Building a Bilingual WordNet-Like Lexicon: the New Approach and Algorithms

Yang Liu, Shiwen Yu, Jiangsheng Yu
Institute of Computaitional Linguistics, Peking Unviersity
Beijing, 100871, China
{liuyang, yusw, yujs} @ pku.edu.cn

Abstract

A bilingual concept MRD is of significance for IE, MT, WSD and the like. However, it is reasonably difficult to build such a lexicon for there exist two ontologies, also, the evolution of such a lexicon is quite challenging. In this paper, we would like to put forth the new approach to building a bilingual WordNet-like lexicon and to dwell on some of the pivotal algorithms.

A characteristic of this new approach is to emphasize the inheritance and transformation of the existent monolingual lexicon. On the one hand, we have extracted all the common knowledge in WordNet as the semantic basis for further use. On the other hand, we have developed a visualized developing tool for the lexicographers to interactively operate on to express the bilingual semantics. The bilingual lexicon has thus gradually come into being in this natural process.

ICL now has benefited a lot by employing this new approach to build CCD (Chinese Concept Dictionary), a bilingual WordNet-like lexicon, in Peking University.

1 Introduction

As the processing of content information has nowadays become the center of NLP, a bilingual concept MRD is of increasingly great significance for IE, MT, WSD and the like. And it is for sure that the computational linguists would find such a lexicon indispensable and useful as semantic information when facing ambiguities in languages in their applications.

At the same time, Princeton University's WordNet, after so many years' development, has exerted a profound influence on semantic lexicons [Vossen, 1998].

When building a Chinese-English bilingual concept MRD, we must take the issue of compatibility with WordNet into account. In other words, for each English concept in WordNet, there should exist a corresponding Chinese concept in the bilingual lexicon and vice versa. Such a bilingual lexicon can offer better reusability and openness.

The Institute of Computational Linguistics (ICL), Peking University, with this point of view, has launched the Project CCD (Chinese Concept Dictionary).

The expectant CCD might be described as follows [Yu et al, 2001]: it should carry the main relations already defined in WordNet with more or less updates to reflect the reality of contemporary Chinese, and it should be a bilingual concept lexicon with the parallel Chinese-English concepts to be simultaneously included.

Such a bilingual WordNet-like lexicon of Chinese-English concepts can largely meet our need of applications.

However, it is by no means easy to build such a lexicon. It is quite obvious that there synchronously exist two ontologies in the same lexicon. One is in the English culture and the other is in the Chinese culture. As there might be different concepts and relations in each language, the mapping of the relevant concepts in different languages is inevitable. Also, the evolution of such a lexicon with passing of time, an issue linked closely to the mapping issue, is quite challenging.

In conclusion, it's a quite demanding job to build such a lexicon, especially for the design of the approach and the realization of the developing tool. Any fruitful solution should give enough consideration to the complexity of these issues.

2 The New Approach to Building a Bilingual WordNet-Like Lexicon

The distinct principles of organization of WordNet can be described below: concepts, viz. synsets, act as the basic units of lexical semantics, and the hyponymy of the concepts acts as the basic relation among others. Upon this tree structure of hyponymy, there also exist some other semantic relations like holonymy, antonymy, attribute, entailment, cause, etc., which further interweave all the concepts in the lexicon into a huge semantic network, say 99,643 synset nodes all told in WordNet 1.6.

What really counts and takes a lot of trouble in building WordNet itself is how to set up all these synsets and relations properly, and, how to maintain the semantic consistencies in case of frequent occurrences of modifications during the revision [Beckwith et al, 1993]. As the desirable developing tool based directly on a large-scale network has not yet appeared, due to the connatural complexity of net structure, this problem is all the way a Gordian knot for the lexicographers.

To build a Chinese WordNet in the same route just as Princeton had taken and then to construct the mapping between these two WordNets may be not a satisfying idea.

So, it is crucial that we had better find an approach to reusing the English common knowledge already described in WordNet as the semantic basis for Chinese when building the bilingual lexicon. And this kind of reusing should contain some capabilities of adjustments to the bilingual concepts besides word-for-word translations. If we can manage it, not only the building of the monolingual Chinese lexicon benefits but also the mapping between Chinese-English [Liu et al, 2002]. Actually, the practice of mapping has now become a direct and dynamic process and the evolution of the bilingual lexicon is no longer a problem. A comparatively high efficiency may be achieved.

Such are the essential ideas of the new solution. A characteristic of this approach is to emphasize the inheritance and transformation of the already existent monolingual lexicon.

Accordingly, it deals with 2 processes. The first process simply gets the semantic basis for further use and the lexicographers' work always focuses on the second. In fact, the bilingual lexicon has just gradually come into being in this more natural process.

2.1 The Inheritance Process of WordNet

This process is intended to extract the common hyponymy information in WordNet as the semantic basis for future use.

