LAMP-TR-060
CS-TR-4205
UMIACS-TR-2001-01

Building a Chinese-English Mapping Between Verb Concepts for Multilingual Applications

Bonnie J. Dorr, Gina-Anne Levow, Dekang Lin

Language and Media Processing Laboratory
Instititue for Advanced Computer Studies
College Park, MD 20742

Abstract

This paper addresses the problem of building conceptual resources for multilingual applications. We describe new techniques for large-scale construction of a Chinese-English lexicon for verbs, using thematic-role information to create links between Chinese and English conceptual information. We then present an approach to compensating for gaps in the existing resources. The resulting lexicon is used for multilingual applications such as machine translation and cross-language information retrieval.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.
Building a Chinese-English Mapping Between Verb Concepts for Multilingual Applications

Language and Media Processing Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park, MD, 20742-3275

Approved for public release; distribution unlimited

<table>
<thead>
<tr>
<th>1. REPORT DATE</th>
<th>2. REPORT TYPE</th>
<th>3. DATES COVERED</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEC 2001</td>
<td></td>
<td>00-12-2001 to 00-12-2001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. TITLE AND SUBTITLE</th>
<th>5a. CONTRACT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building a Chinese-English Mapping Between Verb Concepts for Multilingual Applications</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. AUTHOR(S)</th>
<th>5b. GRANT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</th>
<th>5c. PROGRAM ELEMENT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language and Media Processing Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park, MD, 20742-3275</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8. PERFORMING ORGANIZATION REPORT NUMBER</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</th>
<th>10. SPONSOR/MONITOR’S ACRONYM(S)</th>
<th>11. SPONSOR/MONITOR’S REPORT NUMBER(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12. DISTRIBUTION/AVAILABILITY STATEMENT</th>
<th>13. SUPPLEMENTARY NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approved for public release; distribution unlimited</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>a. REPORT unclassified</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. ABSTRACT unclassified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. THIS PAGE unclassified</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18
Building a Chinese-English Mapping Between Verb Concepts for Multilingual Applications

Bonnie J. Dorr, Gina-Anne Levow, and Dekang Lin

LAMP TR-060, UMIACS TR-2001-01, CS TR-4205

Abstract:
This paper addresses the problem of building conceptual resources for multilingual applications. We describe new techniques for large-scale construction of a Chinese-English lexicon for verbs, using thematic-role information to create links between Chinese and English conceptual information. We then present an approach to compensating for gaps in the existing resources. The resulting lexicon is used for multilingual applications such as machine translation and cross-language information retrieval.

Acknowledgements:
The University of Maryland authors are supported, in part, by PFF/PECASE Award IRI-9629108, DOD Contract MDA904-96-C-1250, and DARPA/TTO Contract N66001-97-C-8540. Dekang Lin is supported by Natural Sciences and Engineering Research Council of Canada grant OGP121338. We are indebted to an astute anonymous reviewer for catching several errors and inconsistencies in the original draft. We also thank Nizar Habash, Maria Katsova, and Scott Thomas for their assistance with experimental runs on the data and their useful commentary and aid in the preparation of this document.

(In Proceedings of Fourth Conference of the Association for Machine Translation in the Americas (AMTA), Cuernavaca, Mexico, 2000, pp. 1-12.)
1 Introduction

With the advent of the web and increasingly global interconnectivity, the need for online multilingual information has grown significantly in the 5-10 years. This is accompanied by a growing for rapid construction of lexical resources. Cross resources by human labor alone has become inflexible, thus motivating the development of autonomous and semi-automatic approaches to resource acquisition. This paper addresses large-scale construction of a Chinese-English lexicon for verbs, including an approach to compensating for gaps in the existing resources.

The lexicons resulting from our acquisition approach are used for semantic analysis in applications such as machine translation and cross-language information retrieval. The importance of semantic analysis in either of these two applications is clear when one considers the degree of inaccuracy that might result from using a weak alternative, such as access to a bilingual word list.

Our starting point is an existing classification of English verbs called EVCA (English Verbs Classes and Alternations) (Levin, 1993). We couple this with a Chinese conceptual database called HowNet (Zhendong, 1988a; Zhendong, 1988b; Zhendong, 1988c) (http://www.hownet.com), from which we extract thematic-role information (e.g., a mapping between the HowNet “Patient” and the EVCA-based “Th(eme)”) to create links between Chinese and English conceptual information. HowNet currently contains no English translations; thus, we also use a large machine-readable Chinese-English dictionary called Optilex to produce candidate English translations. Although later versions of HowNet are expected to include the English translations, these are not openly available—only the binary versions have been promised and these will be accessible solely through the use of (purchasable) HowNet software. Moreover, we expect our techniques to be generally applicable to other foreign language semantic hierarchies where English translations are not available. We predict this will occur more and more frequently, as online (non-bilingual) linguistic resources continue to be made available in multiple languages.

