September 1995

Verb Semantics for English-Chinese Translation

Martha S. Palmer
University of Pennsylvania, mpalmer@cis.upenn.edu

Zhibiao Wu
University of Pennsylvania

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-95-22.

Verb Semantics for English-Chinese Translation

Abstract

A common practice in operational Machine Translation (MT) and Natural Language Processing (NLP) systems is to assume that a verb has a fixed number of senses and rely on a precompiled lexicon to achieve large coverage. This paper demonstrates that this assumption is too weak to cope with the similar problems of lexical divergences between languages and unexpected uses of words that give rise to cases outside of the precompiled lexicon coverage. We first examine the lexical divergences between English verbs and Chinese verbs. We then focus on a specic lexical selection problem - translating English *change-of-state* verbs into Chinese verb compounds. We show that an accurate translation depends not only on information about the participants, but also on contextual information. Therefore, selectional restrictions on verb arguments lack the necessary power for accurate lexical selection. Second, we examine verb representation theories and practices in MT systems and show that under the fixed sense assumption, the existing representation schemes are not adequate for handling these lexical divergences and extending existing verb senses to unexpected usages. We then propose a method of verb representation based on conceptual lattices which allows the similarities among different verbs in different languages to be quantitatively measured. A prototype system **UNICON** implements this theory and performs more accurate **MT** lexical selection for our chosen set of verbs. An additional lexical module for *UNICON* is also provided that handles sense extension.

Keywords

Verb semantics, lexical divergences, lexical organization

Comments

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-95-22.

Institute for Research in Cognitive Science

Verb Semantics for English-Chinese Translation

Martha Palmer
Zhibiano Wu

University of Pennsylvania
3401 Walnut Street, Suite 400C
Philadelphia, PA 19104-6228

September 1995

Site of the NSF Science and Technology Center for Research in Cognitive Science

IRCS Report 95-22

MT Journal, , 1–32 (1995)

Verb Semantics for English-Chinese Translation

Martha Palmer mpalmer@linc.cis.upenn.edu
Department of Computer and Information Science, University of Pennsylvania, PA 19104

Zhibiao Wu wzb@unagi.cis.upenn.edu
Linguistic Data Consortium, University of Pennsylvania, PA 19104

Editor:

Abstract. A common practice in operational Machine Translation (MT) and Natural Language Processing (NLP) systems is to assume that a verb has a fixed number of senses and rely on a pre-compiled lexicon to achieve large coverage. This paper demonstrates that this assumption is too weak to cope with the similar problems of lexical divergences between languages and unexpected uses of words that give rise to cases outside of the pre-compiled lexicon coverage. We first examine the lexical divergences between English verbs and Chinese verbs. We then focus on a specific lexical selection problem – translating English *change-of-state* verbs into Chinese verb compounds. We show that an accurate translation depends not only on information about the participants, but also on contextual information. Therefore, selectional restrictions on verb arguments lack the necessary power for accurate lexical selection. Second, we examine verb representation theories and practices in MT systems and show that under the fixed sense assumption, the existing representation schemes are not adequate for handling these lexical divergences and extending existing verb senses to unexpected usages. We then propose a method of verb representation based on conceptual lattices which allows the similarities among different verbs in different languages to be quantitatively measured. A prototype system UNICON implements this theory and performs more accurate MT lexical selection for our chosen set of verbs. An additional lexical module for UNICON is also provided that handles sense extension.

Keywords: Verb semantics, lexical divergences, lexical organization

1. Introduction

One of the primary tasks in Machine Translation, MT, is the lexical selection of verbs. A lexical item in the source language must first be associated with a distinct verb sense in that language. Then a corresponding verb sense in the target language that most nearly reflects the same sense must be chosen (sometimes via an interlingua representation). Finally, the corresponding lexical item for the sense in the target language is used in the generation of a sentence which includes the appropriate translations of the verb arguments. This process is illustrated in Figure 1.

A common practice in operational MT and natural language processing, NLP, systems is to assume that a verb has a fixed number of senses and rely on a pre-compiled lexicon to achieve coverage of these senses. For example, in a transfer-based MT system, the verb senses in the source language can be defined by the space of candidate target verbs. The translation of the source verb is limited by the coverage of this pre-compiled dictionary, and usually no other mechanism is

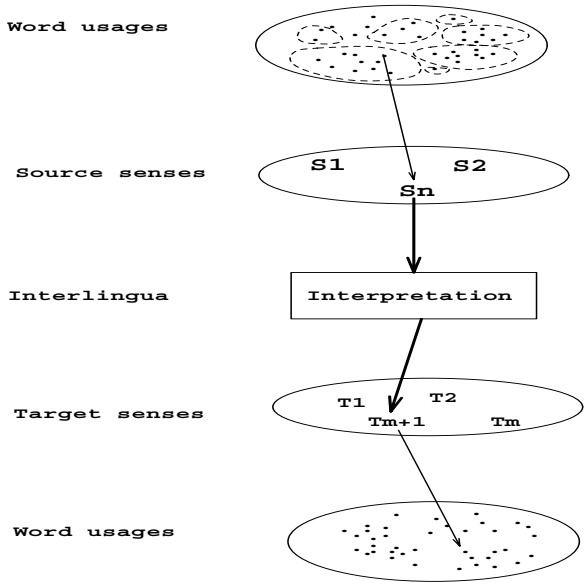

Figure 1. The relation between senses and lexical selection

provided for handling cases that fall outside of the coverage of the dictionary. This solution might be appropriate when an MT system is aimed at a sub-language where the text ranges over very restricted subject matter and is written in a formal, technical style. However, when an MT system is aimed at broader coverage and is used to process large corpora, it is unlikely that the exhaustive listing of verb senses is a realistic goal. The difficulties in obtaining complete coverage that are faced by single language NLP systems are compounded several fold by the task of machine translation. Zipf's law, [2], states that, however large the corpus is, there are always low frequency phenomena outside the corpus coverage. This characteristic of language makes it unlikely that all known senses will ever be identified, much less accounted for.

In MT this law applies to not just one language but to at least two. Each step in the translation process represents an opportunity where gaps in coverage are problematic. Not only can each lexicon be expected to have incomplete coverage, but when the lexicons are mapped together, there is likely to be little overlap between the gaps on each side. In addition, there will always be mismatches, where one language does not capture exactly the same linguistic distinctions as the other. Even if a bilingual lexicon could somehow be built with almost complete coverage for both languages, and with accurate mappings between them, it will still be a static database, and as such is seriously limited in its ability to deal with unexpected usages. One of the inherent properties of a natural language is its flexibility, i.e., the ability of any given sense to be extended to a new usage. The necessity of building

more dynamic lexicons for NLP systems that can cope robustly with the phenomena of unexpected usages is a well-established goal in the NLP community [18],[27], [20]. Less recognition has been paid to the even greater difficulties faced by MT systems. An MT system must first recognize an unexpected usage in the source language, and then must hypothesize an appropriate translation in the target language - an even more daunting task. By unexpected usage, we do not necessarily mean a figurative or metaphorical interpretation, but also an extension of meaning to a broader class of arguments, as in the extension of *break* from *broken wire*, meaning separated into pieces, into *broken insulation*, meaning a separation of the surface, [18] or from *break the fence* to *break the language barrier* [23]. We will illustrate the difficulty of this task with examples involving the translation of English *break* to Chinese.

We propose that the representation of each sense of an individual lexical item must include the ways in which it is related to other similar senses - which semantic concepts are shared, and which are not. In contrast with most interlingua approaches, which try to reduce a verb representation to a single primitive concept, we include several distinct semantic concepts in the representation of a single sense as well as their inter-relations. It is possible for these sets of concepts to overlap with the sets of concepts that represent other verbs. Where even partial overlaps exist, they constitute similarity links between the lexical items in question. We represent the "conceptual relatedness" of the lexical items as a lattice which is organized around hierarchical structures corresponding to the semantic concepts. This allows us to compute a quantitative measure for the similarity between two senses, based on proximity in a hierarchy. The lattice representation also allows us to move gracefully along the links from one sense representation to other closely related sense representations, enabling the system to explore extensions in meaning occasioned by unexpected verb usages.

