Large-Scale Construction of a Chinese-English Semantic Hierarchy

Bonnie J. Dorr[†], Gina-Anne Levow[†], Dekang Lin[‡]

[†]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

Key Words: Semantic Resource Construction, Chinese-English lexicons, thematic roles, machine translation, Information retrieval, Lexical acquisition.

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

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. We are indebted to 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.

1 Introduction

With the advent of the web and increasingly more global interconnectivity, the necessity for online multilingual information has increased significantly in the last 5–10 years. This is accompanied by a growing need for rapid construction of lexical resources. Creating resources by human labor alone has become infeasible, thus motivating the development of automatic and semi-automatic approaches to resource acquisition. This paper addresses large-scale construction of a semantic hierarchy for Chinese 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, 1988c, Zhendong, 1988b, Zhendong, 1988a] (http://www.how-net.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 concepts and English classes. HowNet currently contains no English translations; thus, we also use a large machine-readable Chinese-English dictionary called Optilex to produce candidate English translations.[1] 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. Thus, we see our approach to selecting candidate translations as an important one for addressing the more general problem of constructing large-scale multilingual resources.

Several researchers have investigated the problem of assigning class-based senses to verbs [Dorr, 1997], [Palmer and Rosenzweig, 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]. The work of [Nomura et al., 1994], [Saint-Dizier, 1996], [Jones et al., 1994] indicates that the translation of English classes into other languages is not straightforward, but later work has shown that regularities between different language classifications is evident in online resources [Dang et al., 1998], [Dorr and Jones, 1999], [Olsen et al., 1998].

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: (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. Later work by [Dang et al., 1998] uses an intersective-class technique that partitions English verbs into refined classes using WordNet as a conceptual basis. We adopt a technique that is similar in flavor to this approach, 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-role specifications.

The EVCA classes used in this work relies on extensions by [Dorr, 1997] and [Dorr and Jones, 1999] to a finer-grained set of semantic classes, including 26 new classes. There are 500 total classes in the extended set, each hand-tagged semantic representations and thematic-role specifications. Mapping

[1] Optilex is a large (600k entries) machine-readable version of the CETA Chinese-English dictionary, licensed from the MRM corporation, Kensington, MD.