Several researchers have investigated the problem of assigning class-based senses to verbs (Dorr, 1997; Palmer and Rovenstine, 1996; Palmer and Wu, 1995) using a variety of online resources including Longman’s Dictionary of Contemporary English (LDOCE) (Procter, 1978), EVCA (Levin, 1993), and WordNet (Miller and Fellbaum, 1991). Translation of English classes into other languages has proven difficult (Jones et al., 1994; Nomura et al., 1994; Saint-Dizier, 1996), but regularities between different language classifications can be found in some online resources (Dang et al., 1998; Dorr and Jones, 1999; Olsen et al., 1998).

This work extends previous work which used a concept space to produce a hierarchical organization of Chinese verbs (Palmer and Wu, 1995). We adopt a technique that is similar in flavor to that of (Dang et al., 1998) for partitioning English verbs into refined classes using WordNet, with the following extensions: (1) The use of the entire EVCA database rather than a small set of verbs (the break class); (2) The provision of a thematic-role based filter for a more refined version of verb-class assignments; (3) Concept alignment across two different language hierarchies (Chinese and English); and (4) Mappings between Chinese and English thematic roles.

This work relies on an augmented set of EVCA classes which include 26 new classes (Dorr, 1997). There are 500 total classes in the extended set, each hand-tagged with semantic representations, thematic-role information, and WordNet synset numbers. We will demonstrate that it is possible to produce a lexicon by associating 709 Chinese HowNet concepts with 500 EVCA classes, with a clear concept-to-class correspondence in a large majority of the cases.²

Figure 1 illustrates the relation between existing resources and the mappings we produced. Solid lines represent pre-existing mappings; dotted lines are ones resulting from the application of our tech-

---

²HowNet contains 815 verb HowNet concepts altogether. However, we are not including the 106 HowNet concepts that are not associated with any Chinese words; these are “higher level” conceptual nodes with no Chinese realization (e.g., V.1 [static]).
niques. The most critical of these is the one labeled β-roles (shorthand for “thematic roles”), which associates EVCA classes with HowNet Concepts. The remaining two dotted-line mappings are “transitive closure” biproducts of the other mappings: Once the thematic-role mapping associates EVCA verbs with HowNet verbs, each HowNet verb is associated with Optilex-based English glosses (transla-

tions) and WordNet 1.6 Senses.

We will describe how these correspondences are derived and we will show how this process has provided a framework for compensating for gaps in our online resources.

2 Multilingual Applications

The semantic representations produced semi-
automatically for our multilingual resources are used in machine translation (MT) and cross-language information retrieval (CLIR) applications. Both applications rely on the use of a parser for mapping the input sentence into a syntactic tree. The parser output is semantically analyzed, producing an encoding of semantic and argument-structure information.

The MT approach is interlingual, where the target-language lexicon is searched for appropriate lexical items matching argument-structure information (Dorr et al., 1998). A screen snapshot of a MT example is shown in Figure 2. The CLIR approach relies on the same interlingual representation to transform a user’s query into the document language for information retrieval (Dorr and Katsova, 1998; Levow et al., 2000).

In both of these applications, thematic roles facilitate the selection of appropriate target-language words. For example, the Chinese verb 拉 (la) corresponds to a wide range of English translations—even if we examine only the verb translations: slash, cut, chat, pull, drag, transport, move, raise, help, implicate, involve, defecate, press-gang.4 Our approach provides a framework for disambiguation of such cases. Certain of these possibilities—transport and move—are analyzed as

---

3 The Chinese verbs are additionally associated (for free) with WordNet senses from our previously tagged EVCA verbs. More details are given in (Dorr et al., 2000).

4 The ambiguity in the word 拉 (la) can often be resolved if it is combined with other characters. For example, 拉车 (la che) unambiguously means pull a cart. However, since object dropping is a frequent phenomenon in Chinese, it is not uncommon for verbs like ‘la’ to appear without an argument that easily disambiguates the word. Thus, our approach must allow for multiple possibilities in the lexicon.

---

Figure 2: Translation of a Chinese sentence into English

1. Associate English Optilex glosses with all 12342 Chinese verbs in HowNet, producing 41,324 Chinese-English pairs.

2. Associate each verb-to-concept candidate with at least one of the 500 EVCA classes.3

3. For each HowNet concept, partition the associated Chinese-English pairs into groups whose English glosses correspond to EVCA classes.

