Construction of a Chinese-English Verb Lexicon for Embedded Machine Translation in Cross-Language Information Retrieval

Bonnie Jean Dorr (bonnie@umiacs.umd.edu) and Gina-Anne Levow (gina@umiacs.umd.edu)
University of Maryland Institute for Advanced Computer Studies, College Park, MD 20742

Dekang Lin (lindek@cs.ualberta.edu)
University of Alberta, Edmonton, Alberta, Canada, T6G 2E8

October 1, 2001

Abstract. This paper addresses the problem of automatic acquisition of lexical knowledge for rapid construction of MT engines multilingual applications. We describe new techniques for large-scale construction of a Chinese-English verb lexicon and we evaluate the coverage and effectiveness of the resulting lexicon for a structured MT approach that is embedded in a cross-language information retrieval system. Leveraging off an existing Chinese conceptual database called HowNet and a large, semantically rich English verb database, we use thematic-role information to create links between Chinese concepts and English classes. We apply the metrics of *recall* and *precision* to evaluate the coverage and effectiveness of the linguistic resources. The results of this work indicate that: (1) we are able to obtain reliable Chinese-English entries both with and without pre-existing semantic links between the two languages; (2) if we have pre-existing semantic links, we are able to produce a more robust lexical resource by merging these with our semantically rich English database; (3) In our comparisons with manual lexicon creation, our automatic techniques were shown to achieve 62% precision, compared to a much lower precision of 10% for arbitrary assignment of semantic links.

Keywords: Machine translation, Cross-language information retrieval, Embedded MT Resources, Chinese-English lexicons, Thematic roles, Lexical acquisition

1. Introduction

The growing quantity of online multilingual information, e.g., on the Web, has created an urgent need for rapid construction of lexical resources. Automatic and semi-automatic techniques for lexical acquisition are more critical now than ever before as it becomes infeasible to produce adequate semantic representations on a large scale by human labor alone. We describe a new linguistically-based approach to large-scale construction of a semantic lexicon for Chinese verbs. This resource is used by a machine translation module whose sub-components are embedded in a system for English-Chinese cross-language information

retrieval. We focus specifically on the problem of compensating for gaps in the existing resources and we apply the metrics of *recall* and *precision* to evaluate the coverage and effectiveness of our lexicon.

We leverage off three existing resources: (1) a classification of English verbs based on EVCA (English Verbs Classes and Alternations) (Levin, 1993), in a significantly enhanced form called the LCS Verb Database (LVD) (Dorr, 2001); (2) a Chinese conceptual database called HowNet[1] (Dong, 1988c), (Dong, 1988b), (Dong, 1988a); and (3) a large machine readable Chinese-English dictionary called Optilex.[2] We use thematic-role information (e.g., a mapping between the HowNet "Patient" and the LVD-based "Th(eme)") to create links between Chinese concepts and English classes. Each Chinese-English link is additionally associated with a sense from WordNet (Miller and Fellbaum, 1991), (Fellbaum, 1998), thus producing a new Asian companion to the current (Euro)WordNet initiative. Finally, we use the LVD-based semantic classes, our thematic role mapping, and canonical English words to produce a large set of Lexical Conceptual Structures used in the embedded machine-translation components.

Until this year, HowNet contained no English translations. Thus, our initial experiments used Optilex to produce candidate English translations. In the latest version of HowNet (Dong, 2000), the English translations are included; however, our work has provided the basis for increasing *recall*—acquisition of thousands of correct Chinese-English entries that do not currently exist in HowNet—and, moreover, it has provided a link into the semantic classes underlying a large English conceptual database. Since the new HowNet was released, we have been able to execute a more accurate evaluation of our Chinese-English links—in particular, we use the English translations in HowNet to determine the *precision* of our approach (overall accuracy of the Chinese-English links we have already automatically). Finally, given that our initial work did not make use of the English translations in HowNet, we expect those same 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 (see, for example, (Hovy, 1998)).

The lexicons resulting from our acquisition approach are used for determining word senses in a machine translation (MT) module whose sub-components are embedded in a cross-language information retrieval (CLIR) system—allowing the user to access Chinese documents using

[1] Available at: http://www.keenage.com

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

English as their query language. The importance of determining word senses in embedded MT 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.

The next section relates our work to that of other researchers who have investigated the problem of mapping across semantic hierarchies for construction of MT resources. Following this, we describe *Lexical Conceptual Structure* (LCS)—an interlingual representation used in our cross-language applications—and we illustrate how this representation is used as the basis of our embedded MT approach. After this, we describe the structure of our existing lexical resources, HowNet and LVD, and we show that mapping between these two resources results in a new resource: a Chinese LCS lexicon used for our embedded MT approach. Finally, we demonstrate that our automatic acquisition techniques provide a framework for compensating for gaps in this new resource. We evaluate the coverage and accuracy of the new Chinese LCS lexicon with respect to the pre-existing LVD and HowNet resources. We conclude that: (1) we are able to obtain reliable Chinese-English entries both with and without pre-existing semantic links between the two languages; (2) if we have pre-existing semantic links, we are able to produce a more robust lexical resource by merging these with our semantically rich English database; (3) In our comparisons with manual lexicon creation, our automatic techniques were shown to achieve 62% precision, compared to a much lower precision of 10% for arbitrary assignment of semantic links.

2. Mapping Across Semantic Hierarchies

Several researchers have investigated the problem of assigning class-based senses to verbs (Dorr et al., 1997), (Palmer and Wu, 1995), (Palmer and Rosenzweig, 1996), (Palmer et al., 1997) 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), (Fellbaum, 1998). 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).

Our work extends the techniques described by (Palmer and Wu, 1995), which used a concept space to produce a hierarchical organization of Chinese verbs. We adopt a technique that is similar in flavor to the intersective-class approach of (Dang et al., 1998), with the following

extensions: (1) The construction of an entry for all EVCA verbs—plus those in the enhanced LVD—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) rather than one; (4) Mappings between Chinese and English thematic roles; and (5) Hooks into WordNet 1.6 senses for both languages. We view this work to be a significant enhancement to current efforts in resource building for multilingual applications such as the PropBank effort (Palmer et al., 2001).

Our approach to mapping across semantic hierarchies 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 LVD entries. Each Chinese-English link is subsequently associated with a sense from WordNet 1.6 (Miller and Fellbaum, 1991), (Fellbaum, 1998).

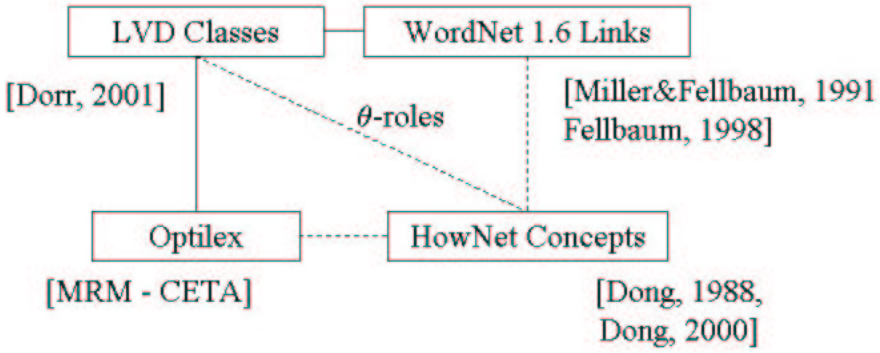


Figure 1. Relation Between Existing Resources and New Mappings

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 techniques. The most critical of these is the one labeled θ-roles (shorthand for "thematic roles"), which associates LVD classes with HowNet Concepts. The remaining two dotted-line mappings are "transitive closure" byproducts of the other mappings: Once the thematic-role mapping associates LVD verbs with HowNet verbs, each HowNet verb is associated with Optilex-based English *glosses* (translations) and WordNet 1.6 Senses. Section 4.3 provides more details about how the correspondences in Figure 1 are derived. Next, we describe the use of our lexical

representations as the basis of the embedded MT modules used in our CLIR system.

