Defining a Methodology for Mapping Chinese & English Sense Inventories

Vicky Tzuyin Lai, Meiyu Chang, Cecily Jill Duffield, Jena D. Hwang, Nianwen Xue, & Martha Palmer
University of Colorado, Boulder
Department of Linguistics
Hellems 290, 295 UCB, Boulder, CO 80309
Vicky.Lai@colorado.edu

Abstract

In this study, we explored methods for linking Chinese and English sense inventories using two opposing approaches: creating links (1) bottom-up: by starting at the finer-grained sense level then proceeding to the verb subcategorization frames and (2) top-down: by starting directly with the more coarse-grained frame levels. The sense inventories for linking include pre-existing corpora, such as English Propbank (Palmer, Gildea, and Kingsbury, 2005), Chinese Propbank (Xue and Palmer, 2004) and English WordNet (Fellbaum, 1998) and newly created corpora, the English and Chinese Sense Inventories from DARPA-GALE OntoNotes. In the linking task, we selected a group of highly frequent and polysemous communication verbs, including say, ask, talk, and speak in English, and shuo, biao-shi, jiang, and wen in Chinese. We found that with the bottom-up method, although speakers of both languages agreed on the links between senses, the subcategorization frames of the corresponding senses did not match consistently. With the top-down method, if the verb frames match in both languages, their senses line up more quickly to each other. The results indicate that the top-down method is more promising in linking English and Chinese sense inventories.

1 Introduction

Currently there exist several lexical resources, whose hierarchical organizations of senses vary. WordNet (Fellbaum, 1998) organizes fine-grained word senses using hierarchical units such as synsets while English Propbank (Palmer, Gildea, and Kingsbury, 2005) differentiates coarse-grained word senses through a comparatively flat subcategorization of predicate argument structures (termed framesets in English Propbank). There are also lexical resources in other languages similar to English ones. The sense organization in Chinese Propbank (Xue and Palmer, 2004) is comparable to that of English Propbank. Linking those resources could broaden the coverage over the senses at the different levels and could potentially improve word alignments predicted by machine learning algorithms. In addition, linking senses cross-linguistically could provide us with a preliminary Interlingua representation system.

In this paper, we explore methods for linking Chinese and English sense inventories using two opposing approaches: creating links (1) bottom-up: by starting at the level of sense group[1] then proceeding to Propbank framesets and (2) top-down: by starting directly with the more coarse-grained frameset levels. We will first describe the pre-existing corpora and the new lexical resources we created in section II. Next, we detail the process by which we select for comparison a set of high frequency communication verbs that are available in both languages' sense inventories. Then, we will describe our two methods of linking English and Chinese sense inventories. The results indicate that the top-down method is more promising.

2 Pre-existing Lexical Resources

2.1 English Proposition Bank

The Penn Proposition Bank, funded by ACE (DOD)

[1] Because of the different methods how those various resources were created, a direct link is unlikely.

, focuses on the argument structure of verbs and provides a corpus annotated with semantic roles, including participants traditionally viewed as arguments and adjuncts. The 1M word Penn Treebank II Wall Street Journal corpus has been successfully annotated and is available via the Penn Linguistic Data Consortium as Propbank I. In addition to the annotated corpus, Propbank provides a lexicon which lists, for each broad meaning of each annotated verb, its "frameset", i.e., the possible arguments in the predicate and their labels and all possible syntactic realizations. An example of one of the framesets for the verb "say" is illustrated in (1).

(1) Frameset say. frame01
Arg0: Sayer
Arg1: Utterance
Arg2: Hearer

"'Well that's odd,' said John of the disappearance of his nose."
Arg1: "Well that's odd"
REL: said
Arg0: John
Arg3-of: of the disappearance of his nose

4,400 framesets were specified for 3,100 verb lemmas. This lexical resource is used as a set of verb-specific guidelines by the annotators, and can be seen as quite similar in nature to FrameNet (Johnson, Fillmore, Petruck, Baker, Ellsworth, Ruppenhofer, and Wood 2002), although much more coarse-grained and general purpose in the specifics. This style of annotation has also been successfully applied to other genres and languages.