Figure 3: Mapping Chinese HowNet Concepts to English EVCA Classes

one semantic representation corresponding to thematic roles (agent, theme, goal, source). Other possibilities—help—are analyzed as a different semantic representation corresponding to thematic roles (agent, theme, mod-poss).

3 Mapping Between Chinese HowNet and English EVCA

Our technique for mapping between Chinese HowNet concepts and English EVCA classes involves associating HowNet thematic roles with those in EVCA. Each HowNet concept (and each EVCA class) is paired with a list of thematic roles, which we call a thematic grid. For example, the HowNet concept [Cure] is paired with the grid (agent, patient, content, tool), as in The doctor (agent) cured the man (patient) of pneumonia (content) using antibiotics (tool). The
corresponding grid in our EVCA database is (ag,th,mod-poss(of)). Although the HowNet and EVCA roles are not in a one-to-one correspondence, they can still be used for a "closest match" prioritization of candidate HowNet-EVCA associations, as we will see shortly.

The three top-level tasks involved in mapping Chinese HowNet concepts to and English EVCA classes are given in Figure 3. (See (Dorr et al., 2000) for more details.) For the purposes of this discussion, we focus on the last of these three tasks, which involves a massive filtering of spurious class assignments. This task involves three steps:

- Order the candidate EVCA classes so that the highest-ranking classes are those that contain the highest number of English verbs matching the Optilex glosses.

- In cases where a tie-breaker is needed, reorder the candidate EVCA classes according to the degree to which the thematic grid in HowNet concept matches that of the relevant EVCA class. The matching procedure relies on correlations derived from approximately 200 seed mappings.\(^5\) Figure 4 shows a small subset of these mappings.

- For each Chinese-English entry associated with the HowNet concept, assign the highest ranking candidate EVCA class.

Consider the case of the multiply ambiguous Chinese verb 拉 (la). Two of the HowNet concepts associated with this verb are [Help] and [Transport]. The thematic grid associated with [Help] is (agent, patient, scope) (as in John helped him with his work). This grid most closely matches that of the Equip EVCA Class (where 拉 (la) is translated as help) which has the grid _ag_th,mod-poss(with); thus, the [Help] HowNet concept is associated with the Equip EVCA Class, and the mapping between the two is (agent->ag), (patient->th), (scope->mod-poss).

On the other hand, the [Transport] HowNet concept is associated with the thematic grid (agent, patient, LocationIni, LocationFin, direction) (as in John transported the goods from Boston to New York (westward)). This grid most closely matches that of the Send EVCA Class (where 拉 (la) is translated as transport);

thus, the [Transport] HowNet concept is associated with the Send EVCA class, and the mapping between the two is (agent->ag), (patient->th), (LocationIni->src), (LocationFin->goal). The end result is that the English glosses associated with 拉 (la) are filtered down to help in the EVCA's Equip class and transport in EVCA's Send class; the corresponding semantic representations are assigned from the EVCA database.

The massive filtering of spurious assignments is evident when we examine each individual HowNet concept. Consider the [Establish] HowNet concept. This concept is ultimately associated with only two EVCA classes, 29.2.c and 26.4.a (Characterize and Create), but it initially had 29 potential EVCA class assignments. One EVCA class that was ruled out is the Change of State class. 45.4.a, associated with the Optilex translation colonize for the Chinese verb 殖民 (zhimin). Although this is a perfectly valid EVCA class assignment for the HowNet concept [Colonize], it is not appropriate for the [Establish] HowNet concept. Because this class is ranked 8th for [Establish]—as opposed to 1st and 2nd place ranking for 29.2.c and 26.4.a, respectively—this assignment is ruled out by our algorithm.

4 Compensating for Resource Deficiencies

As part of our effort to produce a complete alignment between HowNet and EVCA, we built an EVCA-based canonical specification for each of the 709 HowNet concepts so that we could compensate for certain types of resource deficiencies. The canonical specification consists of an EVCA class coupled with its associated prototype verb. These canonical specifications provide a mapping between a HowNet concept and an EVCA class/prototype-verb pair.

Each canonical specification was automatically generated according to the highest ranking EVCA class using steps 3.a and 3.b in Section 2. All such specifications were hand-verified at a rate of 80 per hour for 709 classes). In most cases, the prototype verb names the HowNet concept, e.g., transport for the [Transport] HowNet concept. In other cases—where the HowNet concept is not an English word—the prototype word is a realization of that concept, e.g., belittle for the [PlayDown] HowNet concept. A sample of the canonical specifications is given in Figure 5.