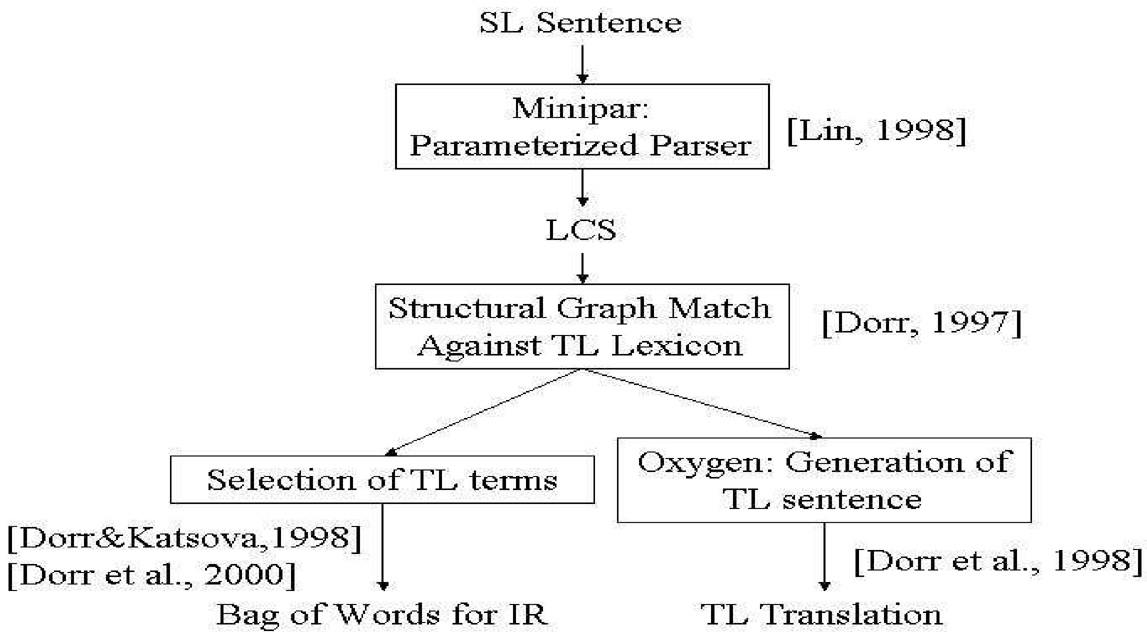


Figure 1: MT and CLIR Applications that use Multilingual Resources

English thematic roles to their Chinese counterparts is the primary aid in selecting the appropriate entry in EVCA. 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] We will describe how this correspondence is extracted 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; see Figure 1. Both applications rely on the use of a parser for mapping the input sentence into a syntactic tree. We currently use an in-house parser called REAP [Weinberg et al., 1995], which will be replaced this year by a parameterized parser called Minipar [Lin, 1998]. Minipar produces English parse trees on a large scale, and Chinese on smaller scale; one of the benefits of this parser is its ease of portability to new languages.

Both applications operate in the Chinese-English direction. The parser output is semantically analyzed, producing an encoding of semantic and argument-structure information called *lexical conceptual structure* (LCS) [Dorr, 1997].

The MT approach is interlingual, adhering to a graph-matching scheme, where the target-language lexicon is searched for appropriate lexical items matching the LCS representation. The most recent version of the generator is called Oxygen, a variant of Nitrogen [Langkilde and Knight, 1998a, Langkilde and Knight, 1998b, Langkilde and Knight, 1998c] which uses our own linearizer imple-

[2] HowNet contains 815 verb concepts altogether. However, we are not including the 106 concepts that are not associated with any Chinese words; these are "higher level" conceptual nodes with no Chinese realization (e.g., V.1 |static|).