2.2 Chinese Proposition Bank

Framesets are also created to support the semantic annotation of the Chinese Proposition Bank. The frameset for a verb are created by examining all of its instances in the Chinese Treebank. Sentences are extracted and organized by the subcategorization frames. The frameset creator then examines these subcategorization frames and determines which frames realize the same set of semantic roles. The posited frameset is then specified with roles. This process is reiterated until all instances of the verb are accounted for[2]. Generally speaking, each frameset postulated this way corresponds to a major sense of the verb. The number of arguments for different senses of a verb may be different, or even when the number of arguments is the same, they may be different arguments. For the 11,765 verbs in the Chinese Treebank, 12,555 framesets are specified. The vast majority of the verbs (11,185 verbs) have only one frameset, 470 verbs have 2 framesets, and 110 of them have three or more framesets.

2.3 English WordNet

WordNet is an on-line lexical database for English organized in accordance with current psycholinguistic theories. WordNet was developed under the direction of George A. Miller. The basic unit is synonym set, or synset, which represents a lexicalized concept. Synsets are comprised of open class words (nouns, verbs, adjectives, and adverbs) and are connected by bi-directional pointers denoting conceptual-semantic and lexical relations, such as synonymy, antonymy, hyponymy, meronymy, troponymy, entailment, etc. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. WordNet has entries for 11,488 verbs.

3 Newly Created Lexical Resources

3.1 DARPA-GALE OntoNotes and English Sense Inventories

The OntoNotes project (BBN, Penn, Colorado and ISI) focuses on a domain independent representation of literal meaning that includes predicate argument structure, word sense, ontology linking, and coreference. Studies have shown that these can all be annotated rapidly and with better than 90% consistency. Once a substantial and accurate training corpus is available (target date January 2007), trained algorithms can be developed to predict these structures in new documents. This process

[2] A consequence of this frame file creation methodology is that it is possible that not all framesets of this verb were accounted for. The worst case scenario would be that these framesets need to be reorganized when new data need to be annotated. So far, this worst case scenario rarely happens and new framesets has been added without affecting the existing framesets.

begins with parse (TreeBank) and semantic (Propbank) structures, which provide normalization over predicates and their arguments. Word sense ambiguities are then resolved, with each word sense also linked to the appropriate node in the Omega ontology (Philpot, Hovy, and Pantel, 2005), which provides for storage and inheritance of relevant axioms. Coreference is also annotated, linking together the entity mentions that are propositional arguments. The sense inventories being used for the most frequent 700 verbs and 1100 nouns in a 300K subset of the WSJ Propbank are based on coarse-grained groupings for WordNet fine-grained senses first developed for the Senseval2 annotation (Palmer, Babko-Malaya, and Dang, 2004; Palmer, Dang, and Fellbaum, 2006), in which these more coarse-grained senses led to improved inter-annotator agreement (ITA) and system performance, with current best performance at 86.7% accuracy on OntoNotes data (Chen, Schein, Ungar, and Palmer, 2006).

The process of grouping WordNet senses for verbs has been continued in Colorado by four native English speakers. During our sense grouping process, these linguists (henceforth, “groupers”) cluster fine-grained sense distinctions listed in WordNet 2.1 into more coarse-grained groupings. These rough clusters of WordNet entries are based on speaker intuition, as well as other resources, including PropBank, VerbNet and online dictionaries (Palmer, et. al., 2005, Kipper et al., 2006). The goal is to create senses that are coarse-grained enough to allow annotators to achieve a 90% agreement rate, but fine-grained enough to allow the highest possible amount of information to be extracted.

While annotators have little trouble tagging text with verb senses that fall neatly into intuitive categories, many verbs have fine-grained WordNet senses that fall on a continuum between two distinct lexical usages. In such cases, syntactic and semantic aspects of the verb and its arguments help groupers cluster senses in such a way that annotators can make consistent decisions in tagging the text. Groupers have found syntactic frames, such as those defining VerbNet classes, to be useful in establishing boundaries between sense groupings that are easily understood by all annotators. Focusing on certain syntactic constructions, such as obligatory prepositional phrases, that characterize the alternations in VerbNet classes, are particularly helpful for annotators. When senses of a verb have similar syntactic frames, and usages fall along a continuum between these senses, semantic features of the arguments (including [+/-attribute], [+/-patient], and [+/-locative]), or less often, of the verb itself, can clarify these senses and help groupers draw clear distinctions between them. However, verb features proved less useful than features of nominal arguments, and annotators not familiar with linguistic theory found them to be confusing. Therefore, they are now rarely used to label sense groupings. Such concepts, when used, are more likely to be described in prose commentary for the sake of the annotators. Sense groupings are ordered according to saliency and frequency. Groupers also provide the annotators with simple example sentences from WordNet as well as syntactically complex and ambiguous attested usages from Google search results. These examples are intended to guide annotators faced with similar challenges in the data to be tagged.

