Automatic Construction of an English-Chinese Bilingual FrameNet

Chen, Benfeng and Pascale Fung
Human Language Technology Center,
Department of Electrical& Electronic Engineering,
University of Science and Technology (HKUST),
Clear Water Bay, Hong Kong
{bfchen,pascale}@ee.ust.hk

Abstract

We propose a method of automatically constructing an English-Chinese bilingual FrameNet where the English FrameNet lexical entries are linked to the appropriate Chinese word senses. This resource can be used in machine translation and cross-lingual IR systems. We coerce the English FrameNet into Chinese using a bilingual lexicon, frame context in FrameNet and taxonomy structure in HowNet. Our approach does not require any manual mapping between FrameNet and HowNet semantic roles. Evaluation results show that we achieve a promising 82% average F-measure for the most ambiguous lexical entries.

1 Introduction

Since the early 90's, automatic alignment of bilingual documents and sentences based on lexical and syntactic information has been a major focus of the statistical NLP community as their results are a valuable resource for statistical machine translation, cross-lingual question answering, and other bilingual or cross-lingual tasks. Recently, there has been an increasing trend of using semantic information for these tasks spurred by the availability of various ontology databases such as WordNet, FrameNet, PropBank, etc. Among these, the Berkeley FrameNet database is a semantic lexical resource consisting of frame-semantic descriptions of more than 7000 English lexical items, together with example sentences annotated with semantic roles (Baker et al., 1998). The current version of FrameNet has been applied successfully to English question answering systems (Gildea, 2002). However, the manual development of FrameNet in other languages has been on a small scale (e.g. German, Spanish, Japanese) or unfinished (e.g. Chinese). Since manually annotation is rather time consuming, the main objective of our work is to automatically create multilingual FrameNet to enable semantic analysis in multiple languages rather than in English. Another objective is to quantify the mapping between semantic structures across language pairs for statistical NLP systems. Our basic idea is to coerce the English FrameNet into another language using existing semantic resources and a bilingual lexicon. Our initial target language is Chinese. However, we expect that our technique is applicable to other languages as well. There are two Chinese semantic resources available today--Cilin (tong2yi4ci2ci2lin2) (Mei et al., 1982) and HowNet (Dong and Dong, 2000). Much like WordNet, Cilin is a thesaurus with a hierarchical structure of word clusters, but it does not describe any semantic relationship between words and categories. HowNet, on the other hand, is an ontology with a graph structure of inter-concept relations and inter-attribute relations. In addition, HowNet has been widely used in resolving NLP problems, such as word sense disambiguation (Dang et al., 2002) and machine translation (Dorr et al., 2002). For our work, we choose to align HowNet concepts to lexical entries in FrameNet in order to construct the English-Chinese bilingual FrameNet.

(Dorr et al., 2002) describes a technique for the construction of a Chinese-English verb lexicon based on HowNet and an English verb database called the LCS Verb Database (LVD). They created links between Chinese concepts in HowNet and English verb classes in LVD using both statistics and a manually constructed "seed mapping" of thematic classes between HowNet and LVD. Ngai et al. (2002) employed a word-vector based approach to create the alignment between WordNet and HowNet classes without any manual annotation. In this paper, we present a fully automatic approach to create links between FrameNet semantic frames and HowNet concepts. We also plan to release an on-line demonstration for the community to access the bilingual FrameNet we built.

2 FrameNet and HowNet

FrameNet and HowNet are ontologies with different structures and different semantic role/relation definitions. FrameNet is a collection of lexical entries grouped by frame semantics. Each lexical entry represents an individual word sense, and is associated with

semantic roles and some annotated sentences. Lexical entries with the same semantic roles are grouped into a "frame" and the semantic roles are called "frame elements". For example:

Frame: *Cause_harm*
Frame Elements: *agent, body_part, cause, event, instrument, iterations, purpose, reason, result, victim…..*
Lexical Entries in "*cause_harm*" Frame:
bash.v, batter.v, bayonet.v, beat.v, belt.v, bludgeon.v, boil.v, break.v, bruise.v, buffet.v, burn.v,….
An annotated sentence of lexical entry "beat.v":
[*agent* I] lay down on him and **beat** [*victim* at him] [*means* with my fists].

HowNet is a Chinese ontology with a graph structure of word senses called "concepts", and each concept contains 7 fields including lexical entries in Chinese, English gloss, POS tags for the word in Chinese and English, and a definition of the concept including its category and semantic relations (Dong and Dong, 2000). For example, one translation for "beat.v" is 打:
NO. = 17645
W_C =打
G_C =V
E_C =~架，~斗，~仗，~敌人，~死，~伤，~得好
W_E=attack
G_E=V
E_E=
DEF=fight|争斗

Whereas HowNet concepts correspond roughly to FrameNet lexical entries, its semantic relations do not correspond directly to FrameNet semantic roles.