In the following sections, we first explain lexical semantic divergences between English verbs and Chinese verbs and the not insignificant problem of translating between them, with *break* as our primary example. Then we review issues in the representation of verb semantics by examining two popular interlingua representations. Finally, our conceptual lattice approach is presented and a prototype system implementation, UNICON, is described. Experimental evidence is presented that demonstrates an improvement in the accuracy of lexical selection using this system along with an extension module designed to handle unexpected usages.

2. Lexical-semantic divergences

After close examination of appropriate translations of English *break* expressions into Chinese (Mandarin), we have determined that English and Chinese are quite far apart in their representation of *breaking* events, as in *John broke the window with a hammer*, [23]. There are several factors that contribute to this divergence. The most significant difference is that Chinese uses a compound Verb Adjective construction that makes both the action precipitating the *change-of-state* and the

details of the resulting state explicit. Although English also makes explicit the result, i.e., a *change-of-state* has taken place in which the object in question becomes *broken*, neither the specific action nor the fine-grained details of the resulting state are usually mentioned explicitly. It is, however, possible in English to refer to the details of the resulting state through the use of a prepositional phrase such as *into pieces* as in *John broke the window into many pieces.* A correlate of this structural difference is that Chinese then distinguishes lexically between both different actions and different types of resulting states, and has unique expressions for each possible combination. As a result, the lexical organization of 'break' in Chinese is quite different from the lexical organization of *break* in English. We will first examine the lexical organization of each language, and then discuss the problems in mapping from one to the other.

2.1. English *break*

We have already stated our commitment to using overlaps between semantic components of verbs to make explicit their conceptual relatedness. In later sections we will give examples of a preliminary conceptual lattice for capturing conceptual relatedness. We have based this work on a lexical organization of English verb classes proposed by Levin [15]. For example, *break* and *cut*, although both classed as *change-of-state* verbs, differ in that *cut* also indicates *directed motion* and *contact*. These differences are reflected in the different sub-categorization frames that can be associated with the two verbs. They can both take the middle construction, as in *Crystal vases break easily, This bread cuts easily*, which is normally associated with *change-of-state* verbs. But only *cut* can occur in the conative alternation *John cut at the bread, * John broke at the vase*. Levin's explanation for this is that the conative alternation assumes an underlying semantic component of *directed motion* and the absence of a normally expected semantic component of *contact*. Since *break* has no inherent *directed motion* or *contact* components, it cannot participate in this alternation. Levin groups several other verbs with *break*, and a different set with *cut*, by recognizing that they share these sub-categorization frames, presumably because they also share the same semantic components. However, for our purposes it is important to note that, in English, *break* is a pure change of state verb. In other words, the only semantic component associated with the set of verbs in the *break* verb class is *change-of-state*.

> "the break verbs, unlike the cut verbs, are pure verbs of change of state, and their meaning, unlike that of the cut verbs, provides no information about how the change of state came about." (Levin p. 242)

However, different senses of English *break* can be distinguished according to the type of *change-of-state* that is occurring. The *change-of-state* may be a change in a concrete object's integrity, such as a separation of the surface, or a separation into two or more pieces. Or the *change-of-state* may have to do with a change in continuity or a change in the functionality of the object, assuming it is a mechanical

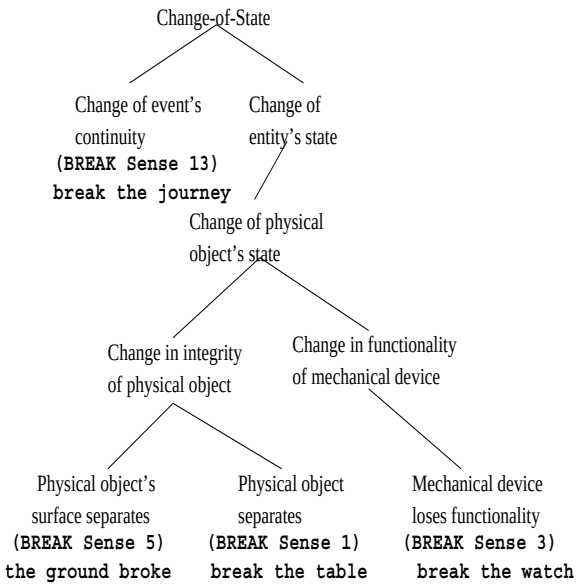

Figure 2. Break senses in change-of-state domain

device of some sort. In Figure 2 we give a conceptual hierarchy for the *change-of-state* domain that is relevant to the senses of *break* discussed here. This will be explained in more detail later.

2.2. Chinese 'break'

The same 'break' situations are described quite differently in Chinese, using *verb compounds* [1], [9]. Not only do these constructions behave very differently from a syntactic point of view, but they also make more specific both the action causing the *change-of-state*, and the resulting state of the object being changed. Recent studies at the University of Maryland indicate that these compounds may actually be serial verb constructions, where the order of the lexical items reflects the temporal ordering of the events [24].

Many Chinese dictionary entries are compound words consisting of several distinct lexical items. The meaning of the complete Chinese expression is usually composed from the meaning of the individual words. This is true of Chinese verb compounds of which there are three types, one Verb Verb (VV) compound, and two Verb Adjective (VA) compounds.

A VV compound, as illustrated below, expresses two distinct actions. In the following example, the VV compound *gan-pao* is composed of two single verbs *gan* and *pao*. The first verb *gan* takes the subject and the object as arguments while

the second verb *pao* takes only the object, and indicates an action that was caused by the action referred to by the first verb.

狗	赶跑	了	猫
Gou	gan-pao	le	mao.
dog	chase-run	Aspect marker	cat

The dog chased the cat and the cat ran away. (VV)

In a VA compound, the resulting state or event can be indicated by an adjective as well as a verb, and this is illustrated by the following two examples. In the first one, *chi-bao* is a VA compound composed of one verb *chi* which takes the subject as an argument, and one adjective *bao* which describes the resulting state of the *subject*.

张三	吃饱	了	饭
Zhangsan	chi-bao	le	fan.
Zhangsan	eat-full	Aspect marker	meal

Zhangsan has eaten his meal and is full. (VA)

In contrast, *da-sui* is a VA compound composed of one verb *da*, which takes the subject and the object as arguments, and one adjective *sui*, which modifies the *object*.

约翰	打碎	了	花瓶
Yuehan	da-sui	le	huaping.
John	hit-into-pieces	Aspect marker	vase

John broke the vase. (VA)

VA compounds are productive, although there are semantic constraints on their formation. A single Chinese verb and a single adjective can be combined to form a new VA compound as long as the resulting state described by the adjective is plausible. Because there are potentially so many combinations, a Chinese dictionary can hardly list them all. For example, native Chinese speakers will agree that the following examples all constitute natural Chinese expressions, although many of them, such as *ji-sui*, are not in the New Chinese Multi-purpose Dictionary [7].

