the meaning of a word, we will always need the meaning information of the word w_{i+1} which is called backwards dependency.

For example, when "بالا" is translated to Chinese, it means "children" and

"disaster", let's look at the sentences:

1) بالاخى بىرى تۇئ. He is still a child.

2) بالاپچىلىككه نىسپەتەن بىر □ ئىش كۇب. To everyone, disaster is

"بالا" in 1) means child, and is decided by "بالا" and "خى بىر "ى تۇئ" right to it.

"بالا" in 2) means disaster, and is decided by "بالا" and "ئىشۇب"

Therefore, to better achieve the context information of word w_i and use it for meaning label, the appearance of the word w_i not only depends on its previous label c_{i-1}, its own label c_i, it also has to depend on the label of the later word c_{i+1}. We need to make some improvement on the traditional HMM model to achieve the above goal, and propose new hypothesis for $P(W_{1,n} \mid C_{1,n})$ in the Hidden Markov Model, that is:

$$P(W_{1,n} \mid C_{1,n}) = \prod_{i=1}^{n} P(w_i \mid c_i, c_{i+1}) \text{, then:}$$

$$C_{1,n} = \arg\max \prod_{i=1,n} P(w_i \mid c_i, c_{i+1}) P(c_i \mid c_{i-1}) \tag{1}$$

The meaning transfer matrix of the improved Hidden Markov Model is $P(c_i \mid c_{i-1})$, vocabulary emission probability is $P(w_i \mid c_i, c_{i+1})$, then to use a quintuple to describe the improved Hidden Markov Model should be:

The improved HMM= $\{N, M, A', B', \pi'\}$, in it:

N represents the meaning number of meaning label concentration.

M represents the total number of vocabulary in training corpus.

A' represents meaning transfer matrix $P\,(c_i \mid c_{i-1})$.

B' represents vocabulary transfer matrix $P(w_i \mid c_i, c_{i+1})$.

$\pi' = \{\pi_i\}$ is initialized probability vector.

Same with traditional HMM, estimation of parameter can use the method of maximum likelihood:

$$P(c_i \mid c_{i-1}) = \frac{count(c_i, c_{i-1})}{count(c_i)} \tag{2}$$

$$P(w_i \mid c_i, c_{i+1}) = \frac{count(w_i, c_i, c_{i+1})}{count(c_i, c_{i+1})} \tag{3}$$

$count(w_i, c_i, c_{i+1})$ means the occurrence number during training corpus, w_i is labeled as c_i and the later word is labeled as c_{i+1}. From formula 2 we can see that through the transformation of traditional Hidden Markov Model (HMM), the transformed model has stronger dependence on the context information, and can conform to the real requirements of meaning label more.

3.3. Improved Viterbi Algorithm

Since meaning transfer matrix and vocabulary launching matrix in HMM model have been improved, traditional Viterbi decoding algorithm cannot satisfy the requirements of disambiguation, so this thesis also makes improvement on traditional Viterbi algorithm.

Suppose the given lexical cluster $span = w_1, w_2, ..., w_n$, and $w_i \in W$, from probability we can know that from w_1 to w_k, there are several routes, there will undoubtedly be a route which makes T_k take the largest value. Using *Viterbi*

algorithm we can calculate this route, and the algorithm is described as follows:

$$\mathrm{T}_k(i) = \max P(c_1, c_2, ..., c_k = t_i, w_1, w_2, ..., w_k)$$

$$2 \le k \le n \quad 1 \le i \le N \tag{4}$$

When the condition transformed from w_i to w_{i+1}, in order to seek the maximum of the whole route, we can rely on the route maximum of w_i for solution, *i.e.*, using recursion to get the value:

$$\mathrm{T}_{k+1}(i) = b_{ij}(w_{k+1}) \max_j [\mathrm{T}_k(j) a_{ij}] \tag{5}$$