A grouping is tested by running it through a sample annotation task that is tagged by two annotators. If the ITA is 90% or above, groupings are approved for actual annotation. Otherwise, the groupings are revised and sent through another round of sample annotation until the desired ITA is reached (85% for second round; 80% for third round). If by the third round of sample annotation the verb cannot achieve the desired ITA, it is discussed by groupers and pushed directly into the actual annotation. Those final groupings for the actual annotation are also tagged by two annotators. All final disagreements are then resolved by one of the three adjudicators.

See appendix for an example of grouping.

3.2 Chinese Sense Inventories

The framesets in Chinese Propbank provide coarse-grained senses as the basis for further refinement of sense distinctions. While framesets give the numbers of arguments and their argument roles for each verb, the sense groups further give specific properties of each argument, e.g., whether a theme is a physical entity or a person might split the senses that originally share the same frameset. Another criterion for Chinese sense grouping is the semantic features, such as agentivity and causativity, e.g., for the verb sheng, the Chinese Propbank has one frameset for both “to give birth to” and “to be born”, which are split into two senses in our sense inventory. After the senses of each verb are defined, they are put into test during the trial annotation. Further revision is done when the target accuracy, 90% of inter-annotator agreement, cannot be achieved. Usually during the sense revision process senses that are too fine-grained for the annotators to tag consistently are merged. So far, we have built 300 sense inventories for the most frequent polysemous verbs from the Chinese corpus and successfully annotated them with 90% or higher accuracy.

4 Selecting Verbs

We chose to examine the following communication verbs in English and Chinese, because 1) they exist in the 700 English verbs that have undergone the sense grouping process and the 300 Chinese verbs that have sense inventories; 2) these verbs are highly frequent and polysemous in the corpora (see Table 1); and 3) each of these verbs in one language has a corresponding counterpart in the other language.

English	Frequency	Chinese	Frequency
Say	10,503	*shuo*	1,776
ask	338	*biao-shi*	441
talk	133	*jiang*	62
speak	69	*wen*	59

Table 1: Frequency of the selected verbs

5 Using English Sense Groups and Chinese Sense Inventories as a Starting Point

In our first method, we used sense groups as a starting point. A native English speaker fluent in Chinese and a native Chinese speaker fluent in English examined the senses of the selected verbs in their native languages and sought out the corresponding senses in the other language separately. Once the senses were distinguished as in (2), the verb-specific framesets for each of the senses were examined as in (3).

(2) shuo.sense01:qing shuo de man yi dian er
please speak MOD slowly one bit ASP
"Please speak slower"
speak.sense01: They spoke in hushed whispers …
talk.sense01: The patient was talking in his sleep.
say.sense01: Will you please say grace for us?
say.sense05: Say 'she sells sea shells by the sea shore' fast.

(3) Framesets that correspond to the senses in (1):

shuo.sense01	speak.sense01	talk.sense01
shuo.frame01	speak.frame01	talk.frame01
Arg0: agent	Arg0: talker	Arg0: talker
Arg1: thing said	Arg1: subject	Arg1: subject
Arg2: source	Arg2: hearer	Arg2: hearer

say.sense01	say.sense05
say.frame01	say.frame01
Arg0: sayer	Arg0: sayer
Arg1: utterance	Arg1: utterance
Arg2: hearer	Arg2: hearer
Arg3: attributive	Arg3: attributive

We can see from (3) that Arg0 and Arg1 match, but "ARG2: text quoted" of shuo.f01[3] in Chinese does not match any of those English verb frames. Furthermore, there is no "ARG3: Attributive" of say.f01 in Chinese. We could argue for this particular instance that the attributive argument in English is often tagged as one of the functional tags in Chinese Propbank. However, the results from other verbs using this method show that although speakers of both languages agreed on the links between senses, the subcategorization frames of the corresponding senses did not match consistently.