We use these canonical specifications to compensate for gaps that arise in our three online resources: (1) EVCA, (2) Optilex, and (3) HowNet.

\(^5\) The seed mappings were done by hand at a rate of approximately 50 mappings per hour; these were verified by a native Chinese speaker in a half day.
4.1 EVCA Gaps

An EVCA gap is detected when an Optilex verb gloss for a Chinese verb does not occur in EVCA. When this occurs, the canonical specification for the Chinese verb is automatically used to assign the verb an appropriate EVCA class. For example, one Optilex gloss associated with the HowNet concept [Establish] (for the verb 重建 (chongjian)) is reconstruct, which does not occur in EVCA. Our technique associates this Chinese verb with the canonical specification “29.2.c Characterize, establish,” and the Chinese verb is then linked with the word sense associated with establish.

An interesting byproduct of the handling of EVCA gaps is that it allows us to enhance our EVCA resource. For example the verb reconstruct can now be added to EVCA Class 29.2.c, on a par with the previously classified EVCA verb establish.

4.2 Optilex Gaps

An Optilex gap occurs when a particular translation for a Chinese verb is missing. For example, the verb 搞 (baibu) has only one Optilex gloss: manipulate. However, the word 搞 is associated with two HowNet concepts, [Decorate] and [Control]. This gloss is only appropriate for the [Control] concept. The decorate meaning of 搞 (baibu) is omitted in Optilex.

Such gaps are detected by means of two types of information: (1) HowNet and EVCA thematic grid; and (2) correlations between the gloss under question and other HowNet concepts. In this particular example, the thematic grid for manipulate in EVCA (ag, exp, instr), which is ranked low (11th out of 28) with respect to the roles (agent, patient) associated with the HowNet [Decorate] concept. By contrast, this same EVCA class has a high ranking (2nd out of 22) with respect to the HowNet [Control] concept due to a close match between (ag, exp, instr) and the HowNet thematic roles (agent, patient, ResultEvent). In addition, the correlation of the gloss manipulate is much higher for HowNet’s [Control] concept than it is for HowNet’s [Decorate] concept (4 occurrences compared to 0). From these two types of information, we can conclude that the decorate sense of 搞 (baibu) is missing from Optilex. As in the case with EVCA gaps, our technique associates the Chinese verb with the canonical specification “9.8.b Fill, decorate” to compensate for this Optilex gap.

In addition to their usefulness in handling of gaps in our lexical resources, the canonical specifications proved useful for assigning EVCA classes to Chinese verbs whose Optilex gloss was not “parsable” by our gloss extraction procedure. For example, the Chinese verb 挨打 (aida) has only a single Optilex translation: take a beating. This verb is associated with the HowNet concept [Suffer], which has as its canonical specification “31.3.d Marvel, suffer.” Thus, our technique associates挨打 verb with this canonical specification.

A similar approach is used for unknown or misspelled words. For example, the translation of输送 (shushong) as in Optilex is misspelled as transport. Because this verb is associated with HowNet’s [Transport] concept, we associated this verb with the canonical specification “11.1 Send, transport.”
4.3 HowNet Gaps

In some cases, the HowNet hierarchy incorrectly associates a Chinese word with a particular concept. For example, HowNet incorrectly associates the two Chinese verbs 扎花 (zhahua) and 绣花 (xiuhua) with the [Decorate] concept. These two verbs are translated as *embroider* in EVCA class 26.1.b (Build), but their meaning is closer to *sew flowers*. That is, the patient is incorporated into the verb, which means the thematic grid _ag_th_goal(into),ben(for) does not match that of the HowNet concept (agent,possession,source).

Discrepancies in HowNet are detected by means of EVCA-class frequency for a particular HowNet concept. Out of the 17 verbs associated with HowNet’s [Decorate] concept, only two of them (the two misclassified Chinese verbs) are associated with an EVCA class that is not 9.9 or 9.8. As in the gap-recovery described approaches above, our technique associates the misclassified verbs with the canonical specification “9.8.b Fill, decorate.”

5 Results

Preliminary results of our classification scheme were reported in (Dorr et al., 2000). This earlier work resulted in 8089 EVCA-classified Chinese entries—about 43% of the number of potential entries. The remaining 10441 entries were accounted for through the compensation techniques described above. Using the canonical specifications, we have achieved a more refined EVCA-to-HowNet mapping, providing an increase in EVCA-classified Chinese words from the previous 8089 entries to the current expanded set of 17284 EVCA-classified Chinese words. The histogram in Figure 6 characterizes the number of EVCA classes required for coverage of 709 HowNet concepts.