However, to extract the full hyponyms for a certain concept is by no means easy. As we have examined, the number of hyponyms for a synset ranges from 0 to 499 with a maximal hyponymy depth of 15 levels in WordNet. This shows the structure of the potential hyponymy tree is quite unbalanced. Due to this high complexity, the ordinary searching algorithm can hardly do. If one inputs the word *entity* as entry in WordNet 1.6 and tries to search its full hyponyms, he will get nothing but a note of failure. Sure enough, if the entry is not *entity* but another word, say *entrance*, the searching will probably do. The cases actually depend on the location of the entry word in the potential hyponymy tree in WordNet. The higher the level of the entry word, the less possibility of success the searching will have.

By now, we have got a refined searching algorithm for getting the full hyponymy information in WordNet [Liu et al, 2002].

By and large, it involves a series of Two Way Scanning action and of Gathering/Sieving and Encoding action, with each round of the series intending to get information of nodes on one same level in the hyponymy tree.

By this special algorithm, the complexity of searching is greatly reduced. We can even get all the 45,148 hyponyms for the topmost entry word *entity*, in 100 or so seconds, on an ordinary PC. People who are interested in it can find more details about the algorithm in [Liu et al, 2002].

2.2 The Transformation Process of WordNet

This process is for the lexicographers to interactively operate on the hyponymy tree to express the bilingual semantics. The bilingual lexicon will gradually come into being in this process.

For this task, we have designed and realized a visualized and data-sensitive tree control with 8 well-defined operations on it, some of the pivotal algorithms for which will be discussed later.

After extracting the hyponymy information for each initial semantic unit in WordNet respectively, we then organize the information into a hyponymy tree by using the above tree control. Every tree node, viz. synset, still carries all other semantic relations already described in WordNet. The lexicographers can now operate on the tree interactively.

The actual practices of the lexicographers are as follows:

(i) For each tree node in English, if there exists a corresponding Chinese concept, the lexicographers simply translate the English concept into Chinese.

(ii) If there does not, cases may be that the English concept is either too general or too specific for Chinese.

(ii_1) For the former case, the lexicographers can create new hyponyms in Chinese for the English concept and link all these new hyponyms in Chinese with the English concept.

(ii_2) For the latter case, the lexicographers just delete the English concept in a special way, which means the English concept has no equivalent in Chinese and only links the English concept with its hypernym.

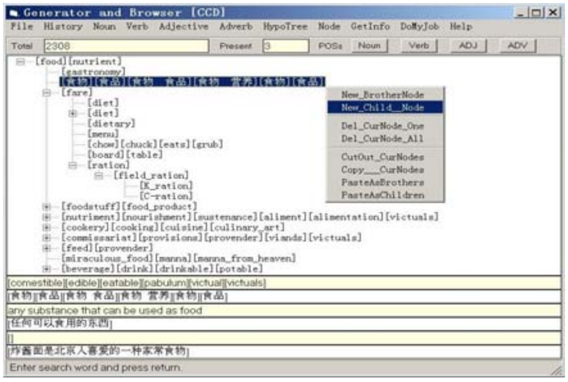
In fact, all the above-mentioned semantic manipulations concerning hyponymy relation have already been encoded into the 8 visualized operations on the hyponymy tree. In addition, in the 8 operations, some other semantic relations already described in the synsets in WordNet are all properly dealt with through systematic and reasonable calculations.

We can see these adjustments clearly in the description of the algorithms.

Now, it is of much significance that the lexicographers need simply operate on the hyponymy tree to express their semantic intention and no longer care for lots of details about the background database, for the foreground operations have already fulfilled all the automatic modifications of the database.

In this way, the problems of mapping between the bilingual concepts and evolution of the bilingual lexicon are dynamically resolved.

Our developing tool for building the bilingual WordNet-like lexicon has come out as below.



Total 2308 Present 3

```
[food][nutrient]
    [gastronomy]
    [食物][食品][食物 食品][食物 营养][食物][食品]
    [fare]
        [diet]
        [diet]
        [dietary]
        [menu]
        [chow][chuck][eats][grub]
        [board][table]
        [ration]
            [field_ration]
                [K_ration]
                [C-ration]
    [foodstuff][food_product]
    [nutriment][nourishment][sustenance][aliment][alimentation][victuals]
    [cookery][cooking][cuisine][culinary_art]
    [commissariat][provisions][provender][viands][victuals]
    [feed][provender]
    [miraculous_food][manna][manna_from_heaven]
    [beverage][drink][drinkable][potable]
```

[comestible][edible][eatable][pabulum][victual][victuals]

[食物][食品][食物 食品][食物 营养][食物][食品]

any substance that can be used as food

[任何可以食用的东西]

[]

[炸酱面是北京人喜爱的一种家常食物]

Enter search word and press return

The interface view shows the hyponymy tree for the entry *food*, which is one of the 25 initial semantic units of noun in WordNet with the category value of 13. For the currently chosen node, the lexicographers can further adopt a proper operation on it when needed.

