Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese

Mari Broman Olsen, Bonnie Dorr, and Scott Thomas
UMIACS
University of Maryland
College Park, MD 20742
phone: +1 (301) 405-6754
fax: +1 (301) 314-9658
{molsen,dorr,scthmas}@umiacs.umd.edu

Abstract

This paper describes a refinement to our procedure for porting lexical conceptual structure into new languages. Specifically we describe a two-step process for creating candidate thematic grids for Mandarin Chinese verbs, using the English verb heading the VP in the subdefinitions to separate senses, and roughly parsing the verb complement structure to match to our thematic structure templates. The procedure is part of a larger process of creating a usable lexicon for interlingual machine translation from a large on-line resource with both too much and too little information necessary for our system.

Keywords: lexical conceptual structure, lexicon acquisition, machine translation, polysemy

1 Introduction

In previous work on Spanish and Arabic [Dorr *et al.*, 1997, Dorr, 1997a], we reported the results of an acquisition process for verb databases in new languages, using automatic assignment of candidate thematic structure templates ("theta grids") and hand verification of the output. A substantial number (39%) of the hand modifications to the output of the acquisition program were eliminations of theta grids that were assigned to target language verbs because the English glosses were polysemous.

This paper reports on acquisition of a Mandarin Chinese verb database from an on-line resource ten times as large as those we had worked with for Spanish and Arabic (600k, rather than 60k). The procedure is part of a larger process of creating a usable lexicon for interlingual machine translation from a large on-line resource with both too much and too little information necessary for our interlingual machine translation system [Dorr, 1997b, Hogan and Levin, 1994].

We attempt to reduce the amount of labor involved in the hand-correction scheme through creating separate entries for polysemous verbs and parsing their English definitions with respect to our database of thematic grids. Specifically, we find the head verb of each subgloss (i.e. 'run' in 'run away calling for help') and assign candidate theta grids associated with *run* in English. Furthermore, we eliminate some candidate grids in favor of those matching the complement pattern in the gloss — the prepositional phrase 'away', above. This procedure results in a reduction of 11% — 15,565 — possible candidate thematic grids that will not have to be evaluated by hand. Furthermore, the separation of sense allows candidate grids to be evaluated with respect to a particular verb sense. Noise that is introduced from polysemy in the English glosses (e.g. 'run a machine' and 'run a race') may therefore be limited to the relevant verb sense, rather than combined in a "bag of grids" attached to a verb.

This reduction has been accomplished without substantive loss of relevant definitions, as evaluated in a preliminary task, assigning theta grids to verbs in 10 sentences from a corpus of 10 Xinhua articles. We will be able to have further evaluation on the 263 verbs from this corpus, which appear in 423 different senses (average 1-2 senses per verb), as well as on the full set of verbs, as the hand-checking progresses.

We see this work as a step in the direction suggested by Dorr, Marti and Castellon [1997], who observe that "The amount of time required for the hand-verification process [for Spanish] would be greatly reduced if the issue of polysemy had been addressed earlier in the process." That research generated 18353 candidate theta grids, representing 3623 verbs in the initial Spanish-English lexicon. Of that, 3025 entries were verified as correct (16.5%), and 15328 (83.5%) had to be modified in some way. There were 6082 deletions of entries, 334 reclassifications (resulting in changes of entries) and 6295 refinements of entries. The refinements included 3648 deletions of non-applicable entries, 2747 changes to prepositions, optional roles made obligatory, etc. 2617 entries (955 verbs) deleted due to rarity of usage and/or disjointness with respect to WordNet concepts 1213 new entries added (1092 verbs not in initial database). That is, there were a total of 9730 deletions, representing 63.5% of the required modifications, and 53% of the total number of candidate grids.

Thus, an automatic process that reduces the number of deletions in a principled way would substantially reduce the hand-correction process. We first describe the role of the theta grids in our system. We then describe our lexicon acquisition procedure, with respect to the verbs, detailing, at every step, how we attempted to deal with polysemy and overgeneration of theta grids.