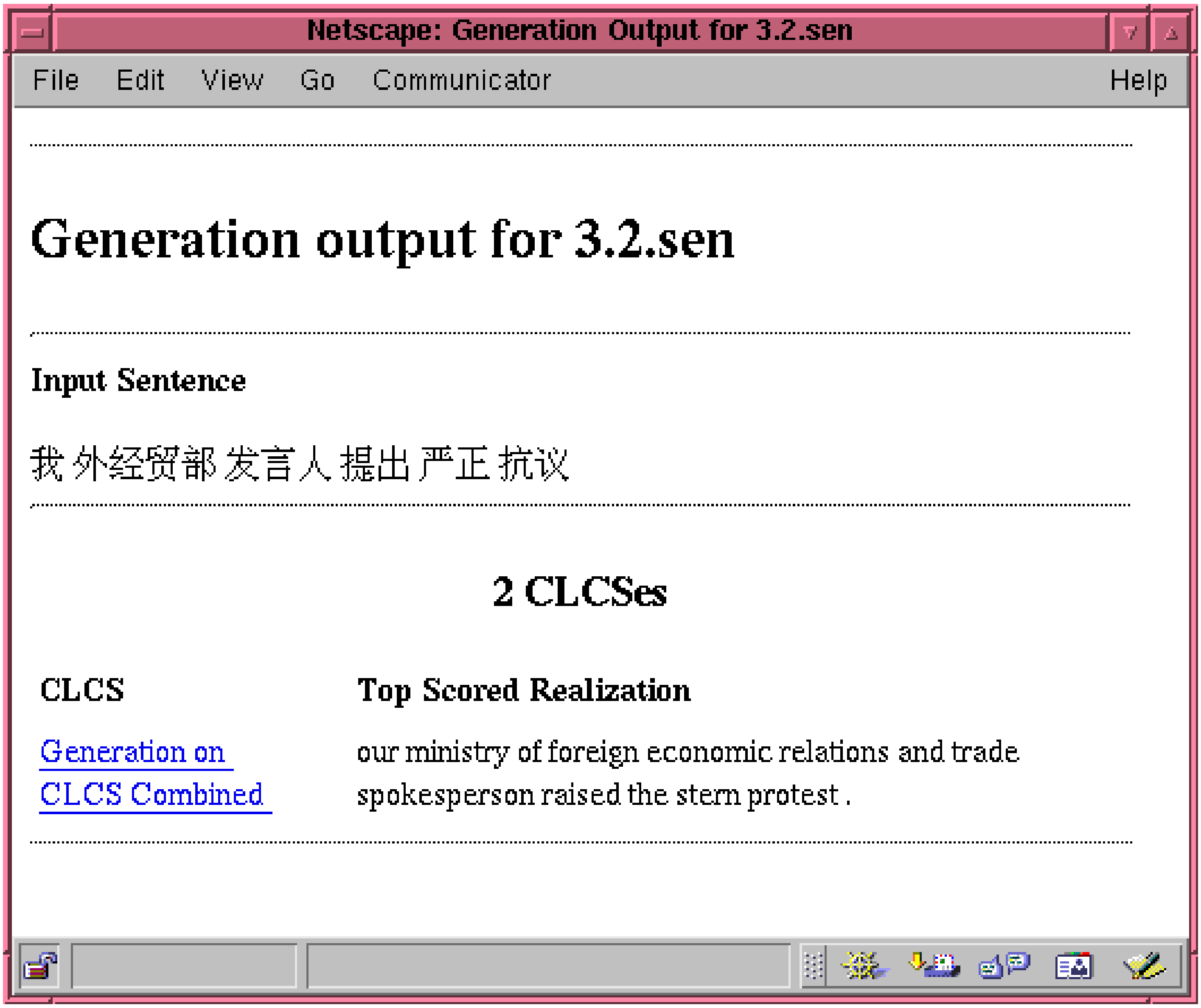


Figure 2: Translation of a Chinese sentence into English

mented in lisp with Nitrogen's statistical extraction module and Nitrogen's morphological generation engine. Thematic-role specifications and their use in generation of natural-language translations are described in [Dorr et al., 1998] as a component of a precursor to the Oxygen system. The English output is produced by means of two steps: lexical selection and syntactic realization. Lexical selection involves a comparison between LCS components and abstract LCS frames associated with words in an English lexicon. Syntactic realization re-casts LCS-based thematic roles as relations in an an unordered tree where the root is a concept and each child is linked by a relation.

A screen snapshot of an MT example is shown in Figure 2. This translation is considerably better than its literal equivalent: *Our Foreign_Economic_Trade_Ministry spokesperson lodge stern protest*.

The CLIR approach relies on the same interlingual representation used for machine translation to transform a user's query into the document language for information retrieval. Query translation involves a structural match of the query against a database of LCS structures, built from a thematic hierarchy. A "bag of words" is then produced as input to a conventional information retrieval engine. For additional details about the use of semantic representations in the CLIR system, see [Dorr and Katsova, 1998], [Dorr et al., 2000].

During the process of mapping the LCS representation into the target language, 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*, *pressgang*.[3] Our approach provides a framework for disambiguation of such cases. Certain of these possibilities—*transport* and *move*—are analyzed as one conceptual representation (e.g., [Cause (W, [GO_{Loc} (X, [$FROM_{Loc}$ (X, Y)], [TO_{Loc} (X, Z)])])]) with thematic roles `(agent,theme,goal,source)`). Other possibilities—*help*—are analyzed as different conceptual representation (e.g., [Cause (W, [GO_{Ident} (X, [$TOWARD_{Ident}$ (X, HELPED)], [$WITH_{Poss}$ (Z)])])] with thematic roles `(agent,theme,mod-poss)`).

We now turn to the construction of multilingual resources for these applications using the HowNet Conceptual Database.

3 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 Figure 3.

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 成为 (*become* under the "static" node V.1.1.1) and 开始 (*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*.

Associated with each HowNet concept 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*.[4] The

[3]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 frequently 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.

