IEEE International Conference on E-Business Engineering

Automatic Concept Mapping between Multilingual Dictionaries and WordNet to Derive Vocabulary for E-Marketplace

Guangyi Xiao, Jingzhi Guo and Zhiguo Gong

Faculty of Science and Technology University of Macau, Macau, China

{ya97409, jzguo, fstzgg}@umac.mo

Abstract - Creating semantically consistent multilingual vocabulary is challenging to concept engineering, natural language process (NLP) and collaborative editing for cross-domain meaning understanding. This paper proposes a novel method of automatic concept mapping (MyLangMapper) between multilingual dictionaries (same as bilingual dictionary, abbreviated as BD) and the Princeton's WordNet (WN) to derive vocabulary for e-marketplace. It is heuristically developed following the analysis on the senses of words appeared in both WN and BD. By this approach, a word association categorization scheme is introduced to categorize words in WN and BD for achievement analysis, and proposed approach with two patterns of automatic words mapping are created to disambiguate the meaning of words being mapping. This approach's implement and experiment result show that the proposed method is a good complement of the existing collaborative concept editing for vocabulary building method. It suggests that collaborative work for vocabulary building can be significantly reduced.

I. INTRODUCTION

E-marketplace is a common business information space (CBIS), which must satisfy four e-marketplace properties of distribution, autonomy, interdependence and emergence [4]. In order to enable buyers and sellers to interoperate with each other in e-marketplaces, researchers arguably propose different methodologies to adopt standard vocabulary, automatically generated vocabulary, and collaboratively created vocabulary [5][14][15]. Regardless which method should be adopted, e-marketplace needs a semantically consistent vocabulary across heterogeneous e-marketplace environments to allow understandable business message exchange. WordNet (WN) [3] is a popular semantic vocabulary including nouns, verbs, adjectives and adverbs grouped into sets of cognitive synonyms (synsets), each expressing a semantic concept. Currently, many Natural Language Process (NLP) research groups have adopted the WN lexical database as the semantic concept network, for example, for machine translation and domain labelling [17].

Although English-based information systems can use synonyms based WN as common vocabulary, there is no existing mapping of the WN onto other language to form WN-alike lexical databases. For example, how to automatically build a Chinese WN is a popular issue in Natural Language Process [7]. Aiming at semantically integrating the WN and any English-Chinese dictionary, this paper proposes a novel generic approach to Automatic Creating Semantic Vocabulary (MyLangMapper), which builds the automatic mapping between words of a bilingual dictionary (BD) and the WN. This approach is useful to build a larger multilingual vocabulary or dictionary based on the existing WN and bilingual dictionaries. It also enables to drastically reduce collaborative editing work on designing a completely semantically consistent vocabulary, which is a mandatory requirement for business message exchange discussed in [5][6].

As noted, each word in the WN contains multiple senses. Each sense has some properties such as part-of-speech, semantic relationships with other word sense, and samples for each synset (sets of cognitive synonyms). Most of the existing bilingual dictionaries (BD) also have multiple word senses for each word, in which its sense also contains some properties such as part-of-speech, English definition, Chinese definition, English samples, and Chinese samples. The semantic relations between the existing WN and an English-Chinese dictionary can be exemplified in Table 1 for the word "fruition" appeared both in the WN and the BD.

TABLE 1: EXAMPLE FOR SEMANTIC MAPPING

	WN	BD
Fruition	1. something that is made real or concrete SS 2. enjoyment derived from use or possession 3. the condition of bearing fruit	1. Realization of something desired or worked for; accomplishment: 实现：实现所期望的或为之奋斗的事情；获得： 2. Enjoyment derived from use or possession. 享用：由于使用或占有而获得快乐 3. The condition of bearing fruit. 长果实的状态
Realization	1. … 2. … 3. something that is made real or concrete (realisation)	1. The act of realizing or the condition of being realized. 实现：实现的行为或处于已实现的状态 2. The result of realizing. 实现的结果

To elicit the research problem of this paper, we provide a motivational example in Table 1, where a synset in WN

978-0-7695-4227-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICEBE.2010.25

contains fruition$_{w}^{1}$, realization$_{w}^{3}$ and realisation$_{w}^{3}$. This synset has the same definition in WN which means “something that is made real or concrete”. The word fruition in WN has two other word senses. Fruition$_{w}^{2}$ means “enjoyment derived from use or possession”, and frution$_{w}^{3}$ means “the condition of bearing fruit”. In the bilingual dictionary fruition also has three different word senses. Fruition$_{d}^{1}$’s English definition is “Realization of something desired or worked for; accomplishment:” and its Chinese definition is “实现：实现所期望的或为之奋斗的事情；获得：”. Fruition$_{d}^{2}$’s English definition is “Enjoyment derived from use or possession.” and its Chinese definition is “享用：由于使用或占有而获得快乐”. Fruition$_{d}^{3}$’s English definition is “The condition of bearing fruit.” and its Chinese definition is “长果实的状态”. Other words’ definitions can be seen in table 1. Now, our problem is how to map word sense in WN onto word sense in bilingual dictionary, since many definitions and samples are different in syntax. An intuitive method is to map a word sense in WN onto a word sense in BD by comparing their respective word definition. This method is easy and efficient. However, two consistent word senses with common definitions in both lexicon databases are not often. Thus, the accuracy for mapping is low. In order to analyse the patterns existing in the problem, we draw a relation diagram for this example in Figure 1. First, we find common word definitions appeared in both WN and BD. For example, the definition for fruition$_{w}^{2}$ and fruition$_{d}^{2}$ is exactly the same syntactically (we can see the mapping in the red line; and the definition of fruition$_{w}^{3}$ and fruition$_{d}^{3}$ is also exactly the same syntactically (mapped by the green line). Second, we find similarity pattern between two definitions from WN and BD（Although the similarity score between fruition$_{w}^{1}$ and fruition$_{d}^{1}$ is not syntactically high, we cannot say fruition$_{w}^{1}$ is consistently mapped onto fruition$_{d}^{1}$ since both of them are remained has not mapped with others. Maybe fruition$_{w}^{1}$ is not converged in the bilingual dictionary and fruition$_{d}^{1}$ can also not converged in WN although this probability is very low. However, we find that the definitions for fruition$_{d}^{1}$ and realization$_{d}^{1}$ are quite similar with each other especially for Chinese definition. So if fruition$_{d}^{1}$ is quite similar to realization$_{d}^{1}$, we can infer that fruition$_{w}^{1}$ is mapped with fruition$_{d}^{1}$ and realization$_{w}^{3}$ is mapped with realization$_{d}^{1}$ because fruition$_{w}^{1}$ and realization$_{w}^{3}$ are contained in the same synonyms set (synset). There are one or more words contained in one synset. So these words’ senses have the same meaning in WN. In essential, these words’ senses BD will also have the possible similarity meaning. So if the similarity scores for these words’ senses in BD is higher, then we can possible infer the accurate word sense mapping.