击碎	ji-sui	hit-into-pieces
击破	ji-po	hit-into-irregularly-shaped-pieces
击开	ji-kai	hit-open
打断	da-duan	hit-into-line-segment-pieces
弄断	long-duan	do-something-resulting-in-line-segment-pieces
折断	zhe-duan	bend-into-line-shape
压断	ya-duan	press-into-line-shape

An important aspect of the use of VA compounds for expressing 'breaking' events is that the Adjectival component expresses the resulting state more specifically than is normally done with English. This can clearly be seen by examining the

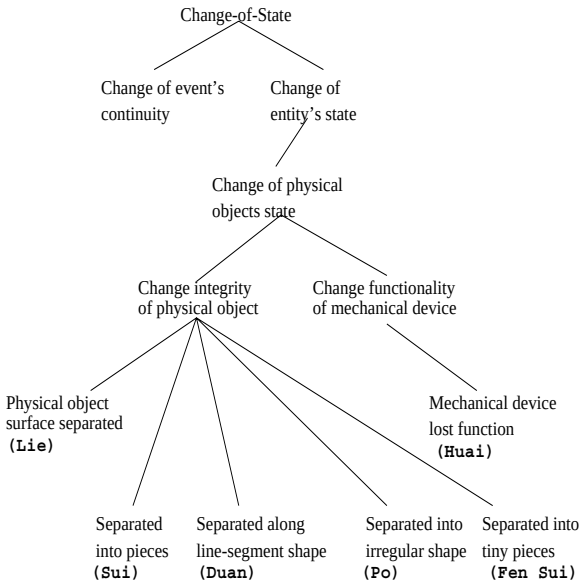

Figure 3. Chinese words in change-of-state domain

examples given above. Chinese makes some of the same distinctions that English makes, with respect to a *change-in-integrity* versus a *change-in-functionality*, but it makes additional distinctions based on the final state of the broken object. We have captured these sense distinctions in the *change-of-state* domain in the Chinese conceptual hierarchy in Figure 3.

Since the verb compounds are productive, it is tempting to assume the individual characters can be treated as stand-alone lexical items, and allowed to compose dynamically. But this is not a random process, and there are semantic constraints on which word can be composed with which other word. For example, the following constructions do not naturally occur in Chinese text, because something cannot be *chased red*, or *bent into pieces*.

* 赶红	gan-hong	chase-red
* 折碎	zhe-sui	bend-pieces

The importance of the VA compound for expressing *change-of-state* events such as *breaking* events in Chinese is brought out by the following experiment. Using the PH corpus (8M bytes), containing publications of the Xinhua News Agency of China during a period from January 1990 to March 1991, a statistical analysis was performed on the occurrences of four adjectives with related "concrete" objects. Over 80% of the constructions occurred as VA compounds, either with or without an explicit grammatical subject [25]. Less than 2% of the constructions occurred as

the A without the V, in an SAO construction, indicating how strongly the Adjective prefers to co-occur with a Verb.

2.3. Semantic Specificity

In addition to the inherent problem of associating single English verbs with Chinese compound verb constructions which have a very different syntactic structure, there is another fundamental difficulty in translating the English verb *break* into Chinese; the problem of semantic specificity. English *break* can be thought of as a general verb indicating an entire set of *breaking* events that can be distinguished by the resulting state of the object being broken. *Shatter, snap, split,* etc., are English verbs which can all be seen as more specialized versions of this general *breaking* event. Since Chinese has no equivalent verb for indicating the entire class of Chinese 'breaking' events, each usage of English *break* has to be mapped on to a more specialized lexical item. This is the equivalent of having to first interpret the English expression into a more semantically precise correlate. For example, *John broke the crystal vase*, and *John broke the stick* could be rewritten as *John shattered the crystal vase* and *John snapped the stick* before translation. Since in Chinese there are lexical matches for *snap* and *shatter*, namely *da-duan* and *da-sui*, this would simplify the translation process. The problem is that there are not always English lexical items corresponding to Chinese specializations of 'break.' In order to determine the most appropriate Chinese translation, the original English sentence must therefore be mapped onto a conceptual level that can then be realized with Chinese lexemes. From now on we will use 'break' to refer to this conceptual level for both English and Chinese.

In addition, as mentioned above, Chinese also makes specific the action involved. In English, if we say *John broke the window with a hammer*, or even *John shattered the window with a hammer*, there is an implicit assumption that what John actually did with the hammer involved hitting the window with it, rather than sliding the hammer against the window, or pressing the window with the hammer, or anything else. In Chinese, that action is made explicit. So, *John broke the window with a hammer* becomes

约翰	用	锤子	咂碎	了	窗子
Yuehan	yong	chuizi	zha-sui	le	chuangzi.
John	uses	hammer	hit-into-pieces	Aspect marker	window

whereas *John broke the window with the vise*, where the implicit assumption is that too much pressure was exerted through the vise, would become,

约翰	用	钳子	夹碎	了	窗子
Yuehan	yong	qianzi	ja-sui	le	chuangzi.
John	uses	vise	clamp-into-pieces	Aspect marker	window

To summarize, English has a single lexical item, *break* that corresponds to a wide range of 'breaking' events, each of which has a unique lexical expression in Chinese composed of at least two lexical items. The Chinese expression, in addition to adding details about the resulting state that are lacking in English, also includes information about the specific action that precipitated the *change-of-state* event.

In the rest of the paper we will look at computational approaches to handling the divergences we presented here. We will begin with selectional restrictions, and discuss their current inadequacies and the potential for improving on this. We will also discuss interlingua approaches, and whether or not they are advantageous. Finally, we will present our implementation of a conceptual lattice, and discuss plans for extending it.

3. The limitations of selectional restrictions

As we have just discussed, there are several inherent obstacles to a simple computational approach to the translation between English 'break' and Chinese 'break.'

- The syntactic structures are fundamentally different.
- Chinese has no lexical item that is representative of the general class of 'breaking' events.
- Chinese is more specific than English with respect to the resulting state.
- Chinese makes the precipitating action explicit and English does not.

The most widely used computational technique for distinguishing between verb senses, especially with transfer-based systems, is *selectional restrictions*; associating the type of each verb argument with membership in a particular class (or classes). In this section we will first discuss inherent strengths and weaknesses in the use of selectional restrictions for the lexical selection of 'break' verbs. We will go on to present an experiment that was performed with a well-known transfer-based system, TranStar. Finally, we will discuss possible enhancements to this system, and their potential for improving performance.

3.1. Selectional restrictions for choosing resulting states

The main factor in determining the correct resulting state in a 'break' event is the object that is undergoing the change-of-state. The most natural manner in which an object will 'break,' for instance, is for the most part determined by what type of object it is. Extremely fragile, brittle, objects such as crystal will break into many

pieces, or *shatter*. More solid concrete objects such as ceramic plates or bowls are less likely to *shatter*, but instead will probably break into a few irregularly shaped pieces. Slightly brittle objects that are originally shaped as line segments, such as wooden sticks, or cinnamon sticks, or candy canes, if they are 'broken', are likely to *snap* into several pieces that are also shaped like line segments. These distinctions can be captured at least partially by associating sets of selectional restrictions with the resulting states that specify the characteristics of objects that are likely to break up in certain ways. It must be acknowledged however, that this will never be completely reliable since a given context can always override normal expectations. An extreme amount of force being applied, (for instance by a steamroller), could *shatter* objects such as trees and bicycles that would normally not be considered brittle. Even in a simple sentence such as *John broke the stick into small pieces*, it must be noted that the prepositional phrase provides information that overrides the expectations normally associated with sticks, that they break up into line segments, and the more accurate Chinese translation would be *da-sui*, (hit-into-small-pieces), instead of the expected *da-duan*, (hit-into-line-segment shaped pieces).

3.2. Selectional restrictions for choosing actions

The importance of context and the limitations of selectional restrictions are highlighted even more in the task of attempting to specify the action involved.

As we have seen, for the sentence *John broke the vase*, a correct translation is *Yuehan da-sui le huaping*. Here 'break' is translated into a VA type verb compound. The action is specified clearly in the translation. An additional example illustrates how the translation can depend on an understanding of the surrounding context.