3 Construction of the English-Chinese Bilingual FrameNet

(Dorr et al. 2002) uses a manual seed mapping of semantic roles between FrameNet and LVD. In this paper, we propose a method of automatically linking the English FrameNet lexical entries to HowNet concepts, resulting in a bilingual FrameNet. We make use of two bilingual English-Chinese lexicons, as well as HowNet and FrameNet. In the following sections 3.1 to 3.3, we use an example FrameNet lexical entry "beat.v" in the "*cause_harm*" frame to illustrate the main steps of our algorithm in Figure 1.

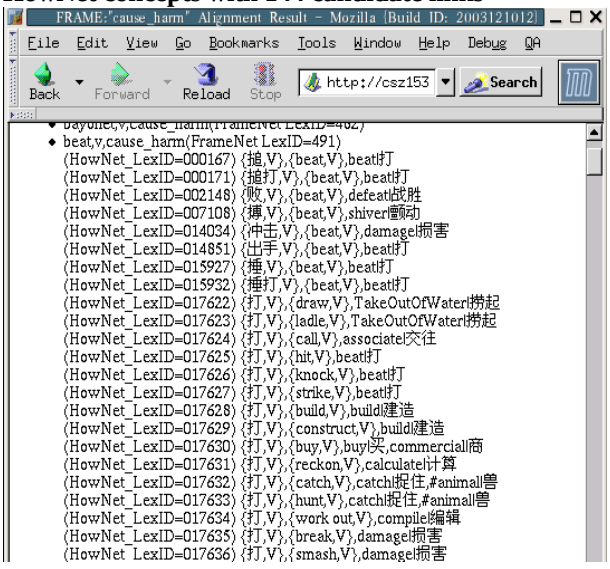
For each lexical entry *l* in FrameNet
  Find translations T1 of *l* in HowNet translations.
  Find translations T2 of *l* in LDC dictionary.
  Combine the T1 and T2 together as T. T= T1∪T2
  Link *l* to all HowNet concepts *LC* whose W_C field is in *T*. *LC*= {c|c.W_C ∈ T}, *c* is any HowNet concept.
For each frame *F* in FrameNet
  Group all the HowNet concepts together *FC* which are linked to the lexical entries in *F*. FC= {c| link(c, l)=true and l ∈ F}.
  Compute the frequency of HowNet categories in *FC*.
  Select the top 3 HowNet categories as valid categories *VA* for frame *F*.
  For each HowNet categories *a*
    If the similarity score between a and one of the top 3 categories is greater than threshold *t*. Sim(a, ta) > *t*, *ta* is any of the top 3 categories.
      Add *a* into *VA*. *VA* = VA∪{a}.
  For each lexical entry *l* in frame *F*
    For each HowNet concept *c* linked to *l*
      If the categories of *c* is not in *VA*
        prune this link.

Figure 1. The algorithm.

3.1 Baseline mapping based on bilingual lexicon

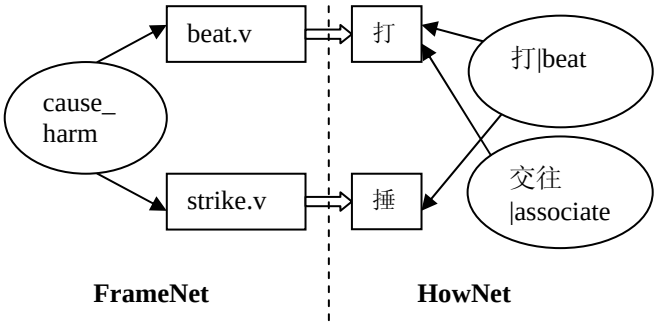
We use the bilingual lexicon from HowNet and LDC dictionary to first create all possible mappings between FrameNet lexical entries and HowNet concepts whose part-of-speech (POS) tags are the same. Here we assume that syntactic classification for the majority of FrameNet lexical entries (i.e. verbs and adjectives) are semantically motivated and are *mostly* preserved across different languages. For example "beat" can be translated into {搥，败，冲击，出手，难倒，骗取，赢，战败…} in HowNet and {打，打败，捣，敲打，赢…} in the LDC English-Chinese dictionary. "beat.v" is then linked to all HowNet concepts whose Chinese word/phrase is one of the translations and the part of speech is verb "v". Figure 2 shows some examples of HowNet concepts that are linked to "beat.v".