2 Thematic Structure: Theta Grids

Thematic structure serves as the interface between the syntactic component (parsing) and the lexical-semantic component, the Lexical Conceptual Structure (LCS). In particular, the assignment of a set of thematic roles to a structure allows a unique interlingual LCS structure to be created, as in the following pair, representing a sentence like *Derek filled the bucket with water*. The thematic structure and LCS gloss roughly as 'Agent causes something to be VERBed with something.'

```
:THETA_ROLES "_ag_th,mod-poss(with)"
:LCS (cause (* thing 1) (be ident (* thing 2) (at ident (thing 2) (!!-ed 9)))
      ((* with 15) poss (*head*) (thing 16)))
```

The theta grids therefore map directly into LCS variables, including ag(ent), th(eme), exp(eriencer), and goal. In our grid scheme, obligatory theta roles are preceded by an underscore, and optional roles with a comma, e.g. `_ag,th` for *John ate (lunch)*.

Verbs in a language can take more than one theta grid. The set of theta grids permitted by a verb allows it to be grouped with others in a class taking the same semantic structures, represented by the LCS [Levin, 1993]. For example, verbs like *fill, carpet, cloak* and *plug* allow the grid `_ag_th,mod-poss(with)`, as in *Derek filled the bucket with water*, and not ∗*Derek filled the water into the bucket* Verbs like *pour, drip, dribble*, and *slosh* allow the latter pattern, but not the former: *,ag_th,src(),goal()*, as in *Derek poured the water into the bucket*, but not ∗*Derek poured the bucket with water*. The grids therefore group verbs by "semantic structure" [Levin and Hovav, 1995]. In contrast to "semantic content," a term used to label idiosyncratic aspects of verb meaning, semantic structure is important in determining syntactic patterning within and across languages [Dorr and Oard, 1998, Grimshaw, 1993, Pinker, 1984, Pinker, 1989].

3 Verb Selection

The assignment of thematic grids is one step in the creation of a lexicon from a large (600k entries) machine readable Chinese-English dictionary. The dictionary was compiled by hand, by the Chinese-English Translation Assistance (CETA) group from some 250 dictionaries, some general purpose, others domain-specific or bilingual (Russian-Chinese, English Chinese, etc.). The CETA group, started in 1965 and continuing into the present decade, was a joint government-academic project. The machine readable version of the resulting *Optilex* dictionary is licensed from the MRM corporation, Kensington, MD.

As is often the case, the information required by our machine translation system is not directly encoded in Optilex, including part of speech tagging. We identified the verbs by a simple process. We parsed the DEF: field in each `optilex.dict` entry derived from 20 of the more general of those dictionaries, using using regular expressions to find English glosses beginning with the infinitival 'to'. If so, the whole entry is used to generate as many new verb entries as there are verbs in the entry's DEF: field. As an example, the excerpt from following `optilex.dict` entry has 4 entries in its DEF: field. PY gives the Pinyin representation, and the DEF: field is the English gloss. Optilex includes other fields not listed, including HWT and HWS fields for Chinese pictograms, STC for the Standard Telegram Code, and REF: for the dictionaries the entry came from.

... PY : bian1 ta4

DEF: 1. to whip, to flog 2. <fig> to chastise, to castigate ...

After processing, each definition has a single entry in the verb-entries.txt file:

(1) ... DEF: to castigate ...

(2) ... DEF: to chastise ...

(3) ... DEF: to flog ...

(4) ... DEF: to whip

Besides missing information, the dictionary contained additional information that needed to be eliminated. Some of the 250 resources used to create the dictionary were very domain-specific, including, for example, *Collier's North China Colloquial Collection*, a publication listing many regionalisms not observed anywhere else in China, and the *Faxue Cidian*, a dictionary of legal terms from Shanghai. We eliminated many archaic and technical verbs by eliminating verbs identified by Optilex as derived from these sources.[1]

Nevertheless, entries varied widely in specificity, from the general verbs (and other words) to the extremely specific, as in the examples below.