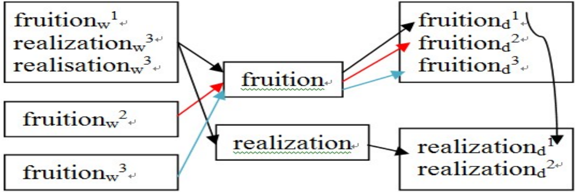


Figure 1: A relation diagram for word senses fruition, realization

In the above two example patterns, the most useful features in the bilingual dictionary are (1) each word sense corresponds to a paired English word and a Chinese word, which means their definitions are semantically equivalent in word sense, and (2) Fruition$_{d}^{2}$ and Fruition$_{d}^{3}$ are syntactically the same as Fruition$_{w}^{2}$ and fruition$_{w}^{3}$, which means they may be fully semantically the same. Based on the above features, we can summarize that the research issue of automatically creating a semantic vocabulary of mapping WN and BD is to find the similarities between words of WN and BD. Thus, to map words between WN and BD, this paper will focus on how to group words and how to compute their similarity between words of WN and BD.

In the remainder of this paper, we first review some related work in section 2, and then in section 3 describe the approach of automatically creating semantic vocabulary. In Section 4, the simulation experiments are presented. Finally, we draw the conclusion with some future work.

II. Related Work

Ontology is a domain-based shared vocabulary. Ontology engineering has a long history for vocabulary editing. It is used to build a semantically understandable domain identified with URLs usable in Internet. Domain-wide ontology design determines different ontology created by different ontology engineers or systems are heterogeneous, since different creators often have diverse background knowledge underlying their own contexts [17]. Thus, ontology engineering is not primarily used to design cross-domain vocabularies.

In [5], Guo proposed a cross-domain vocabulary design approach by introducing collaborative editing on concept agreement. Collaborative editing is helpful to disambiguate the word sense of any concept, but it also needs to pay a considerable collaboration cost.

Automatic cross-domain vocabulary design is a good complement of collaborative editing method in achieving semantic consistency between heterogeneous vocabularies. Natural language process for aligning heterogeneous vocabularies can avoid the problems of the fixed structure based and domain based ontology. Building a multilingual-WordNet for different languages is a challenging issue for aligning heterogeneous vocabularies especially for e-marketplace. The pioneer work for multilingual-WordNet began by the European languages based research group, such as MultiWordnet [1], EuroWordnet [2] and Brazilian Portuguese WordNet [16]. For Asia language process of multilingual -WordNet, researches can be found in [6][8]. The nation natural language process research lab in Korea use the word sense disambiguation approach based on the word sense tag for Korean language and Korean corpus to automatic mapping WN. In [7], Chinese wordnet is also aligned to the Princeton’s WN. This paper used a Chinese synonyms word sets (Cilin), a Chinese tagged corpus (ASBC), some Chinese-English dictionaries, Princeton’s WN and the word sense tag corpus SemCor to automatically build Chinese WN. Christopher C. [9] automatically create cross-lingual thesaurus for multi-knowledge management based on mining internet's multi-knowledge documents.

Changki Lee proposes an automatic WN mapping using word sense disambiguation [9]. This approach presents the automatic construction of a Korean WN from pre-existing lexical resources. A set of heuristic features are extracted to map Korean in MRD to map with WN English. First, for each Korean word sense is translated to m English words. Then these m English words relate to n WN synsets as candidate senses. Actually, he assumes that all the translations in English for the same Korean word sense are semantically similar. Therefore, he uses a maximum similarity feature as the main stream feature for word sense disambiguation. After that, the prior probability feature considerate the case of the translation in English may be has only one synset. The sense ordering feature considerate the statistic result when analysis the WN'S corpus for word sense disambiguation. The Is-A relation means that if two Korean word has a is-a relation then the candidate synsets for WN also has a is-a relation. He translates Korean to English word by bilingual dictionary and then does the mapping work. Our task is mapping the WN sense to bilingual word sense directly. We focus on how to save our time or work on manually justify of the consistence mapping of each word sense.