This new kind of Visualized Auxiliary Construction of Lexicon is characteristic of the inheritance and transformation of the existent monolingual lexicon. We call it Vacol model for short.

As we see, the new approach, in fact, is independent of any specific languages and actually offers a general solution for building a bilingual WordNet-like lexicon.

3 Tree Operations and their Algorithms

As the lexicographers always work on the tool, the visualized, data-sensitive tree control with operations on it is the key to the new approach.

By now, we've schemed a set of algorithms based on the Treeview control in the Microsoft Visual Studio 6.0 and eventually implemented a data-sensitive tree control with operations on it.

3.1 Tree Operations

The 8 operations that we have semantically well defined are listed as follows. When choosing a synset node in the hyponymy tree, these are the operations from which the lexicographers can further adopt one.

[1] To add a synset as brother node;
[2] To add a synset as child node;
[3] To delete the synset node (not including its descendants if exist);
[4] To delete the synset node (including all its descendants if exist);
[5] To cut the subtree;
[6] To copy the subtree;
[7] To paste the subtree as brother node;
[8] To paste the subtree as child node.

These operations are all to edit the tree, with respectively No. 1, 2 for addition, No. 3, 4 for deletion, and No. 5, 6, 7, 8 for batch movement.

In fact, all these operations have been carefully decided on to make them concise enough, capable enough and semantically meaningful enough.

It is easy to prove that any facultative tree form can be attained by iterative practices of these 8 operations.

3.2 Algorithms for the Tree Operations

The data structure of a hyponymy tree with n nodes can be illustrated by the following table:

Pos_1	Ptr_{11}	Ptr_{12}	…	Ptr_{1m}	$BasicInfo_1$
Pos_2	Ptr_{21}	Ptr_{22}	…	Ptr_{2m}	$BasicInfo_2$
…	…	…	…	…	…
Pos_n	Ptr_{n1}	Ptr_{n2}	…	Ptr_{nm}	$BasicInfo_n$

There are 3 parts of information in each record: the structural information $\{Pos_i\}$, the relation information $\{Ptr_{i1}$ (viz. hyponymy), $Ptr_{i2}, \dots, Ptr_{im}\}$ and all other pieces of basic information $\{BasicInfo_i\}$ which are relevant only to the concept proper.

Among these 3 parts of information, $\{Pos_i\}$ is used for the tree structure whereas both $\{Ptr_{i1}, Ptr_{i2}, \dots, Ptr_{im}\}$ and $\{BasicInfo_i\}$ for lexical semantics. It should be noticed that Pos_i only stands for a special encoding for the tree in the foreground and is somewhat different from Ptr_{i1}, a relational pointer of hyponymy, which represents its specific semantics in the background database. And it is the relations in $\{Ptr_{i2}, \dots, Ptr_{im}\}$ that have highly contributed to the dense net structure of WordNet.

After these analyses, we find that each operation should just properly deal with these 3 parts of information. First, it is crucial that two sorts of consistencies should be maintained. One is that of the structural information $\{Pos_i\}$ of the tree and the other is that of the relation information $\{Ptr_{i1}, Ptr_{i2}, \dots, Ptr_{im}\}$ of the lexicon. Following that, the cases of the basic information $\{BasicInfo_i\}$ are comparatively simple for only English-Chinese translations are involved.

Before we can go on to dwell on the algorithms, we still need a little while to touch on the structural information $\{Pos_i\}$. When we say a position Pos_i, we actually mean the location of a certain node in the tree and it serves to organize the tree. For example, a Pos_i by the value “005001002” is to represent such a location of a node in a tree: at the 1st level, its ancestor being the 5th; at the 2nd level, its ancestor being the 1st; and at the 3rd level, its ancestor viz. itself now being the 2nd. In fact, such an encoding onto a linear string does fully express the structural information in a tree and makes all the tree operations algorithms feasible by direct and systematic calculations of the new position.

If we don’t want to badger with much of the details, the algorithms for tree operations can be described in a general way. Although for each line of the pseudocode, there indeed are lots of jobs to do for the programmer.