6 Using English Propbank and Chinese Propbank as a Starting Point

Our second method examines the verb frames first. We listed the number of arguments for all selected verb and their frames in English and Chinese as in Table 2. For verbs that have the same number of arguments, we examined their verb-specific frames and example sentences. Only verbs with matching frames were compared. Following is a step-by-step illustration.

The first step was to examine the verbs with the same number of arguments and their individual argument roles. Looking at the framesets for say.f01, ask.f01, ask.f03, and biao-shi.f01, which have four arguments, we decided that the English say.f01 and the Chinese biao-shi.f01 have the most common argument roles.

[3] Frame 01 is abbreviated to f01 and sense01 is abbreviated to s01 in the rest of the paper.

English	Framesets	Args	Chinese	Framesets	Args
say	say.f01	4	*shuo*	shuo.f01	3
				shuo.f02	2
				shuo.f03	2
				shuo.f04	2
				shuo.f05	2
				shuo.f06	1
ask	ask.f01	4	*biao-shi*	biao-shi.f01	4
	ask.f02	3		biao-shi.f02	2
	ask.f03	4			
speak	speak.f01	3	*jiang*	jiang.f01	2
	speak.f02	1		jiang.f02	2
	speak.f03	2			
talk	talk.f01	3	*wen*	wen.f01	3
	talk.f02	3			

Table 2: Number of arguments of the verbs

(4) Frameset say.f01:
Arg0: sayer
Arg1: utterance
Arg2: hearer
Arg3: attributive

Frameset ask.f01
Arg0: asker
Arg1: question
Arg2: hearer
Arg3: attributive

Frameset ask.f03
Arg0: seller
Arg1: commodity
Arg2: buyer
Arg3: asking price

Frameset biao-shi.f01
ARG0: agent
ARG1: message expressed
ARG2: party receiving
ARG3: topic of message

The second step was to check the example sentences under say.f01 and biao-shi.f01 for matching framesets. In (1), we found that the "ARG3: attributive" of say.f01, the "weird attributive usage", corresponds to the concept of "topic of message", the required argument in Chinese in (5).

(5) Frameset biao-shi.f01
Tanaixing jintian jiu ce shi dui xinhuashe jizhe
Propername today toward this matter to xinhua newswire reporters

biaoshi, zhe ge dianchang ruguo chenggong …
say, this classifier power plant if succeed ...

"Aixing Tan said to the Xinhua Newswire reporters with regard to the matter today, if the power plant succeeds …"
ARG0: Aixing Tan; ARG1: if the power plant succeeds; ARG2: to the Xinhua Newswire reporters; ARG3: With regard to this matter

The third step was to look up the specific senses for say.f01 and biao-shi.f01 separately in the English and Chinese sense inventories. The sense inventory that matches say.f01 is "say.s01: an agent expresses or communicates a concept through words". The sense inventory that matches biao-shi.f01 is "biao-shi.s02: to indicate, to signify (followed by an explanation)". Thus, we determined that the senses are similar to native speakers in both languages.

The last step was to look at the example sentences listed under say.s01 and biao-shi.s02. We found that the example sentences in biao-shi.s02 could be translated with say.s01.

From the communication verbs with four arguments, we found that verbs with similar subcategorization frames in two languages also have similar senses at the level of the English and Chinese sense groups. However, we found that "biao-shi.s02" can always be translated to "say.s01", but not vice versa. It may suggest that the English sense group "say.s01" is not quite at the right level to correspond to the Chinese sense inventory "biao-shi.s02" and could be grouped into several sense groups representing distinctive senses[4]. Lastly, we also looked at verbs with three arguments and two arguments. The results indicated that if the verb frames match in both languages, their senses line up more quickly to each other.

Conclusion

According to the results from the two methods, using verb frames in English and Chinese Propbanks as a starting point is a promising approach to mapping Chinese and English Sense Inventories. Using senses in both languages that are described as similar by native speakers does not necessarily result in similar argument structures. In addition, matching the number of arguments is useful, but verb-specific frames should be consulted for accurate frame matching.