3. Embedded MT: Lexical Knowledge and Translation Approach

One of the types of knowledge that must be captured for use in MT applications is linguistic knowledge at the level of the lexicon, which covers a wide range of information types, such as verbal subcategorization for events (e.g., that a transitive verb such as "hit" occurs with an object noun phrase), featural information e.g., that the direct object of a verb such as "frighten" is animate), thematic information (e.g., that "John" is the agent in "John hit the ball"), and lexical-semantic information (e.g., that spatial verbs such as "throw" are conceptually distinct from verbs of possession such as "give"). By modularizing the lexicon, we treat each information type separately, thus allowing us to vary the degree of dependence on each level, so that we can address the question of how much knowledge is necessary for the success of the particular NLP application. We focus on lexical-semantic knowledge for the remainder of this section.

3.1. Lexical Conceptual Structure

The most intricate component of lexical knowledge is the lexical-semantic information, which is encoded in the form of Lexical Conceptual Structure (LCS) as formulated by Dorr (Dorr, 1993), (Dorr, 1994) based on work by Jackendoff (Jackendoff, 1983), (Jackendoff, 1990). This representation is used in the NLP component of an implemented foreign language tutoring system (Dorr et al., 1997) and an interlingual machine translation system (Olsen et al., 1998).

The LCS approach views semantic representation as a subset of conceptual structure, the language of mental representation, as in (Jackendoff, 1983), (Jackendoff, 1990). This approach includes *types* such as Event and State, which are specialized into *primitives* such as GO, STAY, BE, GO-EXT, and ORIENT. We add a manner component [$_{\text{Manner}}$ JOG+INGLY] to distinguish among verbs, e.g. *run*, *walk*, and *jog*. The full representation for *John jogged to school* is therefore the representation below, roughly 'John went to the school by jogging'.[3]

[3] Note that this representation of the surface sentence does not include the FROM component shown in Figure 8 since there is no *from* phrase in this particular example.

$$
\begin{array}{l}
[_{\text{Event}}\ \text{GO}_{\text{Loc}} \\
\quad ([_{\text{Thing}}\ \text{JOHN}], \\
\quad [_{\text{Path}}\ \text{TO}_{\text{Loc}} \\
\qquad ([_{\text{Thing}}\ \text{JOHN}], \\
\qquad\ [_{\text{Position}}\ \text{AT}_{\text{Loc}}\ ([_{\text{Thing}}\ \text{JOHN}], [_{\text{Thing}}\ \text{SCHOOL}])])], \\
\quad [_{\text{Manner}}\ \text{JOG+INGLY}])]
\end{array}
$$

Given that we have mapped the HowNet entries to LVD entries, we are able to produce the LCS's for Chinese in the same way that we produce entries for English. Figure 2 presents the result of building an LCS entry for the Chinese verb 接触 (*to touch*).[4]

```
(DEFINE-WORD
   :DEF_WORD "touch"
   :CLASS "47.87.f"
   :THETA_ROLES "_th_loc"
   :WN_SENSE (00820743 01832678 00820504)
   :LANGUAGE ENGLISH
   :LCS
      (be loc (* thing 2) (at loc (thing 2) (* thing 11)))
      (touch+ingly 26))
```

Figure 2. Lexicon Entry for *touch* in LVD Class 47.8.f

The number 47.8.f is a subcase of a Levin-based class for "Verbs of Contiguous Location." The thematic grid `_th_loc` indicates that the *theme* and the *location* are both obligatory and should be annotated as such in the instantiated LCS. This list structure recursively associates logical heads with their arguments and modifiers. The logical head is represented as a primitive/field combination, e.g., BE_{Loc} is represented as `(be loc ...)`. The arguments for BE are `(thing 2)` and `(at loc ...)`.

The variables in the representation map between LCS positions and their corresponding thematic roles. In the LCS framework, thematic roles provide semantic information about properties of the argument and modifier structures. In the example above, the numbers 2, 11, and 26 correspond to the roles agent (`th`), theme (`loc`), and manner (`manner`), respectively. These numbers enter into the construction of LCS entries: they correspond to argument positions in the LCS template (extracted using the class/grid/verb specification). Information is filled into the LCS template using these numbers, coupled with the thematic grid tag for the particular word being defined.

[4] The English translation is not used during the MT process; we store it in lexical entries for convenience only (i.e., readability of the lexicon by English speakers).

3.2. LCS-Based Machine Translation

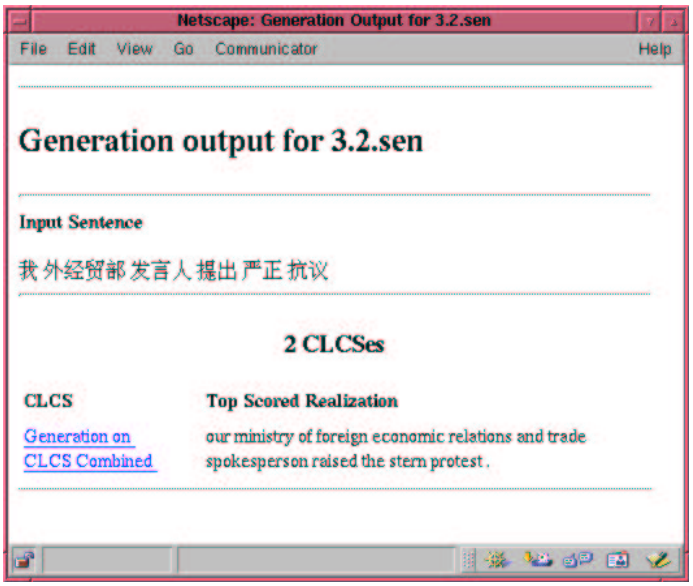


Figure 3. Translation of a Chinese sentence into English

Our approach to machine translation is interlingual, where the target-language lexicon is searched for appropriate lexical items matching a LCS representation. A screen snapshot of a translation by our system on a Chinese example is shown in Figure 3. This translation is more fluent than its literal (gisted) equivalent: *Our Foreign_Economic_Trade_Ministry spokesperson lodge stern protest*. We describe the analysis and generation modules of this system; it is the interlingua and generation modules that we have embedded in our CLIR system (to be described in Section 3.3).

Analysis in our MT system relies on an in-house parser called REAP to produce English parse trees on a large scale, and Chinese on a smaller scale. One of the benefits of this parser is its ease of portability to new languages (Weinberg et al., 1995). The parser output is semantically analyzed, producing an LCS representation, which serves as the interlingua.