here should be a variable which can record when transforming from condition w_k to condition w_{k+1}, which one is the optimal route during the past routes, *i.e.*, to record the optimal meaning label sequence from w_1 to w_k, and the variable is recorded as :

$$\mu_k(i) = \arg\max[\mathrm{T}_{k-1}(j) a_{ij}] \tag{6}$$

In the improved Viterbi algorithm, parameter a_{ij} and b_{ij} respectively represent meaning state transition probability and meaning launching probability. The calculation results of probability are generally decimals less than 1. After several times of multiplication operations, the calculation result of $\mu_k(i)$ tends to be 0, which exceeds the computer's floating point expression scope and cannot conduct operation effectively. Using the algorithm of logarithm, multiplication can be transformed to additive operation. Therefore, take the logarithm of $\mu_k(i)$, and transform the multiplication $\mu_k(i)$ to its additive operation. Since taking the logarithm of decimals will get negative values, we need to negate the operational result to get positive values, and the final budget result is called $Cost(i)$. $\phi_k(i) = \min[w_k Cost(i, j)]$ represents the minimal cost of the word. $sCost(i, j) = -\log(a_{ij})$.

$$w_k Cost(i, j) = -\log(b_{ij})$$

Description of the improved Viterbi algorithm:

Model input: the improved HMM model $\lambda = (N, M, A, B, \Pi)$, N is the number of meaning label.

Model output: $C_1^*, C_2^*, ..., C_n^*$ (optimal meaning label sequence)

Step 1 Initiation: i from 1 to N j from 1 to N

(a) $Cost_k(1) = sCost(0,1) + w_k Cost(1, j)$

(b) $\delta_1(i) = 0$ Means that the first word does not have precursor word, and the value is 0.

(c) $\phi_1(i) = \min[w_k Cost(1, j)]$ Remains the minimum cost of the present word

Step 2 Assign k from 2 to N, calculate (C) repeatedly

Step 3 j,i from 1 to T, recursively compute the optimal route of each meaning label cluster

(a) $Cost_k(j) = w_k Cost(j, i) + \min[Cost_{k=1}(i-1)]$ Gets the cost of optimal route of each moment

(b) $\delta_k(j, i) = \arg\min_i [Cost_{k-1}(i) a_{ij}]$ Save the optimal precursor meaning to the present word.

(c) $\phi_k(j) = \min[w_k Cost(1, j)]$ Remains the minimum cost of the present word

Step 4 When reaching the end w_n word cluster that waits to be labeled, calculate the optimal meaning label of the word.

(a) $P^* = \min_i [Fee_n(i)]$ Means the cost of the chosen optimal label route at the last moment.

(b) $T_n^* = \arg\min_i [Fee_n(i)]$ Saves the boundary state of optimal label route.

(c) $\phi_n(j) = \min[w_k Fee(N, i)]$ Saves the optimal label cost of the last word

Step 5 Using backtracking, from k=N-1 to 1, seeking $T_n^* = \arg\min_i [Fee_n(i)]$ on the optimal label passage, the result is the probable meaning sequence:

$$T_k^* = \delta_{k+1}(T_{k+1}^*)$$

Then $T_1^*, T_2^*, ..., T_n^*$ is the optimal meaning label cluster.

4. EXPERIMENT

4.1. Experiment Corpus and Label Set

Our project team collects the current politics of *Sinkiang Daily* in 1994 as Chinese monolingual corpus, all together 45.9MB. Uyghur ambiguous words disambiguation task refers to all requirements of English word meaning disambiguation task on Senseval-2. It selects 14 nouns as words waiting for disambiguation from Uyghur ambiguous words collection to test the corpus automatic-acquisition translation disambiguation task. For each Uyghur noun, we select sentences including the noun from *Sinkiang Daily* (Uyghur edition) from 1998 to 2000, and select 10 sentences according to the meaning of each word, and collect all together 3540 sentences of Uyghur testing examples. We translated these sentences manually, got 354 Chinese translated sentences, and treated them as testing corpus.