> *The earthquake shook the room violently, and the more fragile pieces did not hold up well. The dishes shattered, and the glass table was smashed into many pieces.*

The translation of the last clause, given below, includes the Chinese verb '震成' (zhenchen) in which the first character means *shake* and has been derived from the first clause of the English sentence:

那	玻璃	桌子	被	震成	了	碎片
na	boli	zhuozi	bei	zhenchen	le	suipian
That	glass	table	Pass.	shake-become	Asp.	pieces

The glass table was shaken until it broke into many pieces

This example illustrates that achieving correct lexical choice requires more than a simple matching of selectional restrictions. A fine-grained semantic representation of the interpretation of the entire sentence that can indicate the contextually implied action as well as the resulting state of the object involved is required. This cannot be provided by selectional restrictions alone, but is indicative of the need for a

knowledge-based understanding approach. The potential for current knowledge-based understanding approaches to handle lexical selection will be discussed later. In the next section we provide an illustration of the limits of an approach based solely on selectional restrictions and an exhaustive listing of verb senses.

3.3. Testing a transfer-based system

In our examination of the potential adequacy of selectional restrictions, we have just seen that, although they should prove fairly adequate for determining the result state, with some exceptions due to contextual overrides, they have little chance of accurately selecting actions. Our next step is to examine an actual implementation of a transfer-based system, to see whether or not it meets our expectations. In this section we present an experiment using the commercial English to Chinese machine translation system TranStar [3]. TranStar uses the verb argument structure for selecting the target verb. This requires that each translation verb pair and the selectional restrictions on the verb arguments be exhaustively listed in a bilingual dictionary. In this way, a verb sense is defined with a target verb and a set of selectional restrictions on its arguments.

In TranStar the English verb *break* can translate into 13 different Chinese expressions, distinguished by selectional restrictions. The selectional restrictions classify the events denoted by the English verb *break* into several sharply divided sub-categories. The relations among different sub-categories are not specified, as illustrated by the following examples:

English	Chinese	Meaning	Selectional restrictions
BREAK	打碎	to break into pieces	Object is brittle
BREAK	决裂	to break (the relation)	Object is a kind of connection
BREAK	打断	to break the continuity	Object is a continuous event
...	...	...	...

In the Brown corpus, we found 246 sentences containing *break, broke, breaking,* and *broken.* After removing most idiomatic usages and verb particle constructions, there were 157 sentences left which were used to test TranStar, with the results given in Table 1. The numbers in the table next to the Chinese characters for each entry are the frequencies with which the 157 sentences were translated into that particular Chinese expression. Most of the zero frequencies represent Chinese verbs that correspond to English *break* idiomatic usages or verb particle constructions which were removed. The accuracy rate of the translations is not high. Only 30 (19.1%) words were correctly translated, as agreed by our four native speakers. The Chinese verb '打碎' *da-sui* acts like a default translation when no other choice matches, but was not usually correct.

Table 1. TranStar break entries

Chinese	打碎 107	破坏 22	间歇 14
Pinyin	da-sui	po-hui	jian-xie
Meaning	to break into pieces	to make damage to	to have a break
Chinese	决裂 5	违反 2	爆发 0
Pinyin	jue-lie	wei-fan	bao-fa
Meaning	to break (a relation)	to against	to break out
Chinese	发生故障 0	闯入 0	打断 0
Pinyin	fa-shen-gu-zhang	chuan-lu	da-duan
Meaning	to break down	to break into	to break a continuity
Chinese	突破 0	得失相当 0	违背 0
Pinyin	tu-po	de-shi-xian-dan	wei-bei
Meaning	to break through	to break even with	to break (a promise)
Chinese	完成绝大部分 0		
Pinyin	wan-chen-jue-da-bu-fen		
Meaning	to break with		

3.4. Potential for performance improvement

The low accuracy rate in the previous section is not due to a fault in TranStar, but is rather an indication of the difficulty of providing accurate, broad-coverage, lexical selection. The same 157 sentences were translated by one of the authors into 68 Chinese verb expressions, many of which occurred only once or twice. These expressions can be listed according to the frequency with which they occurred, in decreasing order. The verb which has the highest rank is the verb which has the highest frequency. In this way, the frequency distribution of the two different translations can be shown in Figure 4.

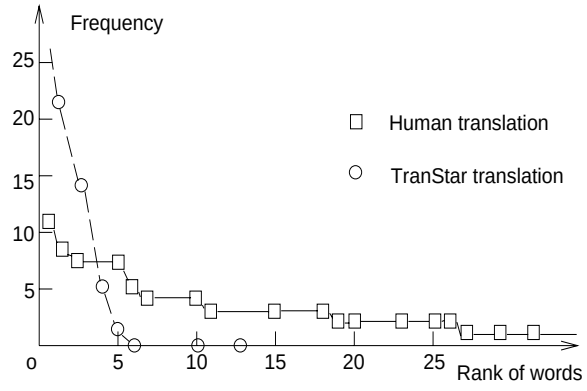

Figure 4. Frequency distribution of translations

Table 2. Human translation results

	Subject 1	Subject 2	subject 3
Total number of translations	148	139	145
Number of translations used by all three subjects	33	33	33
Number of translations used by two subjects	50	14	56
Number of translations used by only one subject	65	92	56

We performed an additional experiment in which we had three native speakers translate the original 246 sentences and compared their results. Each of the translators used an even greater number of different Chinese expressions since these sentences included the idiomatic usages and verb particle constructions. There was a great deal of diversity, and only 33 of the sentences were translated consistently. The results are summarized in Table 2.

The lexical selection task for translation obeys Zipf's law. That means that, for all possible verb usages, a large portion are translated into a few target verbs, while a small portion might be translated into many different target verbs. Clearly, native speakers do not restrict themselves to a fixed set of 13 verbs for lexical selection. Tripling TranStar's number of Chinese verb senses, i.e., to 39, and providing each sense with more detailed selectional restrictions, would still not provide coverage for much more than half of the possible translations. It should have substantially more impact on the accuracy rate, assuming all the high frequency expressions are included. However, given an additional 100 sentences, it is only too likely that many of them will fall outside the coverage of the system. A predetermined exhaustive listing of verb senses, no matter how extensive, cannot guarantee coverage of the phenomena. Human use of language is simply too diverse and too creative. The challenge for lexical semanticists is to contrive a method of verb representation that can model the fluid nature of verb meanings that allows human speakers to contrive and recognize novel usages in every sentence.

4. The limitations of interlingua for lexical selection

In the above sections, we have presented the inherent difficulties in lexical selection that cause problems for standard transfer-based MT systems which rely on selectional restrictions associated with fixed word senses. Interlingua approaches also have limitations when applied to this particular set of problems, which we will discuss here. We will then propose an alternative, and describe our implementation and testing.

The underlying motivation behind an interlingua approach rests on the assumption that a universal semantic representation can be found for a sentence and its translations into different languages. Many interlingua approaches choose a set of primitive concepts and then map everything onto this set [17], [5]. This has been especially effective for handling lexical divergences between languages, when the same concept has different types of syntactic realizations in different languages [5]. One of the main advantages claimed for this approach is that, once the interlingua has been defined, adding an additional language only requires linking the new language to the interlingua representations. The correct generation into the existing languages will follow automatically. There are certainly gains in efficiency of representation that stem from the use of interlingua, but we found on examination that they also had limitations with respect to the particular lexical selection task we had in mind.

In general, an interlingua is expected to be an artificial language consisting of a finite set of primitive concepts. Individual lexical items are considered to be *subconcepts* of the categories represented by the primitives, which are the *superconcepts*. Subconcepts inherit all of the properties associated with their superconcepts, and are considered to be more specialized versions of the superconcepts. They can be distinguished from other subconcepts of the same superconcept through selectional restrictions. This is illustrated by the following example from the Mikrokosmos system of the verb *eat*, which is represented as having two arguments, an AGENT and a THEME [17]:

```
SEM:
        (%ingest
          (AGENT (value ^$var1)
                 (sem *animal))
          (THEME (value ^$var2)
                 (sem *ingestible)
                 (relaxable-to *physical-object))))
```

In this representation, *eat* is mapped onto a superconcept INGEST and two selectional restrictions, ANIMAL and INGESTIBLE are imposed on the verb arguments. In this way the conceptual similarities between verbs such as *eat* and *drink* can be captured, since they both map onto INGEST, with the selectional restrictions being used to help distinguish classes of arguments, i.e., LIQUID vs. SOLID INGESTIBLES. The target verb which shares the same mappings to the superconcept is selected during translation.