[4]Thematic-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|

V.2 |act|
V.2.1 |ActGeneral|
V.2.1.1 |start|
V.2.1.2 |do|
V.2.1.3 |DoNot|
V.2.1.4 |Cease|
V.2.1.5 |Wait|
V.2.2 |ActSpecific|
V.2.2.1 |AlterGeneral|
V.2.2.2 |AlterSpecific|
V.2.3 |AlterRelation|
V.2.3.1 |AlterIsa|
V.2.3.2 |AlterPossession|
V.2.3.3 |AlterComparison|
V.2.3.4 |AlterFitness|
V.2.3.5 |AlterInclusion|
V.2.3.6 |AlterConnection|
V.2.3.7 |AlterCauseResult|
V.2.3.8 |AlterLocation|
V.2.3.9 |AlterTimePosition|

V.2.4 |AlterState|
V.2.4.1 |AlterPhysical|
V.2.4.2 |AlterStateNormal|
V.2.4.3 |AlterStateGood|
V.2.4.4 |AlterQuantity|
V.2.4.5 |AlterStateBad|
V.2.4.6 |AlterMental|
V.2.5 |AlterAttribute|:
V.2.5.1 |MakeHigher|
V.2.5.2 |MakeLower|
V.2.5.3 |AlterAppearance|
V.2.5.4 |AlterMeasurement|
V.2.5.5 |AlterProperty|
V.2.6 |MakeAct|:
V.2.6.1 |CauseToDo|
V.2.6.2 |CauseNotToDo|
V.2.6.3 |use|

Figure 3: HowNet Verb Hierarchy

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

4 Mapping Between Chinese HowNet and English EVCA

The mapping between Chinese HowNet and English EVCA involves three steps:

(1) Produce all possible English Optilex *glosses* (translations) for all 12342 Chinese verbs in HowNet and associate each of the resulting 41,324 Chinese-English pairs with one or more of the 709 HowNet concepts.

Example: The multiply ambiguous Chinese verb 拉 (la) has several different Optilex glosses (*slash*, *cut*, *chat*, *pull*, *drag*, *transport*, *move*, *raise*, *help*, *implicate*, *involve*, *defecate*, *press-gang*) and is associated with multiple HowNet concepts: |Transport|, |Attract|, |Excrete|, |Force|, |Help|, |Include|, |Pull|, |Recreation|, and |Talk|.

(2) Associate each verb-to-concept candidate with one or more of the 500 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).

Example: The Chinese verb 拉 (la) is associated with 22 EVCA classes: Admire (31.2.b, *implicate*, *involve*); 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*).

[Dorr et al., 1998].

(3) For each HowNet concept, partition the associated Chinese-English pairs into groups whose English glosses correspond to EVCA classes. This requires three steps:

   a. 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.

   b. 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. The matching procedure relies on correlations derived from approximately 200 seed mappings. A subset of these mappings are shown in Figure 4.[5]

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

Example: Two of the HowNet concepts associated with the multiply ambiguous Chinese verb 拉 (la) are |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 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 Equip semantic class and *transport* in the Send semantic class; the corresponding semantic representations (LCS's) are assigned from the EVCA database.

The process of associating EVCA classes with Chinese verbs relies on a massive filtering of spurious class assignments. For example, the |Establish| HowNet 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 example of an 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.

5 Preliminary Results

The histogram in Figure 5 characterizes the number of EVCA classes required for coverage of 709 HowNet concepts. The majority of HowNet concepts are covered by 1-4 EVCA classes, although

[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.