Generation from the LCS is achieved by means of a system called Oxygen (Habash, 2000), a variant of Nitrogen (Langkilde and Knight, 1998a), (Langkilde and Knight, 1998b), (Langkilde and Knight, 1998c) that uses our own linearizer implemented in Lisp with Nitrogen's statistical extraction module and Nitrogen's morphological generation engine. 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 unordered tree where the root is an event concept and each child is linked by a relation. Generation of target-language sentences from LCS is described in detail in (Dorr et al., 1998).

Thematic-role information and its use in generation of natural language translations are described in (Dorr et al., 1998) as a component of a precursor to the Oxygen system. Specifically, 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*.[5] Our approach provides a framework for disambiguation of such cases. Certain of these possibilities—*transport* and *move*—are analyzed as 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)`.

3.3. Embedded MT: Cross-Language Information Retrieval

A common approach to transforming documents and queries in different languages into a common indexing space for CLIR is to translate either the document or the queries into a single language (Oard and Dorr, 1996). Due to the time and computational expense of translation, query translation is often preferred over document translation, although document translation often produces superior results. A variety of methods have been applied to translation term selection to cope with the problem of translation ambiguity, where one source language term translates into more than one target language alternative (Oard, 1998), (Ballesteros and Croft, 1997), (Hull and Grefenstette, 1996). These techniques include selecting every translation, the first N translations according to some ranking strategy, and those that co-occur with candidate translation of other terms in the query.

[5] 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.

The technique we adopt—LCS Query Translation (LQT)—uses the same representation that serves as the interlingua in our MT system. LQT employs the LCS MT approach to transform the user's query into the document language for information retrieval. In our current system, we use a structured syntax interface, called MADLIBS (Maryland Action Detection / Language-Independent Browsing and Search), to ensure that the user's query is fully analyzable for application of LQT. Specifically, for each word in the LCS lexicon we produce a simple "composed" LCS for each thematic role structure associated with the word, instantiating each role position with a dummy lexical entry, e.g. "someone-1" or "something-2". We then convert this version of the LCS into a template for user input, by generating a syntactically correct surface sentence realization using the Oxygen module of the MT system described above. We now have a mapping from surface forms to the interlingual structures used in our LQT approach.

The user interface for this system is illustrated in Figure 4. The positions in the sentence realization that correspond to the thematic roles appear as boxes for free-form user input. The interface allows querying of either English or Chinese documents; we will focus on the cross-language variant in the remainder of this discussion.

The LCS representation described in Section 3.1 is an important component of our LQT approach to IR (see Figure 5). In particular, we have embedded a structural graph-matching routine that is used both for full-blown MT (where target-language sentences are generated) as well as selection of terms for cross-language retrieval (where bags of words are produced as query terms).

To construct the query, the surface template retrieves its underlying LCS structure, complete with the input words filling the thematic role positions. This correspondence between surface form and thematic structure performs an initial phase of sense disambiguation, identifying the subset of possible senses with this argument structure. To perform translation of the query, we apply structural matching of the query against a database of LCS structures. Depending on the language choice, we consult different databases and return words with corresponding LCS structure. The structural matching also directly exploits thematic role information associated with entries in the LCS database for the language of choice.

The system permits two forms of matching: exact and relaxed, selected with the pull-down item in the interface. Exact match compares both structure and manner constants and relies only on the database selection. Relaxed match performs a second phase of processing after the structural match, employing the WordNet correspondences produced by the thematic hierarchy. This method computes similarity

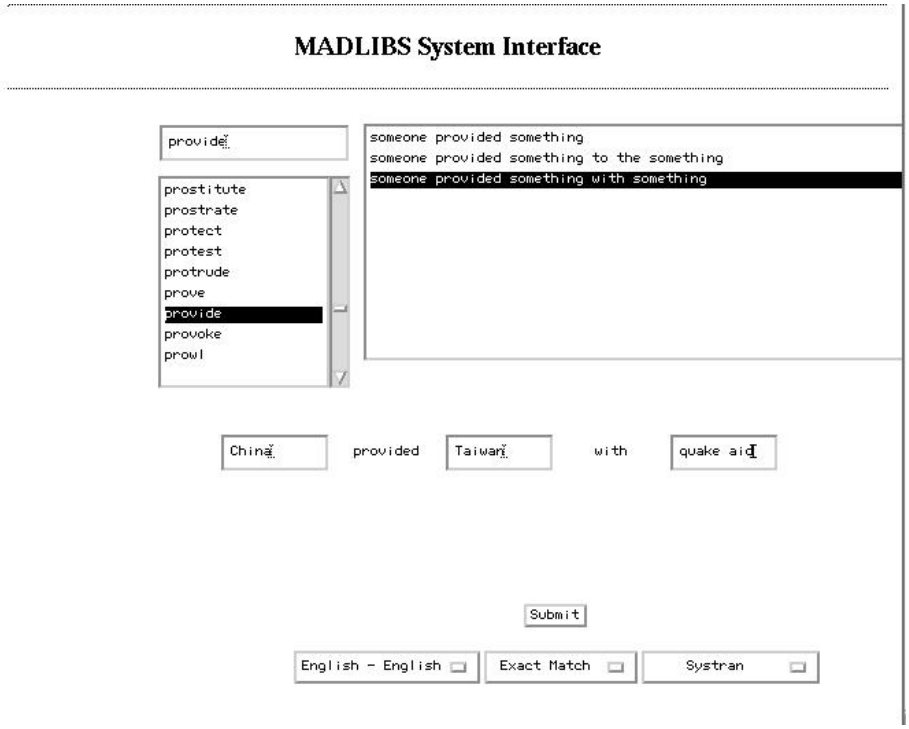


Figure 4. Structured Syntax Input Interface

between the original term and the candidate translations returned by the structural match building on Resnik's (Resnik, 1995) technique for computing taxonomic similarity. In all cases, the top N scoring candidates are returned.

We currently perform no additional analysis of noun phrases entered in the thematic role position, though a fuller treatment of nominalized events is planned. We instead apply a basic matching technique, using a lexicon built from an English-Chinese term list provided by the Linguistic Data Consortium (LDC) augmented with the result of inverting the Optilex lexicon for words with single word translations, for the Chinese document case.[6] Again, we select the top N translation alternatives.

The translation terms identified by structural match, taxonomic match and word-for-word translation form a bag of words that comprise the query to an information retrieval system. We use a version of the Inquery 3.1p1 information retrieval system from the University of Massachusetts, modified for 2-byte encodings of Chinese characters.

[6] See http://www.ldc.upenn.edu for information on LDC.