The success of an interlingua is dependent on the possibility of being able to map all of the semantic distinctions made by individual languages onto the same set of primitive concepts. When one language makes distinctions another language does not make, that were not previously in the interlingua primitives, the primitives must be augmented to allow for the new distinctions. We can illustrate this with our 'break' example. While the superconcept is certainly an important piece of

information, knowing that 'break,' has a superconcept of *change-of-state* is insufficient in selecting Chinese translations that require even more specificity than is found in English. We can see what would be needed more clearly by turning to another system.

An additional significant interlingua system is Bonnie Dorr's UNITRAN system [5] which makes a commitment to the use of Jackendoff's Lexical Conceptual Structures (LCS), [11], as an interlingua representation. The clearly defined mapping rules between the LCS and the different target languages allows UNITRAN to elegantly handle a large variety of both syntactic and semantic divergences between languages. However, similarly to Mikrokosmos, it has not been aimed at capturing the fine granularity of meaning required by the particular types of lexical selection problems we are discussing here. Again, the necessity of decomposing verbs into a pre-defined set of primitives imposes a limitation on the possible range of representation. Since LCS is mainly concerned with syntactic-semantic correspondences, i.e., syntactic realizations, it does not attempt to decompose semantic components such as MANNER and RESULT-STATES. These may not be sensitive to syntactic variation in an individual language such as English, but they are important for resolving semantic divergences in order to achieve accurate lexical selection. In particular, many distinct lexical items have identical conceptual representations, and are distinguished only by inserting the actual lexical item into a MANNER field. For example, the verb *jog* is defined as:

```
(DEF-ROOT-WORDS (GO-LOC Y (FROM-LOC
                  (AT-LOC Y Z1))
                 (TO-LOC (AT-LOC Y Z2)))
        :ROOTS ((JOG (Y (* Y))
                  (Z1 :OPTIONAL ((* FROM-LOC)
                 (AT-LOC (Y) (Z1))))
                  (Z2 (UC (CASE ACC)) ((* TO-LOC)
                  (AT-LOC (Y) (Z2))))
                  (MODIFIER JOGGINGLY))
```

Jog decomposes into several primitives such as GO-LOC, FROM-LOC, AT-LOC, TO-LOC and a MODIFIER JOGGINGLY. This representation scheme captures important parts of the meaning of the verb *jog*. In particular it provides the necessary information for mapping from grammatical roles to the thematic relations, and preserving syntactic-semantic correspondences. However, it attempts to cover a large part of the conceptual meaning through the use of the MODIFIER JOGGINGLY. When similar verbs such as *run*, *walk* and *sneak* are defined, their representations are the same, with different modifiers in the MANNER field, i.e., RUNNINGLY, WALKINGLY, SNEAKINGLY. There is no place in the representation for capturing fine-tuned conceptual differences between these verbs.

The same thing occurs with RESULT-STATES. For example, in the following representations of the English verbs *break* and *die* in UNITRAN, the same seman-

tic primitives, GO-IDENT, TOWARD-IDENT and AT-IDENT, are used for both verbs. The distinctions between the participants of these two different events can be captured in the representation by specifying different selectional restrictions on the arguments. For the *die* event, the participant should be ANIMATE +, while for the *break* event, the participant should be ANIMATE -.

DIE
(DEF-ROOT-WORDS (GO-IDENT Y (TOWARD-IDENT (AT- IDENT Y Z)))
(DIE (Y (UC (ANIMATE +)) (* Y)) (Z DEAD))

BREAK
(DEF-ROOT-WORDS (GO-IDENT Y (TOWARD-IDENT (AT- IDENT Y Z)))
(BREAK (Y (* Y (UC (ANIMATE -)))) (Z BROKEN)))

The differences in the resulting states are reflected as DEAD and BROKEN, which are defined as ROOT-WORDs in the interlingua. This may be sufficient for distinguishing between *die* and *break*, but it is inadequate for capturing the fine-grained semantic distinctions we require for Chinese. It would be necessary, when Chinese verbs are defined based on this interlingua, for the interlingua ROOT-WORDs to include something like SEPARATE-INTO-PIECES, SEPARATE-INTO-LINE-SEGMENTS, and SEPARATE-INTO-IRREGULARLY-SHAPED-PIECES, Then, when *da-sui* is defined with SEPARATE-INTO-PIECES, an explicit connection would have to be made associating BROKEN with SEPARATE-INTO-PIECES. This would require adding an extensive set of ROOT-WORDs, as well as the connections between them, to whatever multilingual ontology is already in place.

In summary, existing interlingua representations cannot handle the semantic divergences we have discussed in the above section without augmentation. The general approach of substituting primitive concepts for lexical items does not provide the enrichment of semantic distinctions that is critical to our lexical choice issues. In the next section we propose an alternative approach that could be seen as a potential augmentation for either one of these systems, or a transfer-based system.

5. Augmenting MT systems with conceptual lattices

In the preceding sections we have discussed two opposing trends in MT verb representation, transfer-based systems and interlingua based systems. One could be characterized as the dreaded "replacement" of lexical items with decompositions, as exemplified by the interlingua approaches. The other could be characterized as the equally dreaded reduction of semantics to basically (syntactic) argument structure with selectional restrictions, as practiced in many transfer systems. In this section we propose an alternative, which relies equally heavily on the selectional restrictions so popular with transfer-based systems and the conceptual primitives so popular with interlingua. However, in our system the conceptual primitives are not seen as replacements for lexical items, but as indicators of class membership, and as point-

ers to conceptually related classes. These conceptually related classes comprise the domains that are organized by our hierarchies, and are used to perform best partial matches for more accurate lexical selection.

5.1. Defining conceptual domains

We see semantic components as an enhancement of the verb representation, rather than comprising the whole of the representation, in agreement with Levin, who stated:

> Numerous arguments have been advanced against the use of predicate decomposition, as in Fodor et al.'s paper "Against Definitions" (1980). Many of their arguments are inapplicable to the discussion of decomposition here. They assume that the decompositions are put to use other than that assumed here. In the works discussed, the decomposition of verbs is proposed for the purposes of accounting for systematic semantic-syntactic correspondences. ... instead, Fodor et al.'s concern is whether the decomposition or definition actually replaces a lexical item whenever it is used. They are not interested in the independent question of whether a decomposition analysis as a lexical semantic representation enters into the statement of linguistic generalizations. [14] p. 39.

In the approach we describe here, we are concerned with making use of linguistic generalizations based on conceptual decompositions that augment, rather than replace, our lexical items. We also rely heavily on the syntactic-semantic correspondences to be found in argument structures and their associated selectional restrictions. Computational linguists have continually sought to simplify lexical semantic representations for more compact system implementations. In contrast, the proposal here is in favor of enriching semantic representations, rather than compressing them.