Hownet	**EVCA/LCS-Based Roles**															
Roles	**ag**	**th**	**exp**	**goal**	**src**	**perc**	**loc**	**info**	**pred**	**prop**	**Instr**	**Poss**	**Pred**	**Purp**	**Loc**	**Ben**
agent	278	77	32	1	2	3	0	0	0	0	4	7	0	11	0	4
beneficiary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
cause	0	0	0	1	0	4	0	0	1	6	4	7	1	11	0	0
content	0	31	1	2	2	14	0	20	3	6	3	0	1	3	0	1
contrast	0	2	0	1	0	1	0	0	0	0	0	1	0	0	0	0
experiencer	13	32	33	0	0	0	0	0	0	0	0	0	0	0	0	0
isa	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0
location	0	1	0	1	0	0	6	0	0	1	2	0	0	0	2	0
manner	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
partner	0	2	0	0	3	3	0	0	0	0	0	11	0	0	0	0
partof	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
patient	0	122	7	7	0	8	0	0	0	0	0	0	0	0	0	0
possession	0	28	0	0	1	2	0	0	0	0	0	3	0	0	0	0
purpose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
range	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
relevant	15	4	4	0	0	1	1	0	0	0	0	0	0	0	0	0
result	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
scope	0	1	0	0	0	2	1	0	0	0	1	2	3	0	0	0
source	0	4	0	0	16	0	0	0	0	0	0	0	0	0	3	1
target	0	7	12	27	1	17	0	0	0	3	0	2	0	0	1	1
ContentProduct	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LocationFin	0	0	0	31	0	0	8	0	1	0	0	2	2	0	8	1
LocationIni	0	0	0	0	24	0	2	0	0	0	0	0	0	0	0	0
StateFin	0	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0
StateIni	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0

Figure 4: Seed Table for mapping HowNet Roles into EVCA Roles

a small number of concepts are represented by as many as 22. These numbers correspond to 8089 EVCA-classified Chinese entries; about 43% of the number of potential entries. The remaining 10441 entries are accounted for through the use of an approach to compensating for resource deficiencies (e.g., the lack of Optilex translations for certain Chinese verbs). Section 7 provides the refined results based on this approach.

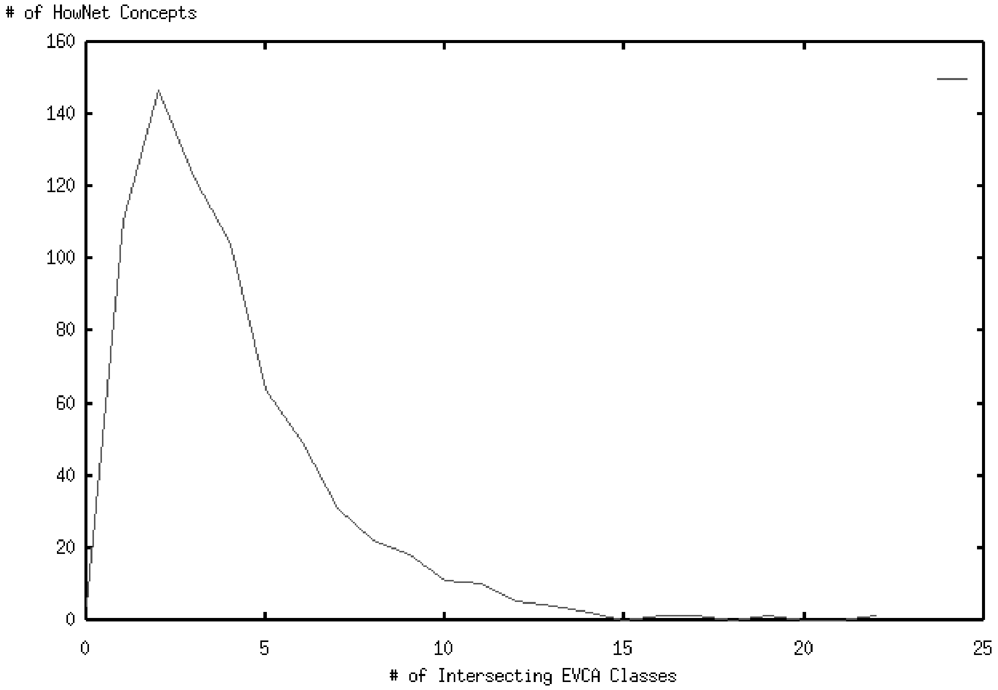
6 Compensating for Resource Deficiencies

As part of our effort to produce a complete alignment between HowNet and EVCA, we built an EVCA-based canonical entry for each of the 709 HowNet concepts so that we could compensate for certain types of resource deficiencies. The canonical entry is specified as an EVCA class coupled with its associated prototype verb. This entry was automatically generated according to the highest ranking EVCA class using steps 3.a and 3.b in Section 2. Each canonical entry was hand-verified (at a rate of 80 per hour for 709 classes). In most cases, prototype word 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 entries is given in Figure 6.