(5) po4_shi3 compel 迫使

(6) po4_shi3 force 迫使

(7) ben1_pao3 run 奔跑

(8) zou3 walk 走

(9) chu1_kou3 speak[2] 出口

(10) bi1_gong1 force_the_sovereign_to_abdicate (th = sovereign) (prop = to_abdicate) 逼宫

(11) ben1_zou3_xiang1_gao4 run_around_spreading_the_news (mod-loc = spreading_news) 奔走相告

(12) ci3 walk_on_the_ball_of_the_foot ◆[3]

(13) chui1_xu1 speak_in_favor_of_somebody_in_exaggerated_terms 吹嘘

4 Pairing Verbs and Candidate Thematic Grids

4.1 English glosses

Next, we needed to assign thematic grids to the verbs. We estimated that creating thematic grids by hand would take an estimated 6 person months for a lexicon on the order of 60k [Dorr *et al.*, 1997]. As mentioned above, for the Arabic and Spanish lexicon, we created candidate thematic grids by pairing target language words with the thematic grids associated with their English gloss, with hand correction over a period of two weeks.

[1]We included verbs from the following sources: BF Chinese-English Dictionary 1978, BE same as BE but Chinese-Chinese 1978, AR Atlas of the PRC 1977 (for Chinese placenames), AO Gazetteer of the PRC (also for Chinese placenames), BQ extra new entries from the first two above BE and BF CJ standardized FBIS translations of Chinese communist terms, CM specialized terms extracted from Mao's works, CU Hong Kong glossary of Chinese communist terms, EJ 1981 idiom dictionary, EK 1982 idiom dictionary, FA Foreign Exchange terms 1963, IP International political economics glossary 1980, IQ Beijing social sciences academy economics terms 1983, NA world place names 1981, PP primary political economics glossary 1956, TM McGraw-Hill general scientific and technical dictionary 1963, VF Lin Yutang's dictionary 1972, VT 1973 Beijing foreign exchange glossary, WB Liang Shih-ch'iu's traditional dictionary 1973, YG Stanford's dictionary of Chinese communist terms 1973.

[2]As in 'to speak ill of someone.' This meaning is the first listed, although John Kovarik (p.c.) claims it is less popular than at least two others, including 'exit', as in exit signs.

[3]The diamond means no glyph mapping is available for the character code.

We did the same initial step for Chinese, as well. However, as described above, the senses had already been separated into different entries. We thus had a candidate theta grid set paired with a specific sense of a verb. The file Chinese-grids was created by first matching the main verb of the English glosses to one or more entries in the English-grids file.

Separating polysemous entries helps us here, since not all grids are associated with all meanings of the verb. For example, a wide range of grids is available for the run verbs.

(14) 26.3 `_ag` 持 chi2 run

(15) 26.3 `_ag_ben_th` 持 chi2 run

(16) 26.3 `_ag_th,ben(for)` 持 chi2 run

(17) 47.5.1 `_ag,mod-loc()` 持 chi2 run

(18) 47.5.1 `_loc_th` 持 chi2 run

(19) 47.5.1 `_th_loc()` 持 chi2 run

(20) 47.7 `_th_goal()` 持 chi2 run

(21) 47.7 `_th_src(from)_goal(to)` 持 chi2 run

(22) 51.3.2 `_ag` 持 chi2 run

(23) 51.3.2 `_th,src(),goal()` 持 chi2 run

In contrast, a relatively small number is available for other meanings of this character. In previous work, all grids were associated to the single entry, with hand-separation of senses necessary, and the opportunity for human error great, with humans deleting or retaining grids depending on which sense of the verb they had in mind.

(24) 31.2 `_exp_perc,mod-poss(in)` 持 chi2 support

(25) 47.8 `_th_loc` 持 chi2 support

In the case at hand, it turns out that 持 chi2 means 'run', as in 'run a business' or 'run a machine', whereas the theta grids were derived from the motion verb *run* in English. Should the grids prove inappropriate in the hand-verification stage, they can be deleted without affecting the entries glossed 'support'. In previous work, the checker was presented with a "bag of grids", without a link to a specific meaning.

4.2 Parsing the Definition

After assigning a candidate set to each gloss, we matched various phrases in each gloss to the theta roles in the grid entries. This permits some automatic modification of the grids that in earlier work had been done by hand. For instance, the gloss 'to force the sovereign to abdicate' matches the English-verb database entry with roles `_ag_th,prop(to)` for the verb *to force*; that is, the English verb takes an agent, theme, and optionally a propositional element, the latter matching "to V ...".