We view a verb meaning as a lexicalized concept which is undecomposable. However, this semantic form can be projected onto a set of concepts in different conceptual domains. Langacker [13] presents a set of basic domains used for defining nouns. It is possible to define an entity such as a *knife* by using the *size, shape, color, weight, functionality* etc. Pustejovsky's qualia structure for defining the different components of a noun's meaning has a similar motivation [20]. We think it is also possible to identify a compatible set of conceptual domains for characterizing events and thus representing verb senses. Initially we are relying on the semantic components suggested by Levin as relevant to syntactic alternations, such as *motion*, *force*, *contact*, *change-of-state* and *action,* etc, [15]. We see these verb classes as closely related to the sets of verbs that share predicate representations in an LCS. For example, verbs defined with GO-IDENT and GO-LOC can be viewed as constituting separate verb classes, both of which are contained in a more general *change-of-state* class. In the work presented here we have made a preliminary attempt to use semantic components relevant to verb classes as conceptual domains

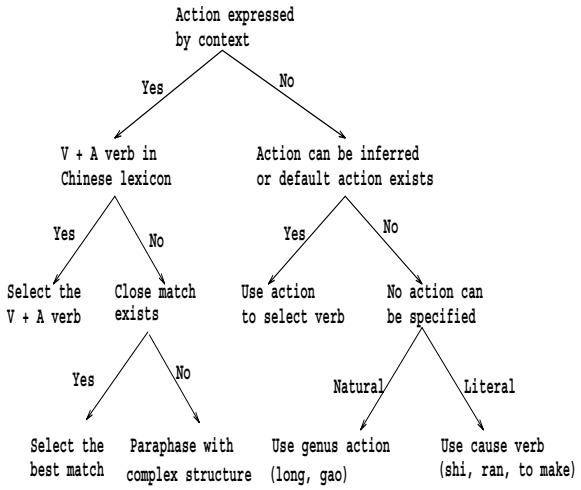

Figure 5. Decision tree for translation

that a verb's meaning can be projected onto. By specifying the inter-relations between the domains, our sense definitions become much less rigid. We can turn to close conceptual neighbors to try and achieve better matches if our first attempts at matching are disappointing. This allows us to respond flexibly to the mismatches occasioned by lexical divergences as well as unexpected usages.

5.2. The lack of suitable contextual information

However, for any existing approach, whether it treats conceptual primitives as definitions or merely indicators of class membership, an explicit representation of the context is required for the selection of *action* lexical items. For anything besides the most limited subdomain, this level of contextual representation is beyond the state of the art. A modern working system must assume that there will be many instances when the context will not be available, and in those instances an algorithm for selecting a default action verb is required. We propose the decision tree in Figure 5 as such an algorithm for choosing a general purpose action verb for the translation of English *change-of-state* verbs into Chinese. This algorithm would be suitable for implementation in any of the systems we have discussed above. The focus of the rest of our paper is on lexical selection of resulting states.

5.3. The relations among verb senses

In the implementation presented here we have merged our English conceptual lattice from Figure 2 and our Chinese conceptual lattice from Figure 3 into a single

interlingua lattice, (see Figure 8), to simplify the matching process. We will first describe a relatively straightforward example, and then explain how the lattices can also be used to hypothesize extensions to verb senses. By this we mean determining an implicit relation between a lexical item and an existing sense definition which was previously outside of the candidate set of verb senses for that lexical item.

The basis for our conceptual lattice for English 'break' comes from Meaning Text Theory, where verbs are assumed to have a core verb sense or basic sense [19]. The **Longman Dictionary of Contemporary English** [16] lists this core verb sense as the first entry,

> *To (cause to) separate into parts suddenly or violently, but not by cutting or tearing: to break a window/a leg. The rope broke when they were climbing. The window broke into pieces.*

and then goes on to list 17 additional related senses. Our analysis views semantically related senses as being either more specific, more general, or analogical to the core sense or other senses. In other words, the senses can be structured together into a lattice as superconcepts, subconcepts and analogies. We have built an IS-A hierarchy under a superconcept of *change-of-state* that relates Longman's 18 verb senses. We displayed a portion of that hierarchy for a few of the most common usages in Figure 2. For a detailed analysis of these 18 *break* senses and their inter-relations see [25]. In the hierarchy presented in this paper, a specialization of sense 1 would be *break off* as in a *branch broke off of the tree*, where there is a separation into pieces but the integrity of the original object is still preserved. Sense 3 is analogical to sense 1 and both of them share the superconcept *change-of-physical-object's-state*. This example illustrates the inter-relations among different senses of the same verb. For the most part, these inter-relations have not been used in existing NLP systems, but we will show the crucial role they play in accurate lexical selection.

We are not claiming that our lattices capture the complete meaning representation of any single lexical item, but rather that the semantic features and conceptual relations that are represented in the lattices form some portion of the verb's meaning that allows useful generalizations to be made.

5.4. Defining meaning similarity

If lexical items can be associated with concepts in an hierarchical structure, it is possible to measure the meaning similarity between words with an information measure based on WordNet [21], or structure level information based on a thesaurus [12]. The reason that the lexical organization is a lattice rather than a hierarchy (as in Mikrokosmos) is that many verb meanings include more than one semantic component. For example, *break* identifies a *change-of-state* event with an optional *causation* conception, while *hit* identifies a complex event involving *motion*, *force* and *contact* domains. Chinese verb compounds with VA constructions always identify complex events which involve *action* and *change-of-state* components. The

separate trees for each semantic component are grouped together into a lattice. Within one conceptual domain, the similarity of two concepts is defined by how far apart they are in the hierarchy for that domain, i.e., their structural relation.

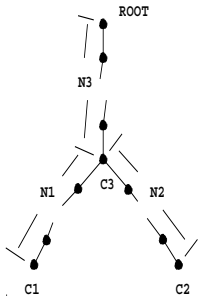

Figure 6. The conceptual relations

The conceptual similarity between C1 and C2 is:

$$ConSim(C1, C2) = \frac{2*N3}{N1+N2+2*N3}$$

C3 is the least common superconcept of C1 and C2. N1 is the number of nodes on the path from C1 to C3. N2 is the number of nodes on the path from C2 to C3. N3 is the number of nodes on the path from C3 to root.

For example, suppose PHYSICAL-OBJECT, WINDOW and KEYBOARD have the structure relation shown in Figure 7, the conceptual similarity between WINDOW and KEYBOARD is $(2 * 6)/(5 + 8 + 2 * 6) = 12/25$.

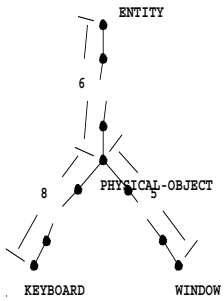

Figure 7. An example of the conceptual relations

After defining the similarity measure in one domain, the similarity between two verb meanings, e. g, a target verb and a source verb, can be defined as a summation of weighted similarities between pairs of simpler concepts in each of the domains the two verbs are associated with.

$$WordSim(V_1, V_2) = \sum_i W_i * ConSim(C_{i,1}, C_{i,2})$$

By making use of a hierarchy for selectional restrictions in the knowledge base, we can also measure the degree of satisfaction for selectional restrictions associated with verb arguments. Suppose the constraint set is:

(IsA con1 @var1)
(IsA con2 @var2)
... ...

We can measure the degree of satisfaction for each of the IsA constraints with the following function:

$$IsA(con1, var1) = ConSim(con1, C1) \tag{1}$$

For example, suppose we have a selectional restriction: (IsA BRITTLE-OBJECT var1) and BRITTLE-OBJECT is the immediate super node of WINDOW. When the variable var1 is set to WINDOW, the value of the IsA function is $(2 * 10)/(0 + 1 + 2 * 10) = 20/21$. If the variable is set to KEYBOARD, the value of the IsA function is $(2 * 6)/(8 + 4 + 2 * 6) = 1/2$.

The following equation measures the complete degree of satisfaction for all of the selectional restrictions of a single argument. N is the number of IsA functions being summed.

$$SatisDegree(ARG, CON) = \frac{\sum_i IsA(con_i, var_i)}{N} \tag{2}$$

In a given argument structure some of the arguments will be mandatory, and some will be optional. If a mandatory argument is missing, we assign -100 as the degree of satisfaction for that argument. If an optional argument is missing, it has no effect on the final degree of satisfaction.

5.5. Defining verb domains

In each conceptual domain, lexicalized concepts can be organized in an hierarchical structure. The conceptual domains for English and Chinese are merged by hand to form interlingua conceptual domains used for similarity measures. When the merge is being done, it is critical that similar concepts are put close together in the network. Figure 8 illustrates a portion of the *change-of-state* domain containing English and Chinese lexicalized concepts. Lexical items, either Chinese or English, are associated with their corresponding conceptual nodes. Some nodes have no lexical items. Some have either Chinese or English, but not both. If the source lexical item is associated with a node that has a target item as well, then this is equivalent to corresponding entries in a bilingual lexicon. Assuming the selectional restrictions are satisfied, the target lexical item will be selected as the translation. If the source lexical item is associated with a conceptual node that has no target

lexical item, then the search must begin for the best partial match, since a total match is impossible.