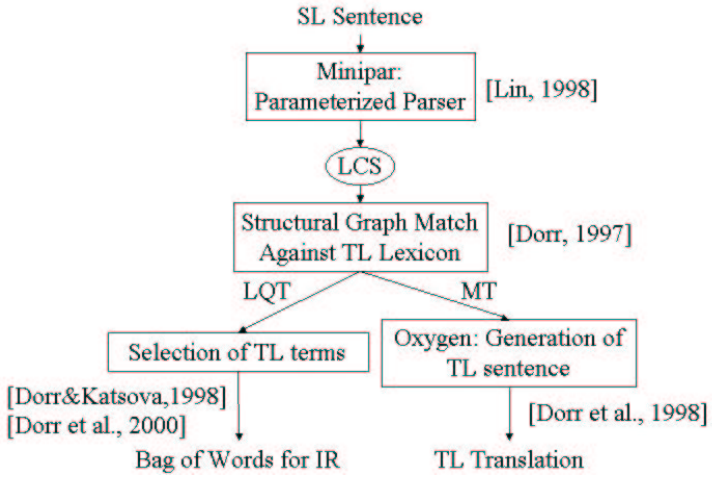


Figure 5. MT and CLIR Applications that use Multilingual Resources

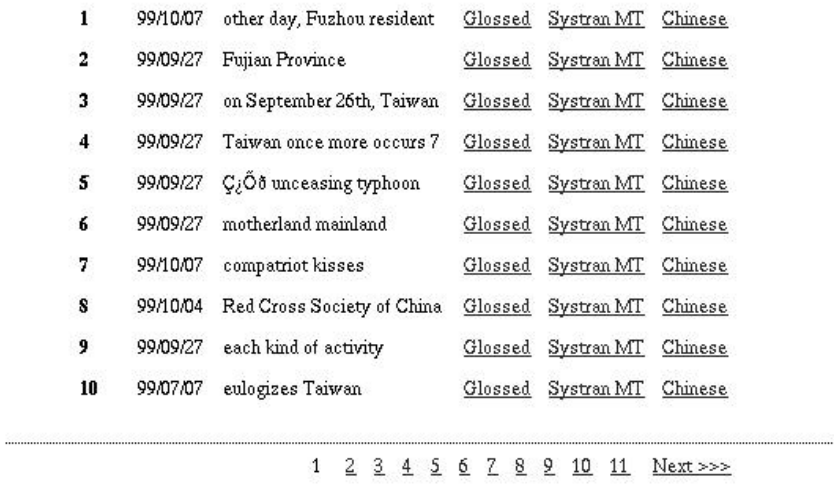


Figure 6. Selection Interface

Results are displayed interactively as well (see Figure 6), in the user's choice of source document language, Systran machine translation, or "gist", a word-for-word translation technique that provides multiple ranked alternate translations (see Figure 7). Additional details about the use of semantic representations in the CLIR system are given in (Dorr and Katsova, 1998), (Levow et al., 2000).

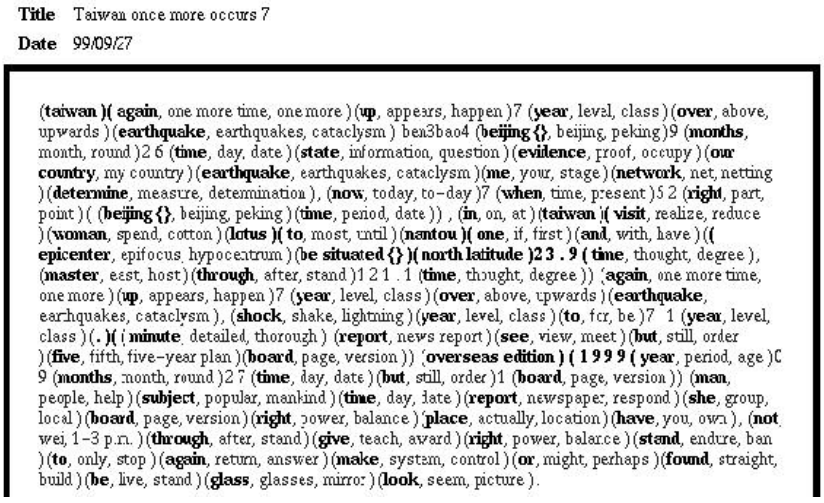


Figure 7. Presentation Interface: "Gisted" Format

4. Automatic Construction of MT Resources

We now turn to the application of automatic techniques for construction of linguistic resources for our embedded MT modules, specifically, the Chinese LCS resources used in our LQT approach. As outlined in Section 2, we leverage off HowNet and LVD, using thematic-role information to create links between Chinese concepts and English classes. We describe these two resources in more detail and then present describing our techniques for mapping between them automatically.

4.1. 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 I.

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*.

Table I. HowNet Verb Hierarchy

V.1 |static|	V.2 |act|	V.2.4 |AlterState|
V.1.1 |relation|	V.2.1 |ActGeneral|	V.2.4.1 |AlterPhysical|
V.1.1.1 |isa|	V.2.1.1 |start|	V.2.4.2 |AlterStateNormal|
V.1.1.2 |possession|	V.2.1.2 |do|	V.2.4.3 |AlterStateGood|
V.1.1.3 |comparison|	V.2.1.3 |DoNot|	V.2.4.4 |AlterQuantity|
V.1.1.4 |suit|	V.2.1.4 |Cease|	V.2.4.5 |AlterStateBad|
V.1.1.5 |inclusive|	V.2.1.5 |Wait|	V.2.4.6 |AlterMental|
V.1.1.6 |connective|	V.2.2 |ActSpecific|	V.2.5 |AlterAttribute|:
V.1.1.7 |CauseResult|	V.2.2.1 |AlterGeneral|	V.2.5.1 |MakeHigher|
V.1.1.8 |TimeOrSpace|	V.2.2.2 |AlterSpecific|	V.2.5.2 |MakeLower|
V.1.1.9 |arithmetic|	V.2.3 |AlterRelation|	V.2.5.3 |AlterAppearance|
V.1.2 |state|	V.2.3.1 |AlterIsa|	V.2.5.4 |AlterMeasurement|
V.1.2.1 |StatePhysical|	V.2.3.2 |AlterPossession|	V.2.5.5 |AlterProperty|
V.1.2.2 |StateMental|	V.2.3.3 |AlterComparison|	V.2.6 |MakeAct|:
	V.2.3.4 |AlterFitness|	V.2.6.1 |CauseToDo|
	V.2.3.5 |AlterInclusion|	V.2.6.2 |CauseNotToDo|
	V.2.3.6 |AlterConnection|	V.2.6.3 |use|
	V.2.3.7 |AlterCauseResult|	
	V.2.3.8 |AlterLocation|	
	V.2.3.9 |AlterTimePosition|	

HowNet contains 815 verb concepts altogether.[7] Each HowNet concept is associated with a thematic-role specification. For example, the verb "cure" has the thematic-roles `(agent,patient,content,tool)`. Consider the sentence *The doctor cured the man of pneumonia with antibiotics*. The roles in the specification have the following binding, respectively, for this sentence : *doctor*, *man*, *pneumonia*, *antibiotics*.[8] The thematic-role specifications are used for prioritizing candidate HowNet-LVD associations, as will be described below.

[7] We have excluded 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|).

[8] Thematic-role specifications and their use in generation of natural-language translations are described further in (Dorr et al., 1998).

4.2. LCS Verb Database

The first public release of the *LCS Verb Database* (LVD) is now available for research purposes (Dorr, 2001).[9] The verbs in this resource were borrowed initially from the publicly available online EVCA index EVCA classification.[10] While Levin's original EVCA work provides a unique and extensive catalog of verb classes, it does not define the underlying meaning components of each class. One of the main contributions of our work is that it provides a relation between Levin's classes and meaning components as defined in the LCS representation (described above in Section 3.1), thematic role information, hand-tagged WordNet synset numbers.