Thus an entry in the Chinese-grids file is created that reads, in part, `... _ag ... (th = sovereign) (prop = to_abdicate)`. That is, the matched theta roles are added, with their thematic assignments, to the end of the entry, resulting in 11,360 distinct theta role assignments. In some cases the original theta-roles list becomes empty; in which case it appears as `_0`, which is the theta grid for verbs with no semantic arguments, such as *rain* in English *It's raining*.

Similarly, PPs in a gloss are matched to a grid element, if possible, and that grid element is removed from the grid. For roles with an unspecified preposition, we heuristically assigned certain roles to certain preps, namely:

from: `src` or `instr`

for: `purp`

with: `instr` or `mod-poss`

without: `mod-poss`

into: `goal`

to: `goal`

against: `goal`

under: `mod-loc`

around: `mod-loc`

along: `mod-loc`

We observed that prepositional phrases with prep = 'of' seem to almost always attach to the preceding phrase, rather than appear as an argument—at least in these glosses. Thus we ignored the possibility of of PP as an argument, always making it a part of the argument that precedes it.

Adverbs, in positions where they typically modify the verb (rather than an adjective)—that is, near the verb or at the end of the gloss—become MANNER components. ('Typical' was determined by looking at the results for these particular glosses.)

A gloss that ends with a dangling preposition was taken as a sign that, where the English verb takes a PP, the Chinese verb fills the same role with a bare NP argument. Thus the parentheses are removed from the grid for that role.

These actions all have the effect of matching elements in the gloss to elements in the grid, and eliminating these elements from the grid, thus reducing the number of grids that have to be hand-checked. We also deleted certain entries entirely, prior to hand checking. In particular, if the set of theta roles lexicalized in the Chinese verbs by an English gloss (which may be the empty set) for one entry of the (polysemous) verb is a proper subset of that for another entry of that verb, the corresponding grid is discarded. This allowed a 11% reduction in the number of entries that needed to be hand checked.

For example, a Chinese verb glossed as *fill* would pull in the theta grid discussed above, shown in the following entry from our theta grid database. (The initial number is the verb class, from Levin [1993].

(26) 9.8 . `_ag_th,mod-poss(with)` ◆漫 mi2_man4 fill

But a verb which incorporates the *with* element in its meaning would not allow that element to appear as another argument in the complement structure, so we automatically eliminate the `mod-poss` role from the grid.

(27) 9.8 . _ag_th 填土 tian2_tu3 fill_in_with_earth (mod-poss = earth)

Similarly, this heuristic allowed four potential candidates to be reduced to one, for the following verb.

(28) Remaining:
29.3 _ag_th 报告 bao4_gao4 make_known (pred = known)

(29) Suppressed:
26.1 _ag_ben_th,instr() 报告 bao4_gao4 make_known
26.1 _ag_th,instr(),ben(for) 报告 bao4_gao4 make_known
26.1 _ag_th_pred(into),ben(for) 报告 bao4_gao4 make_known

That is, in class 29.3 PRED was incorporated in the verb meaning, whereas nothing was incorporated for the other grids. So for this gloss, only the 29.3 grid is used—the empty set being a proper subset of a non-empty set.

This is comparable to what was done by hand for Spanish: In the case of *escribir*, two theta grids were automatically assigned, _ag_th,goal (as in *He wrote his name on the photo*) and _ag_th (as in *He wrote his name*). The latter was left in the database since it provides the most basic thematic requirements for the verb.

5 Results

Using this metric, 15565 thematic grids were eliminated, representing 11% of the total number of candidate thematic grids. We are beginning the process of hand-evaluation of the theta grids, beginning with the verbs in 10 articles from Xinhua, comparable to the Wall Street Journal in content. For the verbs in 10 articles from Xinhua, 124 grids were suppressed for 47 verbs (29 classes) leaving 3041 grids, for 263 verbs (characters, rather than definitions). Number of grids assigned to a given verb (a character set) average 11.6, and range from 0 (for verbs for which a grid cannot be found in our current database) to 183. A set of 51 theta grids were generated for the 13 verbs in ten sentences from these articles. Chinese speakers deleted 17 grids, or 33.3%. Although these results are a tiny subset of the full verb lexicon, this figure compares favorably to the 53% deletion required of the Spanish data. Importantly, none of the relevant grids had been discarded by our algorithm.[4]

As the hand-verification work progresses, we will evaluate the results on a broader scale, tracking necessary modifications to the full set of theta grids for the 120k verb senses. We predict, for example, that the number of deletions should be less than that for the Spanish.