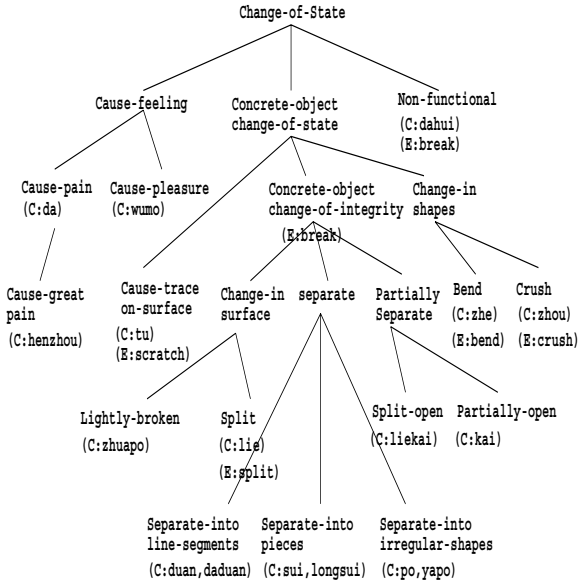

Figure 8. Change-of-state domain for English and Chinese

In addition to the conceptual domains, the representations of the lexical items include the argument structure and the selectional restrictions on each argument.

6. UNICON: An implementation

We have implemented a prototype lexical selection package UNICON where the representations of both the English and Chinese verbs are based on a set of shared semantic domains. This section describes an example in detail. The input to the system is a verb argument structure from a parsed sentence in the source language. Our example is *the man broke the window*, resulting in the following argument structure: (break man-0 window-0). Since this argument structure could conceivably correspond to more than one sense for that lexical item in the source language, the first step is sense disambiguation for the source language.

In our dictionary, English 'break' has seven different senses [19],(out of Longman's 18). Each sense can be illustrated with a sample sentence, as given below:

separated Some physical object is separated.
BREAK-I-1A The branch broke.
BREAK-I-1B Hail stones broke the roof.

BREAK-I-1C John broke the table with a hammer.
BREAK-I-1D The rocket broke into two parts.

discontinue Some continuous event becomes discontinuous.
BREAK-I-2 He broke the song with a solo.

non-functional Some devices lose their functionality.
BREAK-II-1A His watch broke.
BREAK-II-1B The fall broke the watch.
BREAK-II-1C He broke the paper drum.

A predictable set of selectional restrictions, marked with %, is associated with the arguments for each sense, indicated by @VAR1, @VAR2 and @VAR3. Each %SELECTIONAL RESTRICTION corresponds to a node in a conceptual hierarchy for nominals in the knowledge base, the *nominal hierarchy*. Each noun in the lexicon is given a link to the hierarchy. Our 7 English 'break' entries have the following selectional restrictions:

BREAK-I-1A ((UNKNOWN-P @VAR2) (%IS-A %PHYSICAL @VAR1))

BREAK-I-1B ((%IS-A %NATURE-FORCE @VAR1)
(%IS-A %PHYSICAL @VAR2))

BREAK-I-1C ((%IS-A %ANIMATE @VAR1) (OR (%IS-A %PHYSICAL @VAR3)
(%PART-OF @VAR3 @VAR1)) (%IS-A %PHYSICAL @VAR2))

BREAK-I-1D ((%IS-A %SEPARATE-STATE @VAR2) (%IS-A %PHYSICAL
@VAR1))

BREAK-I-2 ((%IS-A %ANIMATE @VAR1) (%IS-A %EVENT @VAR3)
(%IS-A %CONTINUOUS-EVENT @VAR2))

BREAK-II-1A ((UNKNOWN-P @VAR2) (%IS-A %FUNCTIONAL-DEVICE @VAR1))

BREAK-II-1B ((%IS-A %NATURE-FORCE @VAR1) (%IS-A %MECHANICAL-
DEVICE @VAR2))

BREAK-II-1C ((%IS-A %ANIMATE @VAR1) (OR (%IS-A %PHYSICAL @VAR3)
(%PART-OF @VAR3 @VAR1)) (%IS-A %MECHANICAL-DEVICE @VAR2))

The sense disambiguation process uses the selectional restrictions and the *Satis-Degree* equation. Because the nouns human-0 and window-0 are defined in the same hierarchy as selectional restrictions like PHYSICAL, MECHANICAL-DEVICE, etc., the similarities among these entities can be measured. The measure for degree of satisfaction for each candidate verb sense, such as BREAK-I-1A, is given below:

Sense	I-1A	I-1B	I-1C	I-1D	I-2	II-1A	II-1B	II-1C
SatisDegree	-797/16	-11/63	13/28	-47/176	-7/36	-101/2	-8/9	1/12

The lexeme with the highest measure, 13/28, is BREAK-I-1C, so this is chosen as the source verb sense, and the argument variables are instantiated with the verb arguments from the sentence. The representation is: (*change-of-integrity* window-0).

The system then tries to find the target verb realization that most closely matches the source verb sense. If the concepts in the representation do not have target verb realizations, the system examines nearby concepts as candidates to see whether they have target verb realizations. If a possible target verb is found, the selectional restrictions for the target verb arguments are tested against the corresponding source verb argument fillers. This is not expected to be an exact match, but two measurements are used to find the best inexact match. They are the Conceptual Similarity of the source verb and the target verb, and the degree of satisfaction of the selectional restrictions on the verb arguments. Our analysis gives conceptual similarity priority over the selectional restrictions on the arguments. Since there is no Chinese lexical realization for the single concept *change-of-integrity*, the system examines the concepts closest to *change-of-integrity* in the interlingua conceptual hierarchy, given below:

SEPARATE-INTO-PIECES-STATE
 SEPARATE-INTO-NEEDLE-LIKE-STATE
 SEPARATE-INTO-LINESEGMENTS-STATE
 SEPARATE-INTO-IRREGULAR-PIECES-STATE
 SEPARATE-INTO-SHANG-STATE
 SEPARATE-INTO-TINY-PIECES-STATE

For concepts SEPARATE-INTO-LINESEGMENTS-STATE and SEPARATE-INTO-PIECES-STATE, some of the Chinese realizations are:

- 断了 duan le (to separate into line-segment shapes).
- 打断 da-duan (to hit and separate the object into line-segment shapes).
- 碎了sui le (to separate into pieces).
- 打碎 da-sui (to hit and separate the object into pieces).
- 摔碎 suai sui (to throw the object, so it separates into pieces).

In order to compute the degree of satisfaction for the selectional restrictions, the source verb arguments must be associated with the potential argument fillers from the target verb realization. Then the selectional restrictions and the *SatisDegree* equation are used exactly as in the above example. In addition, the WordSim equation is used to measure the distance between the source verb concept and each of the

candidate target verb concepts. These measures are listed under "Conceptual Similarity" below along with the "SatisDegree" measures for the selectional restrictions.

	断了	打断	碎了	打碎	摔碎
	duan le	da-duan	sui-le	da-sui	suai-sui
Conceptual Similarity	1/6	5/7	0	5/7	23/56
SatisDegree	3/16	13/42	-50	9/14	9/24

The Chinese verb *da-sui* has the highest combined score, 5/7 and 9/14, and is chosen as the target lexical item. Although *da-duan* and *da-sui* have the same conceptual similarity measure, 5/7, the constraint satisfaction degree of *da-sui* is higher than *da-duan*. This is because the argument *window* met the selectional restrictions in *da-sui*, which specify that the object must be BRITTLE. The difference in scores between *da-sui* and *suai-sui* is that, even though they have the same result state, *sui*, they have different actions. Since the actions also select for the object, they have their own selectional restrictions, which are included in the equation.

