Chinese-English Semantic Resource Construction
Bonnie Dorr, Gina-Anne Levow, Dekang Lin, Scott Thomas

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

Abstract
We describe an approach to large-scale construction of a semantic lexicon for Chinese verbs. We leverage off of three existing resources— a classification of English verbs called EVCA (English Verbs Classes and Alternations) [Levine, 1993], a Chinese conceptual database called HowNet (Zhendong, 1988c, Zhengdong, 1988b, Zhendong

***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.
Chinese-English Semantic Resource Construction

Bonnie J. Dorr†, Gina-Anne Levow†, Dekang Lin‡, Scott Thomas†

†Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
{bonnie.gina.katsova}@umiacs.umd.edu
Phone: (301)-405-6768
Fax: (301)-314-9658

‡Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada, T6G 2H1
lindek@cs.ualberta.ca
Phone: (780)-492-5198
Fax: (780)-492-1071

Abstract: We describe an approach to large-scale construction of a semantic lexicon for Chinese verbs. We leverage off of three existing resources—a classification of English verbs called EVCA (English Verbs Classes and Alternations) [Levin, 1993], a Chinese conceptual database called HowNet [Zhendong, 1988c, Zhendong, 1988b, Zhendong, 1988a] (http://www.how-net.com), and a large machine-readable dictionary called Optilex. The resulting lexicon is used for determining appropriate word senses in 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/ITO Contract N66001-97-C-8540. Dekang Lin is supported by Natural Sciences and Engineering Research Council of Canada grant OGP121338.
1 Introduction

With the growing quantity of online multilingual information, automatic and semi-automatic techniques for lexical acquisition are more critical now than ever before. We describe an approach to large-scale construction of a semantic lexicon for Chinese verbs. We leverage off of three existing resources—a classification of English verbs called EVCA (English Verbs Classes and Alternations) [Levin, 1993], a Chinese conceptual database called HowNet [Zhendong, 1988c, Zhendong, 1988b, Zhendong, 1988a] (http://www.how-net.com), and a large machine readable Chinese-English dictionary called Optilex.¹

Our approach involves extraction of candidate translations from Optilex for each of the Chinese verbs occurring in HowNet. We then create links between Chinese concepts and English classes using thematic-role mappings between HowNet entries and EVCA-based entries. Each Chinese-English link is subsequently associated with a sense from WordNet [Miller and Fellbaum, 1991], thus producing a new Asian companion to the current (Euro)WordNet initiative. The resulting lexicons are used for determining appropriate word senses in applications such as machine translation and cross-language information retrieval.

Several researchers have investigated the problem of assigning class-based senses to verbs [Dang et al., 1998], [Dorr and Jones, 1999], [Dorr and Jones, 1996] [Dorr, 1997], [Jones et al., 1994], [Nomura et al., 1994] [Olsen et al., 1998], [Palmer and Wu, 1995], [Palmer and Rosenzweig, 1996], and [Saint-Dizier, 1996]. This work extends the techniques described by [Palmer and Wu, 1995], which used a concept space to produce a hierarchical organization of Chinese verbs. The extensions include the use of the entire EVCA database rather than a small set of verbs (the break class) and the provision of a thematic-role based filter. We adopt a technique that is similar in flavor to the intersective-class approach of [Dang et al., 1998], with the following extensions: (1) Concept alignment across two different language hierarchies (Chinese and English) rather than one; (2) Mappings between Chinese and English thematic roles; and (3) Hooks into WordNet senses for both languages. The next section describes the HowNet conceptual database. Following these, we will describe the approach we used to produce the concept-to-class correspondence. Section 4 presents the result of our automatic acquisition experiment.

2 HowNet Conceptual Database

HowNet is an on-line conceptual common-sense knowledge base that contains hierarchical information relating concepts to the associated Chinese word. Our focus is on the verb hierarchy, which has the structure shown in Table 1.