Figure 2. Partial initial alignment of "beat.v" to HowNet concepts with 144 candidate links



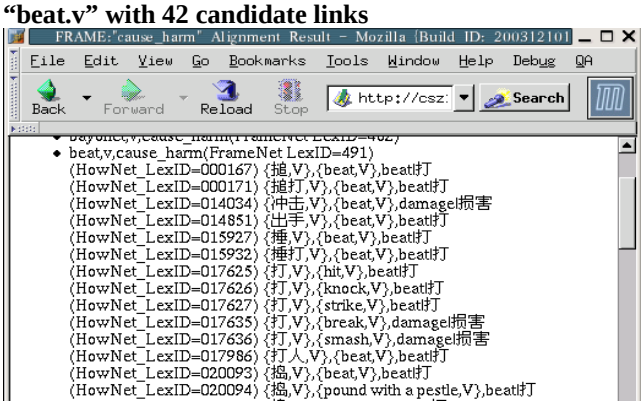
3.2 Disambiguation by semantic contexts in both languages

At this stage, each FrameNet lexical entry has links to multiple HowNet concepts and categories. For example, "beat.v" in "*cause_harm*" frame is linked to "打" in both the "*beat*" category and the "*associate*" category (as in"打电话/make a phone call"). We need to choose the correct HowNet concept (word sense). Many word sense disambiguation algorithms use contextual words in a sentence as disambiguating features. In this work, we make use of contextual lexical entries from the same semantic frame, as illustrated below:



To disambiguate between the above two candidate categories, we make use of the other lexical entries in "*cause_harm*", such as "捶", and their linked categories in HowNet, such as "*beat*" again. Each target HowNet category receives a vote from the candidate links. In our example, "*beat*" receives two votes (from "打" and from "捶"), and "*associate*" only one (from "打"). We choose the HowNet category with the most votes and its constituent concepts to be the valid word sense links to the source FrameNet lexical entry. Consequently, "beat.v" in "*cause_harm*" is linked to all HowNet concepts that are translations of "beat" which are verbs, and which also belong to the HowNet category "beat" (vs. "associate").

Figure 3. Disambiguating HowNet candidates for "beat.v" with 42 candidate links



In our example, Figure 3 shows the top 14 examples of HowNet concepts belonging to two HowNet categories—"beat" and "damage" that are linked to the "cause_harm" frame in FrameNet. Only the concepts in the top N categories are considered as correctly linked to the lexical entries in the "cause_harm" frame. We heuristically chose N to be three in our algorithm.

3.3 Compensating links by HowNet taxonomy structure

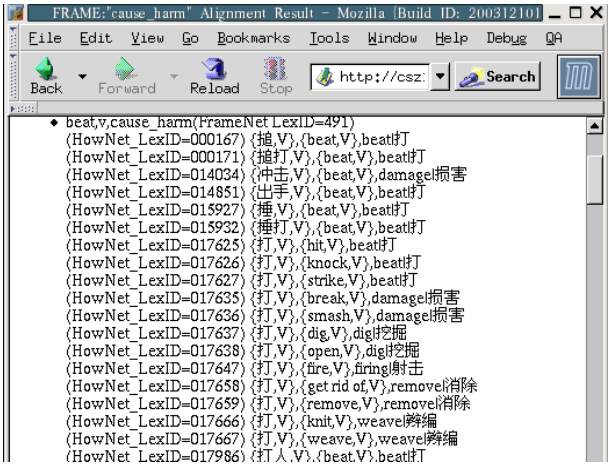
Using frame context alone in the above step can effectively prune out incorrect links, but it also prunes some correct links whose HowNet categories are not in the top three categories but are similar to them. In this next step, we aim to recover this kind of pruned links. We introduce the category similarity score, which is based on the HowNet taxonomy distance (Liu and Li, 2002):

$$\text{Sim(category1,category2)} = \frac{\alpha}{\alpha+\text{d}}$$

Where d is the path length from category1 to category2 in the taxonomy. α is an adjusting parameter, which controls the curvature of the similarity score. We set α =1.6 in our work following the experiment results in (Liu and Li, 2002). If the similarity of category *p* and one of the top three categories is higher than a threshold *t*, the category *p* is also considered as a valid category for the frame.

In our example, some valid categories, such as "firing|射击" is not selected in the previous step even though it is related to the "*cause_harm*" frame. Based on the HowNet taxonomy, the similarity score between "firing|射击" and "beat|打" is 1.0, which we consider as high. Hence, "firing|射击" is also chosen as a valid category and the concepts in this category are linked to the "beat.v" lexical entry in the "*cause_harm*" frame. However, using taxonomy distance can cause erros such as 打 in the "weave" category to be aligned to "beat.v" in the "cause_harm" frame.