We use these canonical entries to compensate for any gaps that arise in our three online resources: (1) EVCA, (2) Optilex, and (3) HowNet. We will describe each of these, in turn.

6.1 EVCA Gaps

An EVCA gap is detected when an Optilex verb does not occur in EVCA. When this occurs, the canonical entry is automatically used as the appropriate EVCA classification for the verb.



EVCA:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
HowNet:	4	111	147	123	104	64	49	31	22	18	11	10	5	4	2	0	1	1	0	1	0	0	1

Figure 5: Distribution of HowNet Concepts by Number of Intersecting EVCA Classes

HowNet Concept	**Canonical Entry**
|Transport|	11.1 Send, *transport*
|BeNot|	22.2.a Amalgamate, *oppose*
|Help|	13.4.2 Equip, *help*
|Moisten|	45.4.a Change of State, *facilitate*
|Excrete|	40.1.2 Breathe, *bleed*
|Apologize|	32.2.a Long, *apologize*
|PlayDown|	33.b Judgment, *belittle*
|Naming|	29.3 Dub, *name*
|Choose|	29.2.c, *choose*
|Announce|	37.7.b Say, *announce*
|Mean|	37.7.a Say, *signify*
|Communicate|	37.9.c Advise *inform*

Figure 6: Sample of Canonical Entries for Filling Resource Gaps

For example, one Optilex gloss associated with HowNet concept |Establish| (for the verb 重建 (chongjian)) is *reconstruct*, which does not occur in EVCA. This is a case where the canonical entry (29.2.c Characterize, *establish*) is associated with the verb.

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 Class 29.2.c and the LCS associated with the verb *establish* can then be linked to this Chinese verb.

6.2 Optilex Gaps

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

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

In addition to their usefulness in handling of gaps in our lexical resources, the canonical entries 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 entry (31.3.d Marvel, *suffer*). Thus, the canonical entry was assigned to this verb.

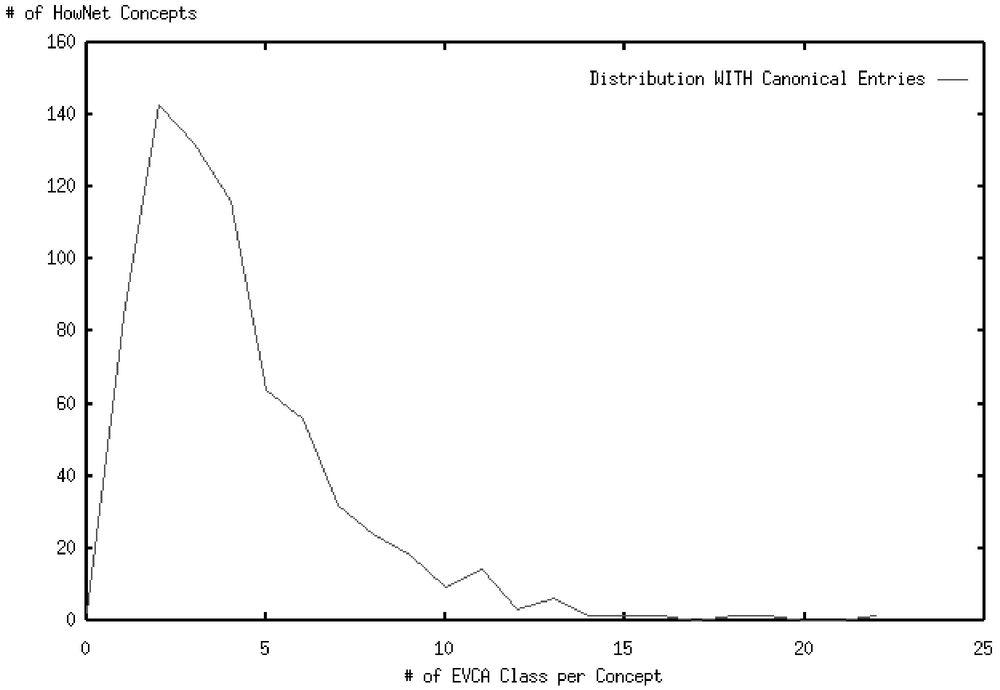
A similar approach is used for unknown or misspelled words. For example, the translation of 输送 (shusong) as in Optilex is misspelled as *tranport*. Because this verb occurs in the |Transport| class, the canonical entry (11.1 Send, *transport*) was assigned to this verb.