6 Conclusions and Future Work

We have described a procedure for automatically reducing the amount of hand-checking necessary for building the thematic grid structure for verbs in Chinese. We anticipate that this procedure will save us time over our original checking procedure. The latter, in turn reduced the amount of time required to create thematic structure from 6 person months (for a lexicon with 60k entries and 3-4k verbs) to approximately two weeks of hand verification. The time savings for our project is

[4]The copular grid for the verb *shi4* was added to the set, using a grid assigned to other copular verbs, namely *wei2* and *zuo4*. Somewhat surprisingly, the absence of the copular grid in our candidate set resulted from an absence in Optilex of the copular meaning for that verb.

even more imperative, since we have some 150k verb-sense entries. This procedure provides further streamlining for the process of acquiring large-scale lexica for NLP applications with non-optimal on-line resources.

In addressing the polysemy problem in this context, we have, as a side-product, produced a sense-to-syntax mapping, tying verb senses of a particular character to a set of grids representing syntactic as well as semantic structure. This resource, in turn, could be used not only for machine translation, but for testing and applying word sense disambiguation algorithms for Chinese.

Acknowledgments

This work has been supported, in part, by NSA Contract MDA904-96-C-1250. The second author is also supported by DARPA/ITO Contract N66001-97-C-8540, Army Research Laboratory contract DAAL01-97-C-0042, NSF PFF IRI-9629108 and Logos Corporation, NSF CNRS INT-9314583, and Alfred P. Sloan Research Fellowship Award BR3336. We would like to thank members of the following lab groups at Maryland: Computational Linguistics and Information Processing (CLIP), and Language And Media Processing (LAMP), particularly Galen Wilkerson for his implementation and description of verb selection, and John Kovarik, on loan from the Department of Defense.

References

[Dorr and Oard, 1998] Bonnie J. Dorr and Douglas W. Oard. Evaluating resources for query translation in cross-language information retrieval. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.

[Dorr *et al.*, 1997] Bonnie J. Dorr, Antonia Marti, and Irene Castellon. Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the MT Summit Workshop on Interlinguas in MT*, San Diego, CA, October 1997.

[Dorr, 1997a] Bonnie J. Dorr. Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring. In *Proceedings of the ACL Fifth Conference on Applied Natural Language Processing (ANLP)*, pages 139–146, Washington, DC, 1997.

[Dorr, 1997b] Bonnie J. Dorr. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–322, 1997.

[Grimshaw, 1993] Jane Grimshaw. Semantic Structure and Semantic Content in Lexical Representation. unpublished ms., Rutgers University, New Brunswick, NJ, 1993.

[Hogan and Levin, 1994] Chris Hogan and Lori Levin. Data Sparseness in the Acquisition of Syntax-Semantics Mappings. In *Proceedings of the Post-COLING94 International Workshop on Directions of Lexical Research*, pages 153–159, Nicoletta Calzolari and Chengming Guo (co-chairs), Tshinghua University, Beijing, 1994.

[Levin and Hovav, 1995] Beth Levin and Malka Rappaport Hovav. The Elasticity of Verb Meaning. In *Proceedings of the Tenth Annual Conference of the Israel Association for Theoretical Linguistics and the Workshop on the Syntax-Semantics Interface*, Bar Ilan University, Israel, 1995.

[Levin, 1993] Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, IL, 1993.

[Pinker, 1984] Steven Pinker. *Language Learnability and Language Development.* MIT Press, Cambridge, MA, 1984.

[Pinker, 1989] Steven Pinker. *Learnability and Cognition: The Acquisition of Argument Structure.* The MIT Press, Cambridge, MA, 1989.