We built our database by subdividing the classes into a more refined set (Dorr and Olsen, 1996), extending this set to include 26 new classes (Dorr, 1997a), constructing an LCS representation for each entry (Dorr, 1997b), assigning WordNet senses to existing verbs (Dorr and Jones, 1999), and adding 3000 WordNet-tagged verbs (Green et al., 2001a), (Green et al., 2001b). Levin's original database contained 3024 verbs in 192 classes numbering between 9.1 and 57—a total of 4186 verb entries. The augmented database contains 4432 verbs in 492 classes with more specific numbering (e.g., "51.3.2.a.ii") and additional class numbers for new classes (between 000 and 026)—a total of 11000 verb entries.

Each semantic class contains a set of verbs that are related by "semantic structure" as defined in the LCS. For example, all the verbs in the semantic class of "Run" verbs have the same semantic structure but vary in their semantic content (for example, run, jog, walk, zigzag, jump, roll, etc.). The lexicon entry for the English verb *jog* in the "Run" class is shown in Figure 8. This entry includes the root form of the word, its semantic class and WordNet sense (for verbs), its thematic roles (clustered into a *grid*), and its LCS representation.

Mapping English thematic roles to their Chinese counterparts is the primary aid in selecting the appropriate verb class(es) in the LCS Database for each concept in HowNet. The next section demonstrates that it is possible to produce a lexicon by associating 709 Chinese HowNet concepts with 492 classes from the LCS Database, with a clear concept-to-class correspondence in a large majority of the cases.

[9] In the work of (Green et al., 2001a), (Green et al., 2001b), the LVD resource was referred to as Levin+. We have since adopted the more precise name "LCS Verb Database" since the resource is not a simple extension to EVCA, but an entirely new database with rich semantic structure, thematic information, WordNet senses, and a significantly more comprehensive set of verbs.

[10] The EVCA index may be found at:
ftp://linguistics.archive.umich.edu/linguistics/texts/indices/evca93.index.

```
(DEFINE-WORD
   :DEF_WORD "jog"
   :CLASS "51.3.2.a.ii"
   :THETA_ROLES "_th,src(),goal()"
   :WN_SENSE (01315785 01297547)
   :LANGUAGE ENGLISH
   :LCS
    (event go loc (* thing 2)
      ((* path from 3) loc (thing 2) (position at loc (thing 2) (thing 4)))
      ((* path to 5) loc (thing 2) (position at loc (thing 2) (thing 4)))
      (manner jog+ingly 26))
```

Figure 8. Lexicon Entry for *jog* in LVD Class 51.3.2.a.ii

4.3. Mapping Between Chinese HowNet and English LVD

Our technique for mapping between Chinese HowNet concepts and English LVD classes involves associating HowNet thematic roles with those in the LCS Database. Each HowNet concept (and each LCS) is paired with a list of thematic roles, which we call a thematic *grid*. For example, the HowNet concept |Cure| is paired with the θ-role 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 LVD database is `(ag,th,mod-poss(of))`. These roles are associated with the first three noun phrases in the sentence; the fourth noun phrase would be considered a modifier and, thus, is not in the LVD grid. Although the HowNet and LVD roles are not in a one-to-one correspondence, they can still be used for a "closest match" prioritization of candidate HowNet-LVD associations, as we will see shortly.

4.3.1. *HowNet-LVD Mapping Tasks*

There are three top-level tasks involved in mapping Chinese HowNet concepts to and English LVD classes:

(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. ["HowNet Concept + Word + Glosses" in Figure 9.]

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

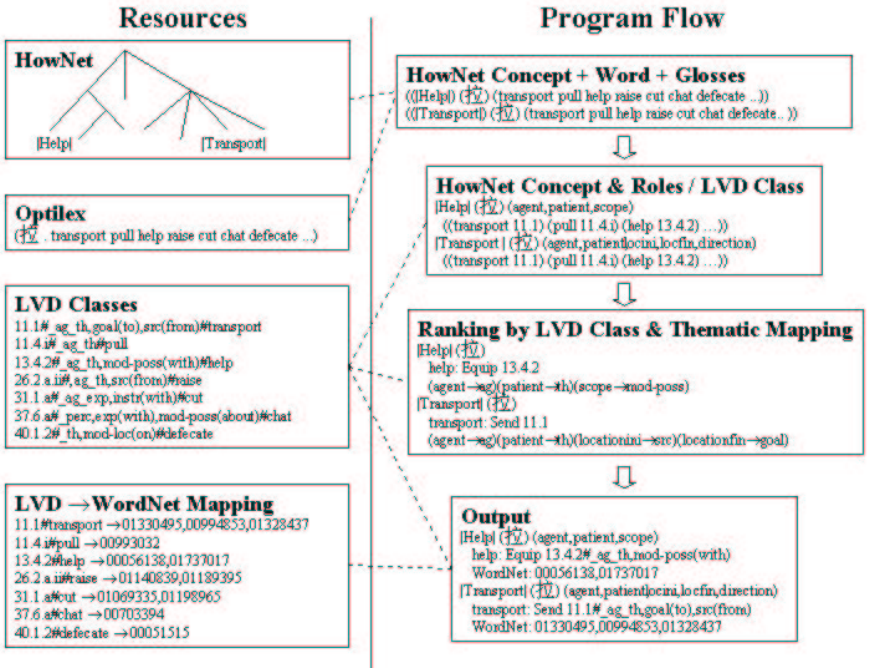


Figure 9. Resources and Processing Stages for Mapping Chinese HowNet and English LVD, including linkages to WordNet Senses

(2) Associate each verb-to-concept candidate with one or more of the 492 LVD 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).[11] ["HowNet Concept & Roles / LVD Class" in Figure 9]

Example: The Chinese verb 拉 (la) is associated with 22 LVD classes: Admire (31.2.b, *implicate*, *involve*); Amuse (31.1.b, *cut*, *move*, *transport*); Braid (41.2.2, *cut*); Breathe (40.1.2, *defecate*); Build (26.1.a, *cut*); Carry (11.4.i, *carry*, *drag*, *pull*); Chitchat (37.6.a, *chat*); Crane (40.3.2, *raise*); Cut (21.1.a, *cut*, *slash*); 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, *cut*, *pull*); Meander (47.7.a, *cut*); Play (009, *pawn*); Put (9.4.a, *raise*); Search (35.2.a, *drag*); Send (11.1, *convey*, *ship*, *smuggle*, *transport*); Slide (11.2.b, *move*); Split (23.2.b, *cut*, *pull*).

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

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

a. Order the candidate LVD classes so that the highest-ranking classes are those that contain the highest number of English verbs matching the Optilex glosses. ["Ranking by LVD Class" in Figure 9]

b. In cases where a tie-breaker is needed, reorder the candidate LVD classes according to the degree to which the thematic-role specification in HowNet concept matches that of LVD class. The matching procedure relies on correlations derived from approximately 200 seed mappings. A subset of these mappings are shown in Table II.[12] ["Ranking by Thematic Mapping" in Figure 9]

c. For each Chinese-English entry associated with the HowNet concept, assign the highest ranking candidate LVD class. ["Output" in Figure 9]

Example: Two of the HowNet concepts associated with the multiply ambiguous Chinese verb 拉 (la) are |Help| and |Transport|. The θ-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 LVD 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 LVD Class, and the mapping between the two is `(agent->ag)`, `(patient->th)`, `(scope->mod-poss)`.