6.3 HowNet Gaps

In some cases, the HowNet classification 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 |Decorate|. 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 semantic-role specification `_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 frequency within the class. Out of the 17 entries associated with the |Decorate| concept, only two of them (the two misclassed Chinese verbs) are associated with an EVCA class that is not 9.9 or 9.8. As in the gap-recovery described approaches above, the misclassed verbs are associated with the canonical entry (9.8.b Fill, *decorate*).[6]

[6] Ultimately, the misclassed verbs should be disassociated from the HowNet concept, but there is currently no way to tease apart such cases from the Optilex gaps. Thus, the two are treated identically.



EVCA:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
HowNet:	2	84	143	132	116	64	56	32	24	18	9	14	3	6	1	1	1	0	1	1	0	0	1

Figure 7: Revised Distribution of HowNet Concepts by Number of Intersecting EVCA Classes using Canonical Entries

7 Revised Results

Using the canonical entries, we have achieved a more refined EVCA-to-HowNet mapping, providing an increase EVCA-classified Chinese words from the previous 8089 entries to the current expanded set of 17284 EVCA-classifed Chinese words. Figure 7 shows that adding new words has not degraded the degree of partitioning.

Examples of the HowNet partitionings into EVCA classes are given in Figure 8, 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 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 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.

8 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

HowNet Concept	EVCA Class(es)
|Transport|	11.1 Send (100%)
|BeNot|	22.2.a Amalgamate (100%)
|Help|	13.4.2 Equip (100%)
|Moisten|	45.4.a Change of State (100%)
|Excrete|	40.1.2 Breathe (100%)
|Apologize|	32.2.a Long (100%)
|PlayDown|	33.b Judgment (89%), 31.2.c Admire (11%)
|Naming|	29.3 Dub (70%), 37.3.b Speaking (30%)
|Choose|	29.2.c Characterize (99%), 30.2.a.ii Sight (1%)
|Announce|	37.7.b Say (92%), 48.1.1.a Appear (6%), 45.4.a Change of State (2%)
|Mean|	37.7.a Say (83%), 29.2.d Characterize (9%), 25.4 Transcribe (4%), 47.8.b Contiguous Location (4%)
|Communicate|	37.9.c Advise (64%), 22.1.b Mix (9%), 9.2.ii Put (9%), 36.1.c Correspond (9%), 45.4.a Change of State (9%)

Figure 8: Examples of HowNet Partitionings with Respect to EVCA

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 709 Chinese HowNet concepts and the 500 EVCA classes. Because each Chinese-English link is additionally associated with a lexical conceptual structure, we see this resource as the first step toward producing a general repository 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 semantic-role specifications 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 classification is associated with a WordNet sense WordNet [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

[Dang et al., 1998] Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig, 1998. 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.

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

[Dorr et al., 1998] Dorr, Bonnie J., Nizar Habash, and David Traum, 1998. A Thematic Hierarchy for Efficient Generation from Lexical-Conceptal Structure. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*. Langhorne, PA.

[Dorr and Jones, 1999] Dorr, Bonnie J. and Douglas Jones, 1999. Acquisition of semantic lexicons: Using word sense disambiguation to improve precision. In Evelyne Viegas (ed.), *Breadth and Depth of Semantic Lexicons*. Norwell, MA: Kluwer Academic Publishers.

[Dorr and Katsova, 1998] Dorr, Bonnie J. and Maria Katsova, 1998. Lexical Selection for Cross-Language Applications: Combining LCS with WordNet. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*. Langhorne, PA.