In our future work, we would like to continue to explore these two methods. Our hypothesis is that using argument structures as a beginning point will allow us to focus on likely mappings between

[4] This may further suggest that when grouping senses in one language, it could be helpful to consult the sense organization in a second language. This could be acceptable fine if the sole purpose of the sense inventories is a bilingual lexicon, but it might introduce a bias that could interfere with other purposes.

Chinese and English verbs in a more efficient and consistent manner. One of our research questions is to devise criteria for evaluating these methods. We would also like to explore any patterns in argument structure that may be used to describe the relationships between verb senses that speakers describe as corresponding. Finally, we would like to test these correspondences by looking at automatic word alignment of parallel corpora. If we find good overlaps between the automatic word alignment and manual mappings, then we can do mutual bootstrapping, which could improve word alignment data based on sense tagged corpora using the manual mapping and vice versa, extend the manual mapping coverage semi-automatically using word aligned corpora.

Acknowledgements

This work is supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

Appendix. English Sense Inventory for "talk".

Sense Group	Description and Commentary	WordNet 2.1 Senses	Examples
Talk.s01	To converse, speak, or use language	1, 2, 3 talk about 1,2 talk of 1 talk over 1 talk shop 1 talk turkey 1	We need to *talk*. Before long, your computer will be able to *talk* back to you. Mary's baby hasn't started *talking* yet.
Talk.s02	To spill the beans: NP[+agent] TALK (PP[+content]) (PP[+recipient]) Note: implies a subject matter that was disclosed to someone.	4, 5	Be careful what you say around here because the secretary does *talk*. The mob put a hit out on the bookie for *talking* to the police.
Talk.s03	To lecture formally to an audience: NP[+agent] TALK [+duration] (PP[+content]) (PP[+recipient])	6	John will *talk* about recent theories of dark energy at the Physics Symposium. Bob gets nervous when *talking* in front of a large audience.
Talk.s04	To persuade, convince or influence someone: NP[+agent] TALK NP[+patient] INTO/OUT OF NP/COMP[+goal]	talk into 1 talk out of 1	John tried to *talk* me into joining his crew for the Bermuda race. We couldn't *talk* them out of painting their house bright orange.
Talk.s05	To belittle or condescend towards: NP[+speaker] TALK DOWN PP[+recipient]	talk down 1, 2	Don't *talk* down to the cleaning lady.
Talk.s06	To answer impertinently: NP[+speaker] TALK BACK PP[+recipient] Note: This sense does not refer to simple responses (sense 3)		Don't *talk* back to your mother! It's dishonorable to *talk* back to a cop.
Talk.s07	To direct or control to the ground: NP[+agent] TALK DOWN NP[+patient]	talk down 3	The control tower *talked* down the plane whose pilot had fallen ill.
Talk.s08	To negotiate the terms of an agreement: NP[co-agent] TALK TERMS (PP[+goal])	talk terms 1	Are they *talking* terms about this house?

References

Chen, J., Schein, A., Ungar, L., & Palmer, M. (2006). An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation. In Proceedings of the Human Language Technology conference of the NAACL, Main Conference (pp 120--127), New York, NY: ACL

Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: The MIT Press.

Johnson, C., Fillmore, C., Petruck, M., Baker, C., Ellsworth, M., Ruppenhofer, J., & Wood, E. (2002). FrameNet: Theory and practice. Version 1.0, http://www.icsi.berkeley.edu/framenet/.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer. (2006). Extensive Classifications of English Verbs. Proceedings of the 12th EURALEX International Congress. Turin, Italy

Palmer, M., Dang, H., & Fellbaum, C. (2006). Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically. To appear in Journal of Natural Language Engineering.

Palmer M., Gildea D., & Kingsbury P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics, 31(1), 71--106.

Palmer, M., Babko-Malaya, O., & Dang, H. (2004). Different Sense Granularities for Different Applications, In Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems, at HLT/NAACL-04, Boston, MA.

Philpot, A., Hovy, E., Pantel, P. (2005). The Omega Ontology. In IJCNLP workshop on Ontologies and Lexial Resources (OntoLex-05) Jeju Island, South Korea.

Xue, N. & Palmer, M. (2004). Annotating the Propositions in the Penn Chinese Treebank. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing (pp 47--54), Morristown, NJ: ACL.