The number labels given here are our own; we use these for indicating the level of each concept in the HowNet database. Note that the highest two concepts in the verb hierarchy are “static” (V.1) and “act” (V.2). These correspond, respectively, to verbs such as 成為 (become under the “static” node V.1.1.1) and 開始 (start under the “act” node V.2.1.1). The levels go much deeper than these, with the lowest ones at 8 levels deep, e.g., V.1.2.1.6.3.3.1.15 itch.

Within each of the HowNet classes is a thematic-role specification. For example, the verb “cure” has the thematic-role specification (agent, patient, content, tool). Consider the sentence The doctor cured the man of pneumonia using antibiotics. The roles in the specification have the following binding, respectively, for this sentence: doctor, man, pneumonia, antibiotics.² The

¹Optilex is the machine-readable version of the CETA dictionary, licensed from the MRM corporation, Kensington, MD.
²Thematic-role specifications and their use in generation of natural-language translations are described further in
V.1 [static]
V.1.1 [relation]
V.1.1.1 [isa]
V.1.1.2 [possession]
V.1.1.3 [comparison]
V.1.1.4 [suit]
V.1.1.5 [inclusive]
V.1.1.6 [connective]
V.1.1.7 [CauseResult]
V.1.1.8 [TimeOrSpace]
V.1.1.9 [arithmetic]
V.1.2 [state]
V.1.2.1 [StatePhysical]
V.1.2.2 [StateMental]

Table 1: HowNet Verb Hierarchy

<table>
<thead>
<tr>
<th>Number of EVCA Classes per Concept:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of HowNet Concepts:</td>
<td>2</td>
<td>371</td>
<td>71</td>
<td>20</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2: Partitioning of HowNet Concepts into EVCA Classes

thematic-role specifications are used for prioritizing candidate HowNet-EVCA associations, as will be described below.

3 Approach

We have associated 478 Chinese HowNet concepts with 485 EVCA classes, demonstrating a clear concept-to-class correspondence in a large majority of the cases. The mapping between Chinese HowNet and English EVCA (hence WordNet) involves three steps.

The first step is to produce all possible English Optilex glosses (translations) for all 12342 Chinese verbs in HowNet and associate each Chinese verb with one or more of the 478 HowNet concepts, forming 48,884 verb-to-concept candidates. For example, there is a common Chinese verb 拉 (la) that is multiply ambiguous, corresponding to 13 Optilex-based English glosses: *slash, cut, chat, pull, drag, transport, move, raise, help, implicate, involve, defecate*, and *pressgang*. This verb is associated with 9 HowNet concepts: [Transport], [Attract], [Excrete], [Force], [Help], [Include], [Pull], [Recreation], and [Talk].

The second step involves associating each verb-to-concept candidate with one or more of the 485 EVCA classes, forming an average of 2 thousand verb-to-class entries per HowNet concept (on the order of 1 million verb-to-class candidates, total). For example, the Chinese verb 拉 (la) is

[Dorr et al., 1998].

3There are actually more than 800 concepts in HowNet that define events. The number was reduced to 478 for the purpose of this preliminary experiment; a more in-depth acquisition process is currently underway to fill out the final 300+ concepts. See [Dorr et al., 2000].
<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.ii Motion by Vehicle</td>
</tr>
<tr>
<td>PlayDown</td>
<td>33.b Judgment (75%), 31.2.a Admire (25%)</td>
</tr>
<tr>
<td>Establish</td>
<td>29.2.c Characterize (90%), 26.4.a Create (19%)</td>
</tr>
<tr>
<td>Decorate</td>
<td>9.8.b Fill (50%), 26.1.b Build (43%), 9.9.ii Butter (25%)</td>
</tr>
<tr>
<td>Buy</td>
<td>10.5 Steal (08%), 13.5.1.a Get (30%), 13.5.1.b.ii Get (54%), 13.5.2.d Get (46%)</td>
</tr>
<tr>
<td>Teach</td>
<td>29.2.c Characterize (24%), 33.b Judgment (71%), 37.9.a Advise (29%), 37.1.a Transfer Message (45%), 31.1.a Amuse (19%)</td>
</tr>
</tbody>
</table>