Different from traditional classification tasks like part-of-speech tagging, label set of fine grit meaning is huge, which results in the over-slowness solution process and is not suitable for practical use; while meaning label set of coarseness cannot correctly distinguish meaning. Therefore, it is an important question to set appropriate label set to correctly distinguish meaning information.

This article adopts the subclass of classification system in *Chinese Thesaurus* (improved version by Harbin Institute of Technology) as label set, and there are 1400 meaning labels.

4.2. Evaluation Standard

This experiment adopts 3 evaluation functions: meaning label accuracy, meaning label recall rate, and F value, as is shown in formula (7), (8) and (9). In the formula, a represents words with accurate meaning label, b represents the number of vocabulary, and c represents the number of all words in the corpus.

$$\text{Meaning label accuracy} \quad p = \frac{a}{b} \times 100\% \tag{7}$$

$$\text{Meaning label recall rate} \quad r = \frac{a}{c} \times 100\% \tag{8}$$

$$\text{F value} = \frac{2 \times p \times r}{p + r} \times 100\% \tag{9}$$

4.3. Experiment Process

Step 1 Uyghur-Chinese meaning translation. Using Uyghur-Chinese Dictionary and find the corresponding Chinese translation of each meaning of the Uyghur ambiguous word e.

Step 2 Determine the corresponding classification code of Chinese translation word. By searching Chinese Thesaurus, find the subclass of the word and use the code as the label code.

Step 3 Achieve translation label corpus, use the corresponding Chinese translation of Uyghur ambiguous word as key words, and search the corresponding semantic examples in Chinese monolingual corpus.

Step 4 Use the improved Hidden Markov Model to establish disambiguation model and train each parameter.

Parameter training algorithm:

Input: labeled Chinese meaning label corpus

Output: the value of parameter A=MartirxA, B=MatrixB, $\pi = TagTreqs$ in improved HMM

Step1: using external circulation and read sentences in training corpus. When read to the end of the document, external circulation ends.

Step2: using internal circulation to process words in the sentence. When read to the end of the document, internal circulation ends.

Step3: if the word exists in the vocabulary, word frequency= word frequency+1, otherwise the word should be put in the vocabulary, and the word frequency=1.

Step4: if the word's meaning label and the later word's meaning label exist in the meaning chart, word frequency= word frequency+1. Otherwise, put the word's meaning and the later word's meaning in the meaning chart, word frequency=1

Step5: add 1 to the corresponding sign record in TagTreqs.

Step6: add 1 to the corresponding sign record in $MatrixA$.

For initialized parameter $\pi = TagTreqs$, establish empty sign (h_0, c_0) in the experiment. By statistics of the label, we can get the meaning distribution of the first word in the sentence. Using this method, the transition matrix from c_0 to c_1 and the frequency of c_1 in the beginning of the sentence is the initial probability of c_1.

Computational formula of meaning probability:

$$a_{ij} = MatrixA[i][j] / TagFreqs[i] \tag{10}$$

Computational formula of vocabulary probability:

$$b_{ij} = (double)\, pfreq / MatrixB[i][j] \tag{11}$$

Step 5 Use Uyghur-Chinese machine translation system to translate sentences including Uyghur ambiguous word 3 to Chinese sentences.

Step 6 Obtain the context information of the Chinese translated word.

Step 7 Disambiguation, use improved Viterbi algorithm to disambiguate Chinese translation.

4.4. Experiment Result and Analysis

From the experiment result, we can prove that Uyghur-Chinese translation disambiguation framework based on automatic-acquired corpus is effective. But because of the low accuracy of automatic-acquired corpus, the accuracy of translation disambiguation is not satisfying.