On the other hand, the HowNet concept |Transport| is associated with the thematic-role specification

`(agent,patient,LocationIni,LocationFin)`

(as in *John transported the goods from Boston to New York (westward)*). This specification most closely matches that of the Send LVD Class (where 拉 (la) is translated as *transport*); thus, the HowNet concept |Transport| is associated with the Send LVD 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*

[12] 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.

in the Send semantic class. The corresponding WordNet senses are then assigned from the hand-tagged LVD database—these are senses 1 and 3 in the case of *help* [indexed as 01737017 and 00056138 in Figure 9] and senses 1, 2, and 4 in the case of *transport* [indexed as 01330495, 00994853, 01328437 in Figure 9]:

- **help:**
  Sense 1: assist
  Sense 3: aid
- **transport:**
  Sense 1: transport
  Sense 2: carry
  Sense 4: send, ship

The process of associating LVD 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 LVD classes, 29.2.c (Characterize) and 26.4.a (Create), but it initially had 29 potential LVD class assignments. One example of an LVD 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 LVD 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. Compensating for Resource Deficiencies

The techniques described in the previous section creates a bridge between entries in the Chinese HowNet conceptual hierarchy and the LVD semantic classes. We now demonstrate how the HowNet thematic role and LVD semantic class mappings are combined to produce a richer lexical resource for multilingual applications, with a focus on the use of event structure for word sense disambiguation.

Recall that our approach in Section 4.3 consists of three top-level tasks. Application of this approach resulted in 8089 LVD-classified Chinese entries—about 43% of the number of potential entries. The histogram in Table III characterizes the number of LVD classes required for coverage of 709 HowNet concepts. The majority of HowNet concepts are covered by 1-4 LVD classes, although a small number of concepts are represented by as many as 22.

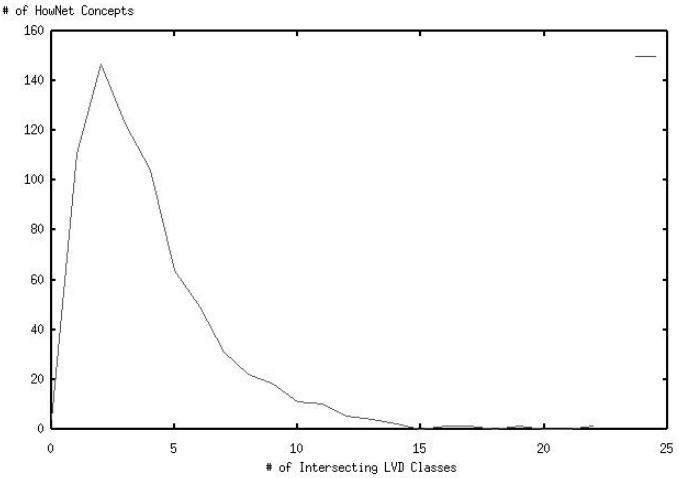
Table II. Seed Table for mapping HowNet Roles into LVD Roles

Hownet Roles	LVD Roles														
	ag	th	exp	goal	src	perc	loc	info	pred	prop	Instr	Poss	Pred	Purp	Ben
agent	278	77	32	1	2	3	0	0	0	0	4	7	0	11	4
beneficiary	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
content	0	31	1	2	2	14	0	20	3	6	3	0	1	3	1
contrast	0	2	0	1	0	1	0	0	0	0	0	1	0	0	0
experiencer	13	32	33	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
location	0	1	0	1	0	0	6	0	0	1	2	0	0	0	0
manner	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
partof	0	1	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
possession	0	28	0	0	1	2	0	0	0	0	0	3	0	0	0
purpose	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
relevant	15	4	4	0	0	1	1	0	0	0	0	0	0	0	0
result	0	1	1	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
source	0	4	0	0	16	0	0	0	0	0	0	0	0	0	1
target	0	7	12	27	1	17	0	0	0	3	0	2	0	0	1
ContentProduct	0	9	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	1
LocationIni	0	0	0	0	24	0	2	0	0	0	0	0	0	0	0
StateFin	0	0	0	5	0	1	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

The remaining 10441 entries are accounted for through the use of techniques that allow us to compensate for resource deficiencies (e.g., the lack of Optilex translations for certain Chinese verbs). These techniques allow us to produce a more complete alignment between HowNet and LVD.

In order to induce this enhanced version of the algorithm, we built an LVD-based canonical specification for each of the 709 HowNet concepts so that we could compensate for certain types of resource deficiencies.

Table III. Distribution of HowNet Concepts by Number of Intersecting LVD Classes



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

The canonical specification consists of an class coupled with its associated prototype verb. These canonical specifications provide a mapping between a HowNet concept and a LVD class/prototype-verb pair.

Each canonical specification was automatically generated according to the highest ranking LVD class using steps 3.a and 3.b in Section 4.3. 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 Table IV.

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

Table IV. Sample of Canonical Specifications for Filling Resource Gaps

HowNet Concept	**Canonical Specification**
|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*

5.1. LVD Gaps

An LVD gap is detected when an Optilex verb gloss for a Chinese verb does not occur in LVD. When this occurs, the canonical specification for the Chinese verb is automatically used to assign the verb an appropriate LVD 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 LVD. 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 LVD gaps is that it allows us to enhance our LVD resource (and, additionally, the original EVCA index). For example the verb *reconstruct* can now be added to LVD Class 29.2.c, on a par with the previously classified LVD verb *establish*.

5.2. Optilex Gaps

An Optilex gap occurs when a particular translation for a Chinese verb is missing. For example, the verb 摆布 (baibu) has only one Optilex gloss: *manipulate*. However, the word 摆布 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 roles and LVD thematic grid; and (2) correlations between the gloss under question and *other* HowNet concepts. In this particular example, the thematic grid for *manipulate* in LVD is `(ag,exp,instr)`, which is ranked low (11th out of 28) with respect to the roles `(agent, patient)` associated with the HowNet concept |Decorate|. By contrast, this same LVD 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 摆布 (baibu) is missing from Optilex. As in the case with LVD 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 LVD 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 (shusong) as in Optilex is misspelled as *tranport*. Because this verb is associated with HowNet's |Transport| concept, we associated this verb with the canonical specification "11.1 Send, *transport*."

5.3. 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 LVD 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 LVD-class frequency for a particular HowNet concept. Out of the 17 verbs associated with HowNet's |Decorate| concept, only two of them (the two miscategorized Chinese verbs) are associated with an LVD class that is not 9.9 or 9.8. As in the gap-recovery described approaches above, our technique associates the miscategorized verbs with the canonical specification "9.8.b Fill, *decorate*."[13]

6. Results

Using the gap compensation techniques described above, we have achieved a more refined HowNet-to-LVD mapping, providing an increase in LVD-classified Chinese words from the previous 8089 entries to the current expanded set of 17284 LVD-classified Chinese words. This section presents a description of our coverage, comparing HowNet to LVD, and then provides a quantitative evaluation.