The algorithms described below are suitable for the non-batch-movement operations, viz. operations [1, 2, 3, 4]. And the batch-movement operations, viz. operations [5, 6, 7, 8], can be regarded as their iterative practices.

```
The lexicographers trigger an action on node$_i$;
IF the action is in operations [1, 2, 3, 4]
  CASE the action
    Operations [1]:
      Add a node with its Pos = NewBrother (Pos$_i$);
    Operations [2]:
      Add a node with its Pos = NewChild (Pos$_i$);
    Operations [3]:
      Delete the node with Pos = Pos$_i$;
    Operations [4]:
      Delete all the nodes with their Pos satisfying conditions of being descendants of node$_i$;
  END CASE
  Recalculate Pos of the rest nodes in the table according to the operation and current Pos$_i$;
  Replace all relevant Ptr$_{j1}$, Ptr$_{j2}$ , … , Ptr$_{jm}$ with new ones according to the operation and current node$_i$;
  Refresh the tree;
ELSE IF
The lexicographers translate current BasicInfo$_i$ from English to Chinese;
END IF
```

The algorithms have some nice features.

Since the structural information {Pos}, defined as the primary key of the table, is kept in order, the maintenance of tree structure can always be completed in a single pass.

The maintenance of consistencies of the relation information $\{Ptr_{j1}, Ptr_{j2}, \dots, Ptr_{jm}\}$ in the lexicon is also limited to a local section of the table.

4 Conclusions

ICL, Peking University has launched the Project CCD since Sept., 2000. Due to the nice features of the new approach, we do have benefited a lot by employing it to build CCD. By now, we have fulfilled more than 32,000 Chinese-English concept pairs in noun.

In the near future, ICL wants to come to a total amount of 100,000 or so bilingual concepts, which might largely meet our need of applications.

What is more, as the byproducts of the new approach and experiences, we have even found some errors and faults of semantic expressing with WordNet 1.6.

For example, in the lexicon there are many occurrences of a node with multiple-father in the identical category (772 times in noun, e.g. {*radish*}) or a node with single-father in the other category (2,172 times in noun, e.g. {*prayer_wheel*}).

In verb, there even exists a node with father being oneself (e.g. {*reserve*, *hold*, *book*}).

These phenomena are quite abnormal and puzzling according to the specification of WordNet. Something may have gone wrong with the classification or implementation.

There are also many undisciplined locations of relational pointers (e.g. "@" and "~", respectively 7 and 451 times in noun) in DAT files and some other problems.

Acknowledgements

This work is a component of researches on Chinese Information Extraction funded by National Foundation of Natural Science No. 69973005 and Project 985 in Peking Univ.

We are especially grateful to Prof. WANG Fengxin and Prof. LU Chuan, our linguistics advisors, for their unforgettable discussion and support. Many thanks go to the fellows who have participated in and collaborated on the work, among whom we would like to mention Mr. ZHANG Huarui, Ms. SONG Chunyan, Dr. LI Zuowen, Ms. ZAN Hongying and others. Thanks also to the participants to the 1st Global WordNet Conference 2002, Mysore, India, for their valuable advice and comment.

References

Beckwith, R., Miller, G. A. and Tengi, R. (1993) *Design and Implementation of the WordNet Lexical Database and Searching Software*. Description of WordNet.

Carpuat, M. and Ngai, G. et al. (2002) *Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet*. GWC2002, India, pp 284-292.

Chang, J. S. and You, G. N. et al. (2002) *Building a Bilingual Wordnet and Semantic Concordance from Corpus and MRD*. WCLS2002, Taipei, China, pp 209-224.

Cook, G. and Barbara, S. (1995) *Principles & Practice in Applied Linguistics*. Oxford: Oxford University Press.

Fellbaum, C. (1993) *English Verbs as a Semantic Net*. Description of WordNet.

Fellbaum, C. (1999) *WordNet: an Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Kamps, J. (2002) *Visualizing WordNet Structure*. GWC2002, India, pp 182-186.

Keil, F. C. (1979) S*mantic and Conceptual Development: an Ontological Perspective*. Cambridge, Mass.: Harvard University Press.

Liu, Y., Yu, J. S., Yu, S. W. (2002) *A Tree-Structure Solution for the Development of ChineseNet*. GWC2002, India, pp 51-56.

Miller, G. A. (1993) *Noun in WordNet: a Lexical Inheritance System*. Description of WordNet.

Miller, G. A. et al. (1993) *Introduction to WordNet: An On-line Lexical Database*. Description of WordNet.

Pavelek, P., Pala, K. (2002) *VisDic – a New Tool for WordNet Editing*. GWC2002, India, pp 192-195.

Touretzky, D. S. (1986) *The Mathematics of Inheritance Systems*. Los Altos, Calif.: Morgan Kaufmann.

Vossen, P. (1998) *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer.

Wong, S. H. S. and Pala, K. (2002) *Chinese Characters and Top Ontology in EuroWordNet*. GWC2002, India, pp 122-133.

Yu, J. S. (2002) *Evolution of WordNet-Like Lexicon*. GWC2002, India, pp 134-142.

Yu, J. S. and Yu, S. W. et al. (2001) *Introduction to CCD*. ICCC2001, Singapore, pp 361-366.