From the analysis of the mistakes in experiment result, we can find that there is an important reason which leads to the low efficiency of translation disambiguation, that is, Uyghur meaning polarization. For example, the Uyghur ambiguous word"قىزىق", when translated to Chinese, means "burning hot" and "interesting", for example, "چاي قىزىق ئىچى□" (drink a cup of burning tea.) "كىنو بەكۈب قىزىق" (This film is interesting). According to statistics, the number of "burning hot" takes up 62% of the meaning. In this translation label corpus, example sentences including "burning hot" are distinctly more than sentences including other meanings. In the training system, the disambiguation framework obtains more the translation examples which are translated to "burning hot". Therefore, in the disambiguation process, decoding algorithm prefers to choose translated words like "burning hot". To improve the disambiguation accuracy, we can start from 2 aspects. On the one hand, we can enlarge the scale of the translation label corpus and include more translation label examples that include different meanings, and decrease meaning polarization rate. On the other hand, we can design filter algorithm, and filter label examples that do not conform to requirements out so as the increase the accuracy of the label library.

5. SUMMARY

This section introduces automatic-acquired corpus by means of Uyghur polysemy and Chinese translation mapping

Table 1. Experiment result.

Size Of The Training Set	Accuracy (%)	Recall Rate (%)	F Value (%)
30 (ten thousand words)	27.51	28.12	30.29
50 (ten thousand words)	30.28	32.56	34.92
70 (ten thousand words)	34.20	38.63	38.91

relation. The method can solve the knowledge acquisition bottleneck problem in Uyghur-Chinese translation disambiguation. On this basis, it proposes translation disambiguation framework and improved the meaning probability definition on the previous HMM model, and increases the model's backwards dependence. It also strengthens the model's ability to achieve context information, and makes it adapt to the parameter changes of the improved HMM by the improvement of Viterbi algorithm, and satisfies the requirements of translation disambiguation tasks. The experiment shows that Uyghur-Chinese translation disambiguation framework based on automatic-acquired corpus is effective. In order to improve the accuracy of the translation, we can make improvement from two aspects, one is to enlarge the scale of the translation corpus, and the other is to filter the noise examples of the corpus.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Sponsor acknowledgment:

(1) Ministry of Education, Humanities and social science projects (No: 12XJJC740006).

(2) Scientific research in Colleges and universities in Xinjiang key projects (No: XJEDU2013127)

(3) Key Laboratory of Xinjiang Normal University tender subject (No: WLYQ2012106)

(4) The key of xinjiang normal university bidding project (No: 12XSXZ0606)

REFERENCES

[1] Carroll X W A J. Word Sense Disambiguation Using Sense Examples Automatically Acquired from a Second Language: Proceedings of Human Language Technology Conference and Conference on Empirical Mehtods in Natural Language Processing, Vancouver, 2012[C].

[2] Moldovan R M A I. A Method for Word Sense Disambiguation of Unrestricted Text: In Proceedings of the 37th annual meeting of the Association for Computational, 2010[C].

[3] Milhalcea R. Bootstrapping Large Sense Tagged Corpora: In Proceedings of the International Conference on Languages Resources and Evaluation, 2013[C].

[4] Martinen E A A D. Unsupervised WSD based on automatically retrieved examples:The importance of bias: In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, 2009[C].

[5] C. Leacock M C G A. Using Corpus Statistics and WordNet Relations for Sense Identification[J]. Computational Linguistics, 2013,24(1):147-166.

[6] Moore A W.Hidden Markov models[D].School of Computer Science Carnegie MellonUniversity,2004

[7] Rabiner L.A tutorial on hidden Marlov models and selected applications in speech recognition[C]//Proceedings of IEEE, 1989

[8] Hinrich Schutze,Yoram Singer.Part-of-speech tagging using a Variable Memory Marlov Model[C].In proceeding of the 32thACL,NM,USA,1994:181-187.

[9] Scott M.Thede,Mary P.Harper. A Second-order Hidden Markov Model for Part-of-speech Tagging[C].In proceeding of 37thACL,MC,USA,1999:175-182.

[10] Zollmann, A., Venugopal, A., Vogel, S.: Bridging the inflection morphology gap for Arabic statistical machine translation. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York City, USA, pp. 201-204 (2006).

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014