Figure 4. Final HowNet candidates for "beat.v" with 54 candidate links



4 Evaluation

We evaluate our work by comparing the results to a manually set golden standard of links for the most ambiguous lexical entries in FrameNet, and use the precision and recall rate as evaluation criteria. To show the lower bound of the system performance, we chose six FrameNet lexical entries with the most links to HowNet concepts as the test set. Since each link is a word sense, these lexical entries have the most ambiguous translations. Such lexical entries also turned out to be mostly verbs. Since the number of lexical entries in a FrameNet parent frame (i.e. frame size) is an important factor in the disambiguation step, we analyze our results by distinguishing between "small frame"s (a frame with less than 5 lexical entries) and "large frame"s. 24% of the frames are "small frames". Results in Tables 2 and 3 have a weighted average of *(0.649*0.24+0.874*0.76)=82%* F-measure.

lexical entry	Parent frame	#candidate HowNet links	#lexical entries in parent frame
beat.v	cause_harm	144	51
move.v	motion	132	10
bright.a	light_emission	126	44
hold.v	containing	145	2
fall.v	motion_directional	127	5
issue.v	emanating	124	4

Table1. Lexical entries test set

lexical entry	Precision best/baseline	Recall best/baseline	F-measure best/baseline
beat.v	88.9/36.8%	90.6/100%	89.7/53.8%
move.v	100/49.2 %	72.3/100%	83.9/66.0%
bright.a	79.1/54.0%	100/100%	88.3/70.1%
Overall	87.1/46.3%	87.6/100%	87.4/52.3%

Table 2. Performance on large frames

lexical entry	Precision step3/step1	Recall best/baseline	F-measure best/baseline
hold,v	22.4/7.6%	100/100%	36.7/14.1%
fall,v	87.0/ 49.2 %	81.1/100%	83.9/66.0%
issue.v	31.1/12.3%	100/100%	47.5/20.3%
Overall	52.1/25.0%	85.9/100%	64.9/40.0%

Table 3. Performance on small frames

	Baseline Alignment	Category Ranking	Category Ranking+ Taxonomy
Precision	36.81%	95.24%	88.89%
Recall	100%	75.47%	90.56%
F-measure	53.81%	84.21%	89.72%

Table 4. Average performance on "beat.v" at each step of the algorithm.

Table 4 shows the system performance in each step of the alignment between the most ambiguous FrameNet lexical entry "beat.v" to HowNet concepts with the final F-measure at 89.72.

5 Conclusion and Discussion

The alignment results can be found at http://www.cs.ust.hk/~hltc/BiFrameNet. Our evaluation shows that our method has achieved an 82% average F-measure in aligning the most ambiguous FrameNet lexical entries to HowNet concepts. This paper describes the first stage in our project towards creating a bilingual English-Chinese FrameNet, by aligning lexical entries between FrameNet and HowNet. The next step is to automatically extract semantically annotated Chinese sentences based on the annotated English sentences in FrameNet, the aligned FrameNet lexical entries, and bilingual corpora. We expect the final bilingual FrameNet will provide a valuable resource for multi-lingual or cross-lingual natural language processing.

Acknowledgment

This work is partly supported by CERG #HKUST6213/02E of the Hong Kong Research Grants Council (RGC).

References

Collin F. Baker, Charles J. Fillmore and John B. Lowe. (1998).The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Hoa Trang Dang, Ching-yi Chia, Martha Palmer, and Fu-Dong Chiou. Simple Features for Chinese Word Sense Disambiguation. In *Proceedings of COLING-2002*, Taipei Taiwan, August 24 - September 1, 2002.

Dong, Zhendong., and Dong, Qiang.(2000). HowNet [online]. Available at http://www.keenage.com/zhiwang/e_zhiwang.html

Bonnie J. Dorr, Gina-Anne Levow, and Dekang Lin.(2002).Construction of a Chinese-English Verb Lexicon for Machine Translation. In *Machine Translation, Special Issue on Embedded MT*, 17:1-2.

Daniel Gildea and Daniel Jurafsky.(2002).Automatic Labeling of Semantic Roles. In *Computational Linguistics*, Vol 28.3: 245-288.

Liu Qun, Li, Sujian.(2002).Word Similarity Computing Based on How-net. In *Computational Linguistics and Chinese Language Processing*, Vol.7, No.2, August 2002, pp.59-76

Mei Jiaju and Gao Yunqi.(1983). tong2yi4ci2ci2lin2. Shanghai Dictionary Press.

Grace Ngai, Marine Carpuat, Pascale Fung.(2002).Identifying Concepts Across Languages: A First Step towards a Corpus-based Approach to Automatic Ontology Alignment". In *Proceedings of COLING-02*, Taipei, Taiwan.