III. MyLangMapper Approach

In this section we present our approach for automatic mapping the WN's word senses (or synset) onto the English-Chinese Bilingual Dictionary's Definitions. The result of this automatic word mapping is used for later collaborative concept mapping mandatorily required by Collaborative Concept Exchange (CONEX) [5], which is out of the discussion scope of this paper. The relationship between the automatic word mapping proposed in this paper and the collaborative concept mapping is that the more automatic mapping work is done, the less collaborative concept mapping is required. This implies that a better research result suggested in this paper, less workload will be needed in collaborative vocabulary editing work later.

To achieve the above automatic word mapping goal, our MyLangMapper approach is heuristically developed following the analysis on the senses (i.e. the definitions) of synonymous and homonymous words appeared in both WordNet (WN) and Bilingual Dictionary (BD).

Before discussing the MyLangMapper approach in details, we first review the basic concepts for WN and Bilingual Dictionary. WN contains some synsets which involved a set of synonyms with some word. There are some special relational links between each synset such as hypernym, hyponym, antonym, etc. Besides those, each synset may have some samples. These samples are bidding with some special word in this synset. We can substitute the special word with any word in the same synset. On the other way, for each English word in the BD, there is one to more definitions. One English definition will follow one Chinese definition. For each definition, there are also one or more samples. One English sample will also follow one Chinese sample. The overview statistic data for the WN and the BD can be shown in Table 2, which explains the computational complexity.

Table 2: Statistic data for WordNet and BiDictionary

	WordNet	BiDictionary
synset	117659	unknown
sense	206904	175238
word	147306	99890

Based on the above information, we discuss the MyLangMapper approach in the following.

A. WN-to-BD Association Categorization

Heuristically analysing the WN and the BD overviewed in Table 2, we can formalize some features for word sense disambiguation (WSD) between the WN and the BD, shown in Figure 2. Through this Figure, we can find the association between WordNet synset and Definition in bilingual dictionary corresponding to each English word. Each word *i* W_i not only connect to synset *j* S_j in WordNet (WN) but also connect to different Definition $D_{i,\ k}$ in bilingual dictionary (BD).

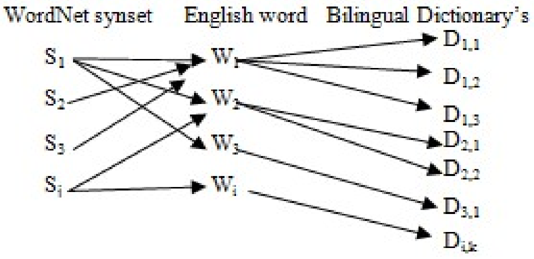


Figure 2: WordNet-to-Bidictionary Assosiation

1) Association Categorization Analysis

From Figure 3, some WN-to-BD association categories can be found from the analysis of the data statistics of WN and BD. The first association category is *the single synset*, denoted as C_1, which means that a set of words is only defined by one word sense of a synset.

The second category is *the two connected synsets*, denoted as C_2, which means that for any two sets of words defined by two senses of synset *i* and synset *j*, there exists at least one word has both senses *i* and *j*.

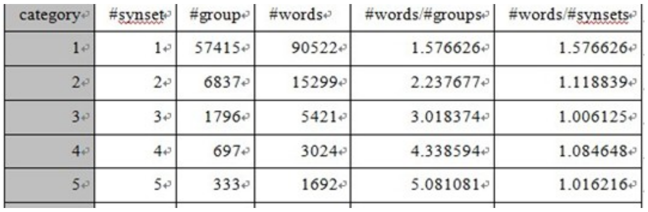
Likewise, we have the third category of *the three connected synsets*, denoted C_3, which means that for any three sets of words defined by three senses of synset *i*, *j* and *k*, there exists at least one word has both senses of *i* and *j* and there at least one word has both senses of *j* and *k*.

Generically, an Nth category is the N-connected synsets, denoted C_n, which means that N sets of words, defined by the senses of synsets S_1, S_2, …, S_n, are semantically connected at least by their individual two connected synsets.

2) Association Categorization Result

The results of the distribution of the categories with different number of connected synsets are showed in Table 3, and Table 4. Column #synset is the number of connected synsets; column #group is the number of groups in the category; column #words is the number of words in these category; column #words/#groups is the number of words divided by the number of groups which is the word density of each group in this category; and column #words/#synsets is the number of words divided by the number of synsets which is the word density of each synset in this group.

TABLE 3: STATISTIC DATA FOR GROUP OF THE CONNECTION OF SYNSETS



category	#synset	#group	#words	#words/#groups	#words/#synsets
1	1	57415	90522	1.576626	1.576626
2	2	6837	15299	2.237677	1.118839
3	3	1796	5421	3.018374	1.006125
4	4	697	3024	4.338594	1.084648
5	5	333	1692	5.081081	1.016216

In Table 3, we separate the synset groups by the number of synsets that are connected together. The statistics shows that if the number of connected synsets is fewer, the computational complexity for these synsets is then lower and the accuracy for automatic mapping between the WN and the BD is higher. The first category's groups are very large in number, and these groups are associated with 90522 words. The last category's group is also very large. This category only has one group but it associates with 32971 synsets with word density as 1.5766. This group is associated with 27276 words with word density as 0.8272. The detailed information about word density for each synset for all categories is showed in Figure 3.