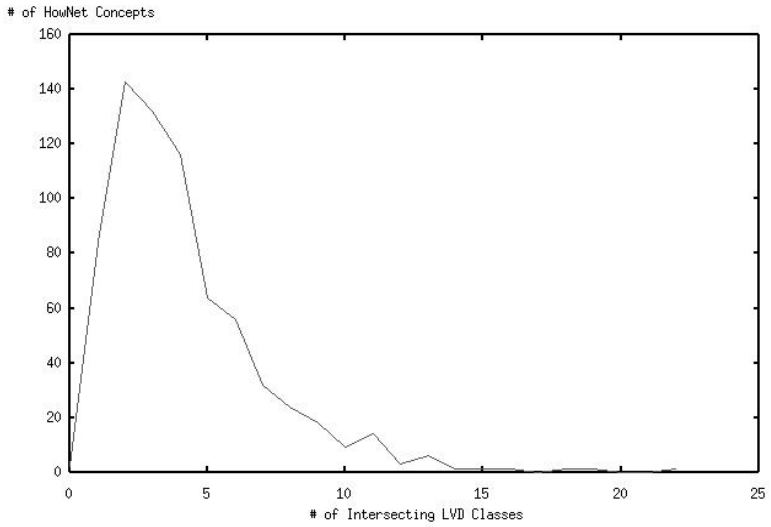
6.1. Coverage of HowNet/LVD Alignment

The histogram in Table V characterizes the number of LVD classes required for coverage of 709 HowNet concepts. We considered this initial experiment to be a success for several reasons: (1) In 359 cases (50% of the HowNet concepts), the partitioning corresponded to 3 or fewer LVD classes; (2) Most HowNet concepts with 2 or more partitions had a very heavy association with a single LVD class (60% or higher), with most other partitions falling around 20% or lower; (3) Only 2 cases did not correspond to any LVD class (i.e., degenerate HowNet concepts for which no correlations with LVD could be found); (4) There were virtually no partitionings (a handful of single HowNet concepts) exceeding 13 LVD classes.

At the time of this initial experiment, the HowNet resource did not include English translations. Although the translation resource we used was the CETA/Optilex dictionary, our technique was developed to accommodate any arbitrary translation resource for mapping between

[13] Ultimately, the miscategorized 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.

Table V. Distribution of HowNet Concepts by Number of Intersecting LVD Classes using Canonical Specifications



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

HowNet concepts and LVD classes. However, the most recent release of HowNet associates English glosses with each word in a class. To assess the impact of additional translation resources, we performed a three-way comparison, performing the HowNet-LVD mapping with: (1) Optilex translations alone; (2) HowNet translations alone; and (3) a merged resource including translations from both HowNet and Optilex. We then computed precision and recall measures for HowNet-LVD mapping for each of the individual resources relative to the merged resource and to each other. We also assess the impact of using canonical grid information to aid in the assignments. The results appear in Table VI below.

We find that the HowNet resource achieves higher precision, as might be expected since the available translations are limited to those the designer believed appropriate for each class. The Optilex resource achieves higher recall by drawing from a wider variety of alternate

Table VI. Precision and Recall for HowNet-LVD Mappings (with and without Canonical Grid Information)

Contrast	Precision		Recall	
	w/o Canon	w/ Canon	w/o Canon	w/ Canon
HowNet vs Optilex	0.61	0.65	0.46	0.55
Optilex vs HowNet	0.42	0.51	0.61	0.65
HowNet vs Merged	0.79	0.82	0.61	0.67
Optilex vs Merged	0.71	0.75	0.79	0.80

translations. The canonical grid information improves all measures and smooths differences between the resources.

If we further examine the translations used to make these assignments, we find 7653 Chinese-word/English-gloss pairs in common, 17609 pairs from HowNet, and 14252 from Optilex, from a total of 24205 assigning pairs. The results indicate that a merged translation resource, drawing from both HowNet and LVD/Optilex, can produce a richer and more robust mapping among the concept classes. For example, the HowNet concept |WeatherChange| is associated with three verbs, 下雨 (rain), 下雪 (snow), and 普降 (fall all over the area). Whereas the first two verbs have translation equivalents that link directly into our thematic-grids (and, hence, our WordNet senses), the third verb is WordNet linked solely by virtue of our thematic-grid matching routine. This routine allows us to determine that the closest English equivalent for 普降 is *precipitate*—a verb that does not show up in the HowNet hierarchy. Thus, our integration of LVD/Optilex with the HowNet resource has provided a more comprehensive linking to thematic grids and WordNet senses than would be available in either resource alone.

6.2. Quantitative Evaluation

In addition to the qualitative descriptions of coverage and comparisons of HowNet-LVD alignment described above, we implemented a quantitative analysis of the effectiveness of our semi-automatic lexicon creation process. We compare the automatic assignments of LVD class and theta grid to Chinese word with manual assignments by two Chinese language experts.

We compare the manually and automatically assigned LVD class and theta grid labels for a set of 272 separately hand-tagged Chinese

Table VII. Precision and Recall of our HowNet-LVD Alignment

Criterion	Precision	Recall
LVD Grid - Auto	0.62	0.25
LVD Grid - Random	0.10	0.0049
LVD Grid - Most Freq	0.357	0.07
LVD Class - Auto	0.63	0.29
LVD Class - Random	0.035	0.02
LVD Class - Most Freq	0.18	0.043

verbs. For these verbs, the semi-automatic linkage technique proposes 577 assignments of class or grid, some of which are duplicates. Manual assignment resulted in 1188 distinct theta grid labels and 1282 LVD class labels. We report precision and recall measures for these two types of labels relative to manual assignment. We also contrast the effectiveness of two plausible naive strategies as baselines: random assignment of LVD class and theta grid labels and assignment of most frequent label based on an English verb lexicon with more than 10000 entries.

Our criteria for agreement between manual and automatic assignments are as follows:

- LVD classes are said to agree if the major and minor LVD classes match exactly, e.g. 40.7.ii.a matches 40.7.i.

- Theta grids agree if the same roles appear in the same order, without regard with obligatory versus optional distinctions, e.g. `_ag,th` matches `_ag_th`.

These results appear in Table VII.

We achieve precision of approximately 0.62 for both LVD class and theta grid assignment; recall levels are lower, at approximately 0.24. As the table illustrates, the automatic technique we have developed substantially outperforms either random or most frequent LVD class assignment and random theta grid assignment. While still large, the contrast with most frequent thematic grid assignment is less dramatic. The relatively good performance of most frequent theta grid assignment

is easily explained by the fact that 28% of the verbs can appear as the basic transitive, the most common thematic grid. Thus, the transitive grid assignment produces precision of 0.357.

The relatively lower numbers for recall are best understood in terms of two features. First, this technique is more focused on high precision than on recall. Second, the "majority rules" strategy for selection among alternative likely assignments will tend to prefer more common class assignments, when less frequent acceptable class assignments are available.