Examples of the HowNet partitionings into EVCA classes are given in Figure 7, with a focus on the cases where 1 partition was found. Percentages are given with respect to the number of Chinese verbs associated with each EVCA class.

We consider the approach to be a success for several reasons: (1) In 359 cases (50% of the HowNet concepts), the partitioning corresponded to 3 or fewer EVCA classes; (2) Most HowNet concepts with 2 or more partitions had a very heavy association with a single EVCA class (60% or higher), with most other partitions falling around 20% or lower; (3) Only 2 cases did not correspond to any EVCA class (i.e., degenerate HowNet concepts for which no correlations with EVCA could be found); (4) There were virtually no partitionings (a handful of single HowNet concepts) exceeding 13 EVCA classes.

6 Summary and Future Work

We have presented an approach to aligning two large-scale online resources, HowNet and EVCA. The lexicon resulting from this approach is large-scale, containing 18530 Chinese entries. The technique for producing these links involves matching thematic grids in HowNet with those in EVCA. Our results indicate that the correspondence is very high between the 709 Chinese HowNet concepts and the 500 EVCA classes. We see our techniques as the first step toward a general approach to building repositories for interlingual-based NLP applications.

We are currently investigating the use of the lexicon for word-sense disambiguation in machine-translation and cross-language information retrieval. As we saw above the Chinese verb 拉 (la) has several possible translations, but not all of these will be appropriate in every context. If we can determine which HowNet concept corresponds to 拉 (la), then we will translate it appropriately. For example, if the HowNet concept is [Transport], the translation would be *ship or transport*, but not *slash, chat, implicate*, etc. We can detect which HowNet concept is appropriate by examining the other words in the sentence. If those words co-occur with other Chinese verbs associated with a particular HowNet concept (as determined through a corpus analysis), then it is likely that that HowNet concept is the appropriate one for the Chinese verb. That is, if we find other verbs from a given HowNet concept occurring in the same context, then we can hypothesize that this particular verb has the meaning of this HowNet concept.

The algorithm for mapping between HowNet concepts and EVCA classes requires a “training” step—i.e., the seed mappings given earlier. However, it is possible to produce a ranked mapping between thematic grids by counting correspondences between EVCA-based roles and the HowNet-based roles across the entire concept space. This approach is also currently under investigation.

Another area of investigation is the use of a WordNet-based distance metric (e.g., the information-content approach of (Resnik, 1995)) for additional pruning power in the HowNet-to-EVCA alignment. Because each of the entries in the EVCA
Figure 6: Distribution of HowNet Concepts by Number of Intersecting EVCA Classes using Canonical Specifications

<table>
<thead>
<tr>
<th>HowNet Concept</th>
<th>EVCA Class(es)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport</td>
<td>11.1 Send</td>
</tr>
<tr>
<td>Help</td>
<td>13.4.2 Equip</td>
</tr>
<tr>
<td>Apologize</td>
<td>32.2.a Long</td>
</tr>
<tr>
<td>Naming</td>
<td>29.3 Dub</td>
</tr>
<tr>
<td>Judge</td>
<td>29.4 Declare</td>
</tr>
<tr>
<td>Moisten</td>
<td>45.4.a Change of State</td>
</tr>
<tr>
<td>Excrete</td>
<td>40.1.2 Breathe</td>
</tr>
<tr>
<td>TakeVehicle</td>
<td>51.4.2.a.i Motion by Vehicle</td>
</tr>
<tr>
<td>PlayDown</td>
<td>33.5.b Judgment (75%), 31.2.a Admire (25%)</td>
</tr>
<tr>
<td>Establish</td>
<td>29.2.c Characterize (90%), 28.4.a Create (19%)</td>
</tr>
<tr>
<td>Decorate</td>
<td>9.8.b Fill (50%), 26.1.b Build (43%), 9.9.ni Butter (25%)</td>
</tr>
<tr>
<td>Buy</td>
<td>10.5 steal (98%), 13.5.1.a Get (30%), 13.5.1.b.i Get (54%), 13.5.2.d Get (46%)</td>
</tr>
<tr>
<td>Teach</td>
<td>29.2.c Characterize (24%), 33.5.b Judgment (71%), 37.9.a Advise (29%), 37.1.a</td>
</tr>
<tr>
<td></td>
<td>Transfer Message (45%), 31.1.a Amuse (19%)</td>
</tr>
</tbody>
</table>

Figure 7: Examples of HowNet Partitionings with Respect to EVCA
classification is associated with a WordNet sense (Miller and Fellbaum, 1991), it is possible to rule out certain class assignments for a given HowNet concept by examining semantic distance between the Optilex glosses for a particular Chinese word and the glosses for other words associated with that concept.

References