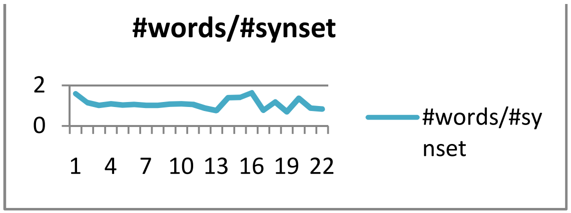


FIGURE 3 THE WORD DENSITY FOR THE SEPARATED GROUPS

By the analysis of the connection of synsets and word distribution for two dictionaries, we find that handling each separated diagram respectively is achieved. First, in each connected diagram we can use the glossary method such as how many common words involved in two text-based strings to computing the similarity score between WN and BD for all the candidate maps. Second, in the glossary method and some special finding such as common synonyms in b these two special dictionaries we can find one most confident candidate map to each synset with the highest similarity score. Third, in order to increase the accurate of mapping result, for the last ranking score for each candidate not only account for glossary similarity score between WN and DB but also account for glossary similarity score between this DB and the most confident map's DB. At last, in order to increase the better efficiency for accurate map and computing time, the clipping pattern is used to clip redundancy maps.

B. Approach for automatic word senses mapping

1) MyLangMapper Approach Overview

We now give a specification of what we call the automatic mapping semantic concept between Multilingual Dictionaries and WordNet to Derive Vocabulary for E-Marketplace (abbreviated as MyLangMapper). In this approach, we just map the common words involved in both dictionaries and each word is indexed by i. W_i has lots of senses in WN related to each synset $S_{i,h}$. Similarly, W_i also has lots of senses in BD related to each $D_{i,k}$. So for each synset $S_{i,h}$ and each explanation $D_{i,k}$ can construct a confident map indicated as $M_{i,k,h}$. For example, $S_{i,h}$'s number of confident maps is the length of word W_i's explanations in DB, which is indicated as $|D_i|$. Similarly $D_{i,k}$'s number of confident maps is the length of word W_i's synsets in WN, which is indicated as $|S_i|$. Beside the only accurate map there are lots of redundancy maps in candidate maps. We will handle and computing the rank score for candidate maps one by one order by some executing order. When we are computing one candidate map's rank score for $S_{i,h}$, if we can find this candidate map is the accurate map then we call this map as confident map. This confident map can be found by some measures such as the rank score is bigger than a threshold which is come from some statistic result, or this confident map is indicated by users. So if we can find the current computing map is the confident map, we can clip some redundancy unhandled maps where unhandled maps' explanation in BD is the same as confident map's explanation in BD. After all the candidate maps have been computing the rank score, the word senses' maps can be ranked by the rank score. So for synset $S_{i,h}$ can find the top one $D_{i,k}$ with the maximum rank score from all $|D_i|$ candidate maps. In general the top one of rank maps is the accurate map. This result cannot get 100% accurate map but can get more than 70% accurate maps in our experiment result. The formulation for calculate rank score for all the candidate map will be illustrated in the next section with first pattern. The formulation for clipping the redundancy candidate maps will also be illustrated in the next section with second pattern.

2) MyLangMapper Approach

Definitions:

i: *The index of word i which is the common words between WN and BD.*

j: *The index of synset j defined in WN.*

$D_{i,k}$: *The word i's k^{th} definition(k^{th} senses) involved in BD.*

$S_{i,h}$: *The word i's h^{th} synset(h^{th} senses) in WN.*

SimScore: *The similarity score for text based contents.*

RScore: *The rank score for each candidate map.*

$M_{i,k,h}$: *A structure for hold each candidate mapping between $D_{i,k}$ and $S_{i,h}$, involving <$S_{i,h}$, $D_{i,k}$, SimScore, RScore>.*

S.maxMap: *Each synset structure can hold a child structure maxMap ($M_{i,k,h}$), which is used to record the highest similarity score among all the candidate maps related to this synset .*

S.words: *Each synset structure can also hold the synonymy words involved in this synset.*

```
Initialization:
  $L_m$<M>← add all the candidate maps to the list, the length of list is $\sum_i|M_i|$.
  $L_s$<S>←add all the synsets to the list involved in WN.
  $L_c$<M>←empty   // set the confident map list to empty
  θ ←0.6   // set the threshold for confident map
Pre-process:
  FOR each $M_{i,k,h}$ in $L_m$<M> DO
    $M_{i,k,h}$.SimScore ← Sim($M_{i,k,h}$ .$S_{i,h}$, $M_{i,k,h}$.$D_{i,k}$)
    // refer to formulation (4) in next section
Pick max map process:
  //Pick a candidate map with max similarity score for
  //each synset
  FOR each $S_j$ in $L_s$<S> DO
    tempScore ←0
    FIND each $M_{i,k,h}$ in $L_m$<M> where $M_{i,k,h}$.$S_{i,h}$ equals to $S_j$.
          DO speScore←SP($M_{i,k,h}$.$S_{i,h}$,$M_{i,k,h}$.$D_{i,k}$)
              //SP() refer to formulation (3) in next section
              //which check the special relation between
              //this synset and this definition
              maxScore ← $M_{i,k,h}$.SimScore+speScore
              //refer to formulation (2) in next section
              IF maxScore >tempScore
                 THEN tempScore ← maxScore
                      $S_j$.MaxMap ← $M_{i,k,h}$
Main Process:
 FOR each $S_j$ in $L_s$<S> DO
    FIND each $M_{i,k,h}$ in $L_m$<M> where $M_{i,k,h}$.$S_{i,h}$ equals to $S_j$.
      DO $M_{i,k,h}$.RScore ←
         ($M_{i,k,h}$.SimScore+ WM($M_{i,k,h}$.$D_{i,k}$, $S_j$.maxMap.$D_{i,k}$) )/2
          \\ refer to formulation (1) for deal Rank Score
          \\WM() refer to formulation (5) forWord Match
         IF $M_{i,k,h}$.RScore > θ THEN
            \\refer to the second pattern in the next section
            \\refer to formulation (6) in the next section:
             $L_c$<M>←$L_c$<M> + $M_{i,k,h}$
             Clipping($M_{i,k,h}$)
 Function Clipping($M_{i,k,h}$ m) DO
        FOR each $M_{i,k,h}$ in $L_m$<M> DO
            IF $M_{i,k,h}$.$D_{i,k}$ equals to m.$D_{i,k}$ THEN
               $L_m$<M>←$L_m$<M> - $M_{i,k,h}$
Final-process:
        Store $L_m$<M>, $L_c$<M>, $L_s$<S>
```

C. Two patterns for automatic words mapping

From the Figure 2, two basic patterns of the automatic mapping are found, which are word similarity pattern and mapped definition clipping pattern.