Table 3: Examples of HowNet Partitionings with Respect to EVCA

associated with 22 EVCA classes: Admire (31.2.b, implicate, invoke); Amuse (31.1.b, transport, move, cut; Braid (41.2.2, cut); Breathe (40.1.2, defecate); Build (26.1.a, cut); Carry (11.4.i, carry, pull, drag); Chitchat (37.6.a, chat); Crane (40.3.2, raise); Cut (21.1.a, slash, cut); Cut (21.1.d, cut); Equip (13.4.2, help); Force (12.a.ii, pull); Get (13.5.1.a, pull); Grow (26.2.a.ii, raise); Hurt (40.8.3, pull, cut); Meander (47.7.a, cut); Play (009, pawn); Put (9.4.a, raise); Search (35.2.a, drag); Send (11.1, smuggle, transport, ship, convey); Send Slide (11.2.b, move); Split (23.2.b, cut, pull).

The final step is to partition each HowNet concept into groups of Chinese-English pairs whose English glosses correspond to EVCA classes. This involves three subtasks:

- 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-role specification in HowNet concept matches that of EVCA class.

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

Consider two HowNet concepts associated with the the Chinese verb 拉 (la): [Help] and [Transport]. The thematic-role specification associated with [Help] is (agent, patient, scope) (as in John helped him with his work). This specification most closely matches that of Equip EVCA Class (where 拉 (la) is translated as help) which has the specification 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-role specification (agent, patient, LocationIni, LocationFin, direction) (as in John transported the goods from Boston to New York (westward)). This specification most closely matches that of the
Send EVCA Class (where 运输 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 运输 are filtered down to help in the
Equip semantic class and transport in the Send semantic class; the corresponding WordNet senses
are assigned (for free) from the hand-tagged EVCA database. These are Senses 1–3 in the case of
transport (i.e., move/carry/displace) and Sense 1 in the case of help (i.e., aid/assist):

- **transport:**
  - Sense 1: transport
  - Sense 2: transport, carry
  - Sense 3: transport, send, ship

- **help:**
  - Sense 1: help, assist, aid

4 Results

Table 2 characterizes the number of EVCA classes required for coverage of 478 HowNet concepts.
We consider the approach to be a success for several reasons: (1) Association of a unique EVCA
class to a HowNet concept was achieved in 371 cases—77% of the HowNet classes; (2) Most of
the other cases partitioned the HowNet entries into 2 EVCA classes; (3) Only 2 cases did not
correspond to any EVCA class (i.e., every word associated with the concept belonged to a different
EVCA class); (4) There were no partitionings exceeding 5 EVCA classes.

Examples of the HowNet partitionings into EVCA classes are given in Table 3, with a focus on
the cases where 1 partition was found. In cases where there is more than 1 partition, percentages
are given with respect to the number of Chinese verbs in each HowNet class.4

5 Summary

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 17284 Chinese-English conceptual
links. The technique for producing these links involves matching semantic-role specifications in
HowNet with those in EVCA. Our results indicate that the correspondence is very high between
the 478 Chinese HowNet concepts and the 485 EVCA classes. Because each Chinese-English link
is associated with a WordNet sense, we see this resource as the first step toward producing

References

Investigating Regular Sense Extensions Based on Intersective Levin. In *ACL/COLING 98,
Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics* (joint
with the 17th International Conference on Computational Linguistics). Montreal, Canada.

---

4The astute reader will notice the percentages don't always total 100%. This is because certain of the Chinese verbs
are assigned to two different “partitionings.” The resulting groups are, thus, not true partitions in the mathematical
sense since they are not necessarily mutually exclusive. In the cases where the percentages total 100%, the resulting
groups are mutually exclusive.