[Dorr et al., 2000] Dorr, Bonnie J., Gina-Anne Levow, Dekang Lin, and Scott Thomas, 2000. Chinese-English Semantic Resource Construction. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*. Athens, Greece.

[Jones et al., 1994] Jones, Douglas, Robert Berwick, Franklin Cho, Zeeshan Khan, Karen Kohl, Naoyuki Nomura, Anand Radhakrishnan, Ulrich Sauerland, and Brian Ulicny, 1994. Verb Classes and Alternations in Bangla, German, English, and Korean. Technical report, Massachusetts Institute of Technology.

[Langkilde and Knight, 1998a] Langkilde, Irene and Kevin Knight, 1998a. Generating Word Lattices from Abstract Meaning Representation. Technical report, Information Science Institute, University of Southern California.

[Langkilde and Knight, 1998b] Langkilde, Irene and Kevin Knight, 1998b. Generation that Exploits Corpus-Based Statistical Knowledge. In *Proceedings of COLING-ACL '98*.

[Langkilde and Knight, 1998c] Langkilde, Irene and Kevin Knight, 1998c. The Practical Value of N-Grams in Generation. In *International Natural Language Generation Workshop*.

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

[Lin, 1998] Lin, D., 1998. Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*. Granada, Spain.

[Miller and Fellbaum, 1991] Miller, George A. and Christiane Fellbaum, 1991. Semantic Networks of English. In Beth Levin and Steven Pinker (eds.), *Lexical and Conceptual Semantics, Cognition Special Issue*. Amsterdam, The Netherlands: Elsevier Science Publishers, B.V., pages 197–229.

[Nomura et al., 1994] Nomura, Naoyuki, Douglas A. Jones, and Robert C. Berwick, 1994. An architecture for a universal lexicon: A case study on shared syntactic information in Japanese, Hindi, Ben Gali, Greek, and English. In *Proceedings of COLING-94*. Kyoto, Japan.

[Olsen et al., 1998] Olsen, Mari Broman, Bonnie J. Dorr, and Scott C. Thomas, 1998. Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*. Langhorne, PA.

[Palmer and Rosenzweig, 1996] Palmer, Martha and Joseph Rosenzweig, 1996. Capturing motion verb generalizations with synchronous tags. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*. Montreal, Quebec, Canada.

[Palmer and Wu, 1995] Palmer, Martha and Zhibao Wu, 1995. Verb Semantics for English-Chinese Translation. *Machine Translation*, 10(1–2):59–92.

[Procter, 1978] Procter, P., 1978. *Longman Dictionary of Contemporary English*. London: Longman.

[Resnik, 1995] Resnik, Philip, 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*. Montreal, Canada.

[Saint-Dizier, 1996] Saint-Dizier, Patrick, 1996. Semantic Verb Classes Based on 'Alternations' and on WordNet-like Semantic Criteria: A Powerful Convergence. In *Proceedings of the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases*. Toulouse, France.

[Weinberg et al., 1995] Weinberg, Amy, Joseph Garman, Jeffery Martin, and Paola Merlo, 1995. Principle-Based Parser for Foreign Language Training in German and Arabic. In Melissa Holland, Jonathan Kaplan, and Michelle Sams (eds.), *Intelligent Language Tutors: Theory Shaping Technology*. Hillsdale, NJ: Lawrence Erlbaum Associates, pages 23–44.

[Zhendong, 1988a] Zhendong, Dong, 1988a. Enlightment and Challenge of Machine Translation. *Shanghai Journal of Translators for Science and Technology*, 1:9–15.

[Zhendong, 1988b] Zhendong, Dong, 1988b. Knowledge Description: What, How and Who? In *Proceedings of International Symposium on Electronic Dictionary*. Tokyo, Japan.

[Zhendong, 1988c] Zhendong, Dong, 1988c. MT Research in China. In *Proceedings of International Conference on New Directions in Machine Translation*. Budapest. Also in New Directions in Machine Translation, 4 Distributed Language Translation edited by Dan Maxwell, Klaus Schubert and Toon Witkam, Foris Publications, Dordrecht.