1) Similarity between WN Synset Sense and BD Definition and Similarity among BD definitions with different words

For automatic mapping problem for non-monosemy words, the basic concept of first pattern is very simple. For any WN synset consisting of multiple words, there will be at least 0 or 1 BN word definition semantically mapping onto the WN synset sense if this word is also appeared in WN.

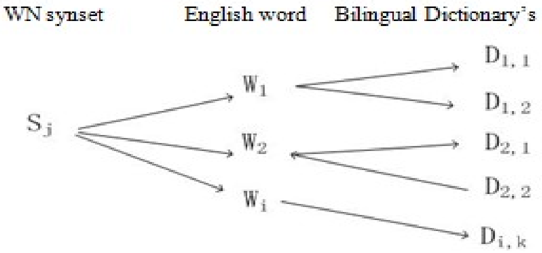


FIGURE 5 GENERIC ASSOCIATED FORMALIZE FOR EACH SYNSET

Generically speaking, for each synset S_j (also refers to its word sense) we have the associated diagram, shown in Figure 6. This synset consists of *m* English words, each denoted as W_i appeared in both WN and BD. The W_i also associates with *n* BD Definitions, each denoted as $D_{i,k}$.

Given the above notation, our objective is to find the highest semantic similarity between S_j and $D_{i,k}$ common to the English word W_i. To measure and rank the similarity, we suggest a rank score that computes the similarity. The rank score is formalized as follows:

$$R_i(S_j, D_{i,k}, D_{\max}) = \frac{Sim(S_j, D_{i,k}) + WM(D_{i,k}, D_{\max})}{2} \quad (1)$$

where Sim(S_j, $D_{i,k}$) is the similarity score for weighed vector referring to the text-based content, such as the English and Chinese word definition in BD, the WN synset sense, and examples of S_j in the WN and the BD. WM($D_{i,k}$, D_{max}) is used to compute the similarity from all the English definitions, Chinese definitions, English samples and Chinese samples between current computing $D_{i,k}$ and current synset's picked out maximum candidate map's explanation D_{max}.

In pick up max map process, D_{max} is stored in the explanation from S_j.maxMap. The formulation for choice one maximum similarity map with the max similarity score from all the candidates for synset S_j is as follows:

$$S_j.\max Map = \arg\max_M (Sim(M.S, M.D_{i,k}) + SP(M.S, M.D_{i,k})) \quad (2)$$

where SP(S_j, $D_{i,k}$) is a special function to pick out the maximum score map among all candidate maps. This special function is just appropriated in this case. In our bilingual dictionary, some word will indicate some relation to some other word. For example in BD, some Definitions $D_{i,k}$ for a word W_i will indicate some link to W_s which is any special word linked to W_i. For example, in W_j's definition there are some special terms W_s such as "see &b{W_s}". If W_s is also consisted in synset S_j's words, then we said this Definition $D_{i,k}$ is the best similarity candidate. For example, W_j's special Definition $D_{i,k}$ looks like that "see W_s", and W_s is also contained in S_j's words. So the formulation of special pick function is as follows:

$$SP(S_j, D_{i,k}) = \begin{cases} 1 & iff(\exists W_s \in D_{i,k} \wedge W_s \subseteq S_j.words) \\ 0 & others \end{cases} \quad (3)$$

The similarity function for candidate maps between synset S_j and $D_{i,k}$ corresponding to English word *i* is the kernel function for automatically mapping work. Fewer definitions for an English word W_i is easier to find the best map from candidate maps, since we propose this fewer definitions for W_i as prior. Besides, $|D_i|$ the number of definitions in BD for word W_i has the negative impact because the number of definitions is smaller, the mapping work will be easier to find the accurate mapping. The most positive impact about similarity score is to calculate the co-match [9] word involved in both WN's synset and BD's explanations, and the cosine similarity about text based weighed vector [12]. The similarity score's function is as follows:

$$Sim(S_j, D_{i,k}) = w_1 \times \frac{1}{|D_i|} + w_2 \times WM(S_j, D_{i,k}) \quad (4)$$

The word match formulation consists of two parts. For computing the similarity between explanations, examples of Synset S_j with the explanations, samples of $D_{i,k}$, the first part for co-match word is good enough [9] which is the number of common words set divide the number of union words sets between WN and BD. For getting similarity score between two Definitions in BD, the cosine similarity measure [12] is used to do it since all the definition and examples come from the same source, which proved promising approach for similarity is computing for text based documents. So the word match similarity score function is as follows:

$$WM(W,V) = \alpha \frac{|W \cap V|}{|W \cup V|} + \beta \frac{W \cdot V}{|W| + |V|} \quad (5)$$

where W and V are word based content such as explanation, sample in WN and BD respectively. The first part is the co-match similarity computing [9], and the second of the formulation is the cosine similarity measure [12].