7. Summary and Future Work

We have presented an approach to aligning two large-scale online resources, HowNet and LVD. 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 LVD. Our results indicate that the correspondence is very high between the 709 Chinese HowNet concepts and the 492 LVD classes. We see our techniques as the first step toward a general approach to building repositories for interlingual-based NLP applications.

Our work has shown that it is possible to combine different types of knowledge from existing resources in ways that improve upon the coverage and robustness of each of these independent resources. One area of investigation that has allowed us to enrich the existing resources is the development and application of gap compensation techniques, allowing us to fill in possible Chinese-English links where none existed previously.

We are currently using 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 LVD 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 LVD-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-LVD alignment. Because each of the entries in the LVD classification is associated with a WordNet sense, 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.

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, and to James Allan for help with the Inquery configuration.

References

Ballesteros, L. and W. B. Croft: 1997, 'Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval'. In: *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Dang, H. T., K. Kipper, M. Palmer, and J. 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, pp. 293–299.

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

Dong, Z.: 1988b, 'Knowledge Description: What, How and Who?'. In: *Proceedings of International Symposium on Electronic Dictionary.* Tokyo, Japan, p. 18.

Dong, Z.: 1988c, 'MT Research in China'. In: *Proceedings of International Conference on New Directions in Machine Translation.* Budapest, pp. 85–91. Also in New Directions in Machine Translation, 4 Distributed Language Translation

edited by Dan Maxwell, Klaus Schubert and Toon Witkam, Foris Publications, Dordrecht.

Dong, Z.: 2000, 'HowNet Chinese-English Conceptual Database'. Technical Report Online Software Database, Released at ACL. http://www.keenage.com.

Dorr, B. J.: 1993, *Machine Translation: A View from the Lexicon*. Cambridge, MA: The MIT Press.

Dorr, B. J.: 1994, 'Machine Translation Divergences: A Formal Description and Proposed Solution'. *Computational Linguistics* **20**(4), 597–633.

Dorr, B. J.: 1997a, '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, pp. 139–146.

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

Dorr, B. J.: 2001, 'LCS Verb Database'. Technical Report Online Software Database, University of Maryland, College Park, MD. http://www.umiacs.umd.edu/~bonnie/LCS_Database_Docmentation.html.

Dorr, B. J., N. Habash, and D. 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, pp. 333–343.

Dorr, B. J. and D. Jones: 1999, 'Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision'. In: E. Viegas (ed.): *Breadth and Depth of Semantic Lexicons*. Norwell, MA: Kluwer Academic Publishers, pp. 79–98. ftp://ftp.umiacs.umd.edu/pub/bonnie/Dorr-1999a.ps.

Dorr, B. J. and M. 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, pp. 438–447.

Dorr, B. J., G.-A. Levow, D. Lin, and S. Thomas: 2000, 'Chinese-English Semantic Resource Construction'. In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*. Athens, Greece, pp. 757–760.

Dorr, B. J., M. A. Martí, and I. Castellón: 1997, 'Spanish EuroWordNet and LCS-Based Interlingual MT'. In: *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCS-97-314*. San Diego, CA, pp. 19–32.

Dorr, B. J. and M. B. Olsen: 1996, 'Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization'. *Machine Translation* **11**(1–3), 37–74.

Fellbaum, C.: 1998, *WordNet: An Electronic Lexical Database*. MIT Press. http://www.cogsci.princeton.edu/~wn [2000, September 7].

Green, R., L. Pearl, B. J. Dorr, and P. Resnik: 2001a, 'Lexical Resource Integration across the Syntax-Semantics Interface'. In: *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*. Carnegie Mellon University, Pittsburg, PA, pp. 71–76.

Green, R., L. Pearl, B. J. Dorr, and P. Resnik: 2001b, 'Mapping WordNet Senses to a Lexical Database of Verbs'. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, pp. 244–251.

Habash, N.: 2000, 'oxyGen: A Language Independent Linearization Engine'. In: *Fourth Conference of the Association for Machine Translation in the Americas, AMTA-2000*. Cuernavaca, Mexico.

Hovy, E.: 1998, 'Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses'. In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain.

Hull, D. A. and G. Grefenstette: 1996, 'Experiments in Multilingual Information Retrieval'. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps.

Jackendoff, R.: 1983, *Semantics and Cognition*. Cambridge, MA: The MIT Press.

Jackendoff, R.: 1990, *Semantic Structures*. Cambridge, MA: The MIT Press.

Jones, D., R. Berwick, F. Cho, Z. Khan, K. Kohl, N. Nomura, A. Radhakrishnan, U. Sauerland, and B. Ulicny: 1994, 'Verb Classes and Alternations in Bangla, German, English, and Korean'. Technical report, Massachusetts Institute of Technology.

Langkilde, I. and K. Knight: 1998a, 'Generating Word Lattices from Abstract Meaning Representation'. Technical report, Information Science Institute, University of Southern California.

Langkilde, I. and K. Knight: 1998b, 'Generation that Exploits Corpus-Based Statistical Knowledge'. 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, pp. 704–710.

Langkilde, I. and K. Knight: 1998c, 'The Practical Value of N-Grams in Generation'. In: *International Natural Language Generation Workshop*.

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

Levow, G.-A., B. J. Dorr, and M. Katsova: 2000, 'Construction of Chinese-English Semantic Hierarchy for Cross-Language Retrieval'. In: *Proceedings of the Workshop on English-Chinese Cross Language Information Retrieval, International Conference on Chinese Language Computing*. Chicago, IL, pp. 187–194.

Miller, G. A. and C. Fellbaum: 1991, 'Semantic Networks of English'. *Lexical and Conceptual Semantics* pp. 197–229.

Nomura, N., D. A. Jones, and R. 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, pp. 243–249.

Oard, D. W.: 1998, 'A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval'. 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, pp. 472–483.

Oard, D. W. and B. J. Dorr: 1996, 'A Survey of Multilingual Text Retrieval'. Technical Report UMIACS TR 96-19, University of Maryland, Institute for Advanced Computer Studies. http://www.glue.umd.edu/~oard/research.html.

Olsen, M. B., B. J. Dorr, and S. 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, pp. 41–50.

Palmer, M. and J. 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, M., J. Rosenzweig, and S. Cotton: 2001, 'Automatic Predicate Argument Analysis of the Penn TreeBank'. In: *Human Language Technologies Conference*. San Diego, CA.

Palmer, M., J. Rosenzweig, and H. T. Dang: 1997, 'Intersective Levin Classes'. In: *Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington, D.C. Presentation at the Working Group on Combining Knowledge Sources for Automatic Semantic Tagging.

Palmer, M. and Z. Wu: 1995, 'Verb Semantics for English-Chinese Translation'. *Machine Translation* **10**(1–2), 59–92.

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

Resnik, P.: 1995, 'Using information content to evaluate semantic similarity in a taxonomy'. In: *Proceedings of IJCAI-95*. Montreal, Canada, pp. 448–453.

Saint-Dizier, P.: 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, pp. 62–70.

Weinberg, A., J. Garman, J. Martin, and P. Merlo: 1995, 'Principle-Based Parser for Foreign Language Training in German and Arabic'. In: J. K. Melissa Holland and M. Sams (eds.): *Intelligent Language Tutors: Theory Shaping Technology*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 23–44.