2) Clipping redundancy candidate maps from the confident mapped definitions in BD.

The second pattern is clipping redundancy candidate maps from the confident mapped definitions in BD, which is delete the mapped definitions for other unhandled synsets' in main process. For example in Figure 1, $realization_d^1$ is confident mapped to $realzation_w^3$. So we clip $realization_d^1$ in the candidate set. Next when we computing the map result for $realiazation_w^1$ and $realization_w^2$, $realazation_d^1$ is not necessary to computing. We just compute the similarity scores from $realization_w^1$ and $realazation_w^2$ to the only one unmapped definition $realization_d^2$. If we can get some confidence map which is account as the similarity score is bigger than a confident threshold about the semantic consistency between a WN's word sense S_j and a BD word definition $D_{i,k}$, then we can clip this word definition $D_{i,k}$ for this word *i*. So for further computing, when next synset S_{next} also connect to this word *i*, the similarity score computing with first pattern can only take other definitions as candidate set which contains no $D_{i,k}$. If $D_{i,k}$ is not clipped as the confident map, this definition will also available for its unhandled maps for computing similarity score for S_{next}. So if we can find the confident map or if we can find the appropriate parameter as a threshold for confident mapping, we can clip the redundancy candidate maps for saving lots of computing time.

For example, see the synset's relationship linked by some words in Figure 2, and Figure 3. For analysing the computational complexity, the method proposed by first pattern is not very efficient. From the motivation example, if we can identify the mapping between $fruition_w^2$ to $fruition_d^2$ and the mapping between $fruition_w^3$ to $fruition_d^3$ since those definitions and Definitions for the same English word is exactly the same. Therefore we can clip the word sense in BD such as $fruition_{d,2}$, $fruition_{d,3}$. After this clipping, we adopt the first pattern approach to do the remainder mapping work which will get more efficient to computation and get more precise rate for mapping. So from this heuristic, the former computation's accurate rate will impact on the latter computation's accurate rate. And the calculation order and prior have some indicate impact on the experiment result. That's why we have done some statistic work in the former section.

In clipping function, we give the clipping formulation as follows:

$$Clip(D_{i,k}) = \begin{cases} D_{i,k} & iff(R_i(S_j, D_{i,k}) > \theta \\ NULL & otherwise \end{cases} \quad (6)$$

where $R_i(S_j, D_{i,k})$ is the current computing similarity score from WN's synset S_j to word w_i's kth explanation $D_{i,k}$, if this computing result is bigger than the confident mapping threshold, then clip the candidate maps which also involved this $D_{i,k}$ for other unhandled synsets.

At last, we should consider the execute order and the prior for each synset. The separate group work is necessary for this work. And then we assign the prior weight to each category of the connected diagrams. After our simulation experiments and lots of analysis, we assign the compute prior to each category as follows:

$$Order(S_j) = \frac{1}{|S_j.words|} \quad (7)$$

where $|S_j.words|$ is the number of words related to synset S_j.

Two formulations of (4) and (7) have thought about all the categories when we analysis the dataset relationship in the beginning of this chapter which is heurist from [9]. In essential, if one synset has fewer words in WN and these words also have fewer explanations in BD, then finding the accurate mapping result will get much easier.

Our proposed approach is not only the automatic approach for word sense disambiguation between concept lexicon and bilingual lexicon but also have good compatible for collaborative editor for semantic vocabulary. If some editors have mapped some synset S_j to one explanation $D_{i,k}$ corresponding to English word W_i. Then we just set the manually mapped $D_{i,k}$ as the confident map for this synset. So our Clipping pattern will clip this $D_{i,k}$ from the unhandled candidate maps. Then it is better for the Automatic Agent to do the next automatic mapping job.

IV. SIMULATION AND EXPERIMENTS

1) Experiment Setting

We adopt nine steps to create a larger English-Chinese bilingual semantic vocabulary as an experiment on the MyLangMapper approach. Particularly, the steps are:

(1) Collected English WN dictionary from Internet and created English-Chinese bilingual dictionary (BD).

(2) Parsed the WN and BD dictionaries to structured data. Each definition is an identity involving the English explanation, Chinese explanation, part-of-speech, and zero or more samples. The structure of WN is organized by a set of synonyms. Each synonym contains one or more English words indicated as a sense, English definition, part-of-speech and some samples. In this proposed approach the relations between synonyms indicated by WN are not applied.

(3) Analysis the threshold for confident map. Because lots of $S_{i,h}$ and $D_{i,k}$ are the one-to-one maps. We first computing the rank score for all the one-to-one maps. And then we compute the threshold of confident map as the average of all the rank scores for all the one-to-one maps. This threshold is almost close to 0.6031. So in our experiment, beside the baseline experiment we also have done 3 experiments set threshold as 0.4, 0.6 and 1 respectively.

(4) For each experiment with our proposed approach, we do the experiment step by step by the procedure of MyLangMapper.

2) Implementation and Evaluation

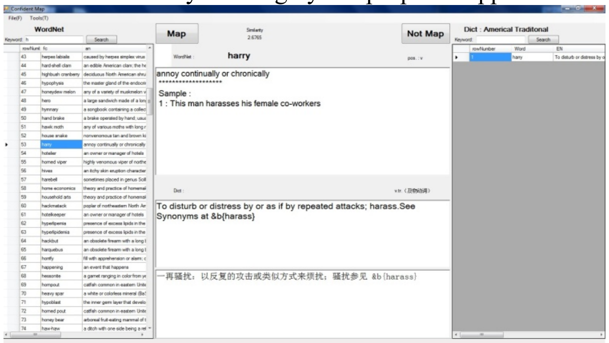
a. Experiment Implementation

In order to improve the efficient of computing, all the experiments are implemented by the c++ programming. For these experiments, the uniform strut for candidate map is adopted for the entity analysis. This entity class involves the synset's id, synset's definition, the common word, and the part-of-speech of WN, the part-of-speech of BD, the English explanation for WN, the English explanation for BD, the Chinese explanation for BD, the samples, similarity scores for three iterations, and the final rank score. First, we generate the product join for this entity. Then, after each step, we work in two aspects for these product join entities or records, which are filter and scoring. The filter is used to delete none mapping entities, such as part-of-speech filtering. The WN's part-of-speech consists of (s, a, n, v, r), which are adjective satellite, adjective, noun, verb and adverb. BD's part-of-speech is very complex, including transitive verb and intransitive verb in BD, which maps onto verb in WN. There are conjunction term, interjection term, adverb and preposition term mapping to adverb in WN. The detailed mapping of part-of-speech can be seen appendixes.

Rank Scores is to making similarity score for one word sense in WN with all the candidate word senses in BD, since we are unable to achieve 100 percent accurate mapping. Essentially, scoring is used to compute the similarity score between two English's explanations and the similarity score between two Chinese explanations. For English explanations we segment the paragraph to a set of words with some weight. This word's weight can use the TFIDF [12] feature and the increasing some words' weights in the special formulation such as "{1552-1618}". For Chinese explanations we should also segment the paragraph of Chinese. In the future for optimization we will make use the part-of-speech tagging feature.

b. Interface for Evaluation

The following Figure 6 is the confident mapping result interface. The collaborative editor can use this interface to evaluate the mapping result according the similarity score which is automatically creating by our proposed approach.



rowNumt	fc	en
43	herpes labialis	caused by herpes simplex virus
44	hard-shell clam	an edible American clam; the h
45	highbush cranberry	deciduous North American shru
46	hypophysis	the master gland of the endocri
47	honeydew melon	any of a variety of muskmelon v
48	hero	a large sandwich made of a lon
49	hymnary	a songbook containing a collec
50	hand brake	a brake operated by hand; usu
51	hawk moth	any of various moths with long r
52	house snake	nonvenomous tan and brown ki
53	harry	annoy continually or chronically
54	hotelier	an owner or manager of hotels
55	horned viper	highly venomous viper of northe
56	hives	an itchy skin eruption character
57	harebell	sometimes placed in genus Scil
58	home economics	theory and practice of homemal
59	household arts	theory and practice of homemal
60	hackmatack	poplar of northeastern North Am
61	hotelkeeper	an owner or manager of hotels
62	hyperlipemia	presence of excess lipids in the
63	hyperlipidemia	presence of excess lipids in the
64	hackbut	an obsolete firearm with a long l
65	harquebus	an obsolete firearm with a long l
66	horrify	fill with apprehension or alarm; c
67	happening	an event that happens
68	hessonite	a garnet ranging in color from ye
69	hornpout	catfish common in eastern Unite
70	heavy spar	a white or colorless mineral (Ba
71	hypoblast	the inner germ layer that develo
72	horned pout	catfish common in eastern Unite
73	honey bear	arboreal fruit-eating mammal of t
74	haw-haw	a ditch with one side being a ret

annoy continually or chronically
****
Sample :
1 : This man harasses his female co-workers

To disturb or distress by or as if by repeated attacks; harass.See Synonyms at &b{harass}

一再骚扰：以反复的攻击或类似方式来烦扰；骚扰参见 &b{harass}

rowNumber	Word	EN
1	harry	To disturb or distress by o

FIGURE 6 EVALUATING INTERFACE FOR CONFIDENT MAPPING

In our confident mapping result interface, we search the keyword of the word. All the mapping results are displayed in the left table order by the similarity score. In the middle of the top, the similarity score is showed. The collaborative editor can evaluate the mapping result according the English explanation and samples of WN and the English-Chinese explanation and samples of BD.

3) Result and discussion

In order to evaluate the result of our proposed approach, the evaluation result will compare with the baseline approach. The evaluation criterions involve the precision and the efficiency. The baseline algorithm is very simple which is computing the similarity score from synset's English definition to BD's English definition, since it is the most efficient one, the similarity scores between all the word senses in WN and word senses in BD are simply accounted as the similarity of English explanations.

There is a balance between the precision and efficiency when we set the threshold θ for selecting the confident mapping. We evaluate the mapping result in distribution 5 which one word have only one sense in WN and have only one sense BD. The least similarity score in distribution 5 for the accurate mapping is 0.86. So we will adjust the threshold according to this analysis. The inaccurate confident mapping will lead to the inaccurate scoring result because of the clipping algorithm. If θ is too small, the clipping rate is bigger, leading to small computing time. However the mapping precision will lose. In other way if θ is too big, the clipping rate is smaller leading to a long time computing. It will increase the mapping precision. The most complex computing is setting θ to the biggest value. So in order to find the better balance, we do the comparison between three experiments with setting as 0.4, 0.6 and 1 respectively. So totally we will compare the precision and efficient between these four experiments.

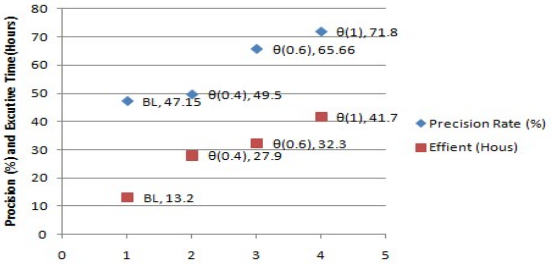


FIGURE 7 PERFORMANCE COMPARISON AMONG BASELINE (BL) AND OUR APPROACH WITH DIFFERENT THRESHOD (θ)

In Figure 7, the baseline experiment's computing time is 13.2 hours and the precision rate is 47.15%. When the threshold θ is 0.4, the computing time is 27.9 hours and the precision rate is 49.5%. When the threshold θ is 0.6, the computing time is 32.3 hours and the precision rate is 65.67%. When the threshold θ is 0.1, the computing time is 41.7 hours and the precision rate is 71.8%. Last tree experiments are finished after the baseline's experiment. Because we should make use the baseline's similarity scores to do the further computing such as picking out the maximum candidate from each synset and the boosting scores from two patterns. After these four experiments, we found our proposed approach is promising. We have increased the precision rate from the baseline approach. Although this precision result is also not very significant for e-marketplace, it will save lot of time for collaborative editor to creating Chinese semantic vocabulary. We will optimise the precision result in future. If the collaborative editor can supply some confident mapping, then our clipping algorithm will work efficient and improve the precision rate automatically. So our proposed approach can help collaborative editor to save lot of time for creating Chinese semantic vocabulary and the collaborative editor's confident mapping can also improve the precision rate and save the computing time for our proposed approach. So we propose simultaneously do the collaborative work and the automatic computing work.

V. CONCLUSION

This paper has proposed an automatic creating semantic vocabulary (MyLangMapper) approach to build semantic maps between the words of English WordNet and English-Chinese bilingual dictionaries. It consists of a method of word categorization and two patterns of automatic mapping. MyLangMapper is an automatic method for word sense disambiguation between concept lexicon and bilingual lexicon. It is a good complement of collaborative concept editing approach for building semantic vocabulary used for e-marketplace. If collaborative editors have mapped WordNet synset S_j onto bilingual dictionary Definition $D_{i,k}$ for English word W_i, the manually mapped $D_{i,k}$ will be set as the maximum score. Consequently, the connected diagram will clip the Definition from the pre-computing candidate. After clipping, Automatic Agent does the next automatic mapping job.

MyLangMapper approach is important. It contributes a method of (1) word association categorization based on connected synsets by words, (2) a word similarity computing pattern, (3) a mapped definition clipping pattern, and (4) a automatic mapping approach. Our five experiments show that this method is promising. The future work of this paper is to propose some methods for mapping unknown words (see word distribution areas 5, 6, 7, 8 in Figure 4) for both lexicons in WordNet and bilingual dictionaries. We plan to adopt some external corpus for mapping the unknown words.

REFERENCES

[1] MultiWordNet: http://multiwordnet.fbk.eu/english/home.php

[2] EuroWordNet: http://www.illc.uva.nl/EuroWordNet/

[3] Princeton's WordNet: http://wordnet.princeton.edu/

[4] Guo, J. (2007). A Term in Search of the Infrastrure of Electronic Market. In: Research and Practical Issues of Enterprise Information System II Volume 2, IFIP Volume 255, pp. 831-840.

[5] Guo, J. Collaborative Conceptualization: Towards a Conceptual Foundation of Interoperable Electronic Product Catalogue System Design. *Enterprise Information Systems* 3(1), pp. 59-94, 2009.

[6] Guo J. Document-Oriented Heterogeneous Business Process Integration through Collaborative E-Marketplace. In: Proc. Of 10th Int'l CONF. on Electronic Commerce (ICEC'08), ACM Press (Innsbruck, Austria, August 19-22).

[7] Building a Chinese-English WordNet for Translingual Applications. HSIN-HIS CHEN, CHI-CHING LIN, and WEN-CHENG LIN. ACM Transaction on Asian Language Information Processing. Vol. 1, No.2, June 2002. Pages 103-122.

[8] Automatic Learning of Text-of-Concept Mapping exploiting WordNet-like Lexical Networks. Dario Bonino, Fulvio Corno, Federico Pescarmona. In proceeding of SAC'05, ACM.

[9] Automatic WordNet mapping using word sense disambiguation. Geundae Lee, Seo JungYun. Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora.

[10] Christopher C. Yang, Chih-Ping Wei, K. W. Li.. Cross-lingual thesaurus for multilingual knowledge management. Elsevier Science Publish

[11] http://www.socialresearchmethods.net/kb/sampprob.ph

[12] Sandeep Tata, Jignesh M. Patel. Estimating the Selectivity of tf-idf based Cosine Similarity Predicates. In proceeding of SIGMOD Record, June 2007.

[13] Darja F. , B. Sagot. Combining Multiple Resources to Build Reliable Wordnets. Springer-Verlag Berlin Heidelberg 2008.

[14] Sebastian P. , Mirella L. Dependency-Based Construction of Semantic Space Models. 2007 Association for Computational Linguistics.

[15] Dario B. Fulvio C. Laura F. Andrea F. Multilingual Semantic Elaboration in the DOSE platform. In: Proc. Of SAC'04, March 14-17, 2004, Nicosia, Cyprus.

[16] Dias-da-Silva, B.C.; Di Felippo, A. and Nunes, M.G.V. 2008. The Automatic Mapping of Princeton WordNet.Br Hierarchical Realtion onto the Brazilian Portuguese WordNet database. In proceedings of the 6th International Conference on Language Resouces and Evaluation. Marrakech, Morocco.

[17] Roberto Navigli. Word Sense Disambiguation: A Survey. ACM Computing Surveys, Vol. 41, No. 2, Article 10, February 200