The measurement of varying degrees of satisfaction is similar in spirit to the well-known tradition of using weights to choose between competing semantic analyses, first labeled as preference semantics by Yorick Wilks [22], and later implemented in several natural language systems, a recent, notably successful implementation being Grishman [8]. However, our work differs from theirs in emphasizing the conceptual relatedness of verb semantic representations required for machine translation.

We extended the coverage of the system to several verbs from the *hit, touch, cut* and *break* verb classes, and used this method to translate sentences from the Brown corpus. Before describing our experimental results, we will first describe an extension of this technique that allows the system to handle previously undefined senses.

7. Extending existing verb senses

We have implemented an extra module for handling unexpected verb usages which is activated when an input sentence cannot be classified according to the existing candidate verb sense categories. In other words, when the constraint satisfaction degree for each candidate sense is less than zero. The module has a different treatment for each of the three methods by which a sense might be extended. These three methods involve the same possible relations, subconcept, superconcept, and analogy that are used to define a conceptual hierarchy. The system does not create entirely new sense definitions, but finds means of associating lexical items with already existing sense definitions that are closely related conceptually, but which had not previously been associated with that particular lexical item. The means of association must be found by examining already existing conceptual links. As such, our process bears certain similarities to the process of recognizing metaphorical allusions [6]. We describe here the methods by which this module hypothesizes an

extension of a verb sense which has either a superconcept relation or an analogical relation to the candidate verb senses.

- Subconcept/Superconcept relation - A verb sense extension can be a sub-concept of a candidate verb sense. This means that the meaning of the candidate verb sense can be specialized in at least two or more ways. For example, the core sense of English *break* can be specialized into several different senses, such as *shatter, snap, etc.* which then correspond to different Chinese serial verb compounds such as SEPARATE-INTO-SMALL-PIECES, SEPARATE-INTO-LINESEGMENTS.

- Analogical relation - A verb sense extension can be an analogy of the candidate verb sense. For example, for the sentence *The car drinks gasoline*, there are analogies between *car* and *human*, and *edible liquid* and *gasoline* that need to be identified. This is the equivalent of coercing *car* to *human* and *gasoline* to *edible liquid (for cars)* so that the selectional restrictions on *drink* can be satisfied. (See [10] on coercion.)

The set of possible inter-relations between an extended verb sense and the existing candidate verb senses are crucial for prediction. When a human encounters a unexpected verb usage, it is natural to try to guess the verb meaning based on verb senses that are already associated with that lexical item. The extended verb sense may use any one of the categories discussed above (or other as yet undefined categories) to form a relation with a candidate sense. Based on the possible relations between a potential extended sense and the candidate verb senses, and the knowledge about the event participants, either the participants can be coerced or a candidate sense can be coerced to find a match. In order to perform coercion successfully in the system, the verb meaning representation must provide all of the possible inter-relations.

7.1. Extending a sense to a superconcept

If the event participants of the unexpected usage come close to satisfying the selectional restrictions for the arguments of a candidate verb sense, then the module will try to relax the selectional restrictions on the verb arguments to include these event participants. One method of relaxation is to coerce the candidate verb sense to its superconcept which usually has more general selectional restrictions, then these restrictions can be applied instead.

For example, using our hand-crafted knowledge base, the system was able to correctly translate the *break* usage in the following sentence from the Brown corpus.

> *No believer in the traditional devotion of royal servitors, the plump Pulley broke the language barrier and lured her to Cairo where she waited for nine months, vainly hoping to see Farouk.*

The input to the system is the verb argument structure (break man-0 lang-barrier-0). It fails to match any of the seven break senses in the system. The numbers here are the satisfaction degree of the selectional restrictions on the arguments for the 7 verb senses.

I-1A	I-1B	I-1C	I-1D	I-2	II-1A	II-1B	II-1C
-797/16	-13/18	-1/12	-21/80	-3/16	-101/2	-8/9	0

The most similar sense is II-1C which means *loss of mechanical functionality*. Its selectional restriction is that the patient should be a MECHANICAL-DEVICE which fails to match *language barrier*. However, in our ontology, a *language barrier* is supposed to be a FUNCTIONAL-ENTITY, and it has been placed in the nominal hierarchy near the concept of MECHANICAL-DEVICE. A possible *loss of functionality* is part of the default knowledge for FUNCTIONAL-ENTITIES. So the system can coerce the *break* sense *loss of mechanical functionality* to *loss of functionality*, acquiring a new set of more general selectional restrictions - i.e., relaxing the original restrictions. The result of this relaxation is:

Old restriction is: (%IS-A %MECHANICAL-DEVICE @VAR2)
New restriction is: (%IS-A %FUNCTIONAL-ENTITY @VAR2)
Old conception is: (%LOSE-MECH-FUNCTION @VAR2)
New conception is: (%LOSE-FUNCTION @VAR2)

Based on this interpretation, the system correctly selects the Chinese verb '打破' *da-po* as the target realization.

7.2. Identifying analogical relations

For analogical relations, the prediction process is a cooperative process between the verb's semantic representation and the built-in knowledge about the event participants. It can be divided into two steps. The first step is to find available information from the discourse model and the knowledge base concerning the event participants, including likely conceptual relationships. In our module, since the implementation is restricted to the verb argument structure level, discourse knowledge is not available, and only the knowledge base information about the event participants is used. The second step is to identify the analogical relations between the candidate verb senses and the likely conceptual relations associated with the event participants in the knowledge base. The similarities between the candidate verb senses and these likely relationships are then measured. The pair which has the highest similarity measure is identified as the most probable coercion, thus identifying the extended verb sense. This is illustrated by the following sentence from the Brown corpus, which translates correctly:

> *Other tax-exempt bonds of State and local governments hit a price peak on February 21, according to Standard & Poor's average.*

In this usage, *the price hitting a certain point* is analogical to an object reaching a point in space. In our system, there is no explicit sense definition of *hit* that would have the appropriate selectional restrictions and conceptual representation for *the price hits a certain point*. However, because we have a multi-domain sense definition, we can find the overlap between the semantic components in the representation of *hit* and in the analogical concept for *reach*.

Hit is defined with the concepts *directed-motion, contact* and *application-of-force*. All of these semantic components have selectional restrictions for PHYSICAL OBJECTS. Clearly *tax-exempt bonds* and *a price peak* are not physical objects and they fail these selectional restrictions. However, the system has the default knowledge that prices can be changed in value and fixed at some value. The requisite concepts are *change-in-value* and *fix-at-value*. It is commonly accepted in the linguistics literature that there are many concepts that are analogous to motion in space, and changes in value can fall into that category - the values can be seen as moving from one point to another [11]. In our implementation it is only necessary for *change-in-value* to be close to *directed-motion*, and for *fix-at-value* to be close to *contact* for these analogical relations to be recognized. The system is able to extend the sense of *hit* to the nearby analogical concepts, and thus inherit a new set of selectional restrictions for application to the sentence. These selectional restrictions require ABSTRACT objects and they are satisfied by *the price*. In this way a new candidate verb sense for *hit* can be formed. Based on the new meaning representation, the correct lexical selection in the target language of 达到 *da-dao* is made. This result is predicated on the definition of *hit* as having concepts in domains that are all structurally related, i.e., nearby in the lattice, to the concepts related to prices.

8. Experimental results

For the testing of the system our coverage was extended to include verbs from the semantically similar *hit, touch, break* and *cut* classes as defined by Levin. Twenty-one English verbs from these classes were encoded in the system. Close to 400 Brown corpus sentences containing these 21 English verbs were selected, among them, 100 sentences with concrete objects that were used as training samples. The verb argument structures (not the entire sentence) were translated into Chinese expressions. The remaining nearly 300 sentences were divided into two test sets. Test set one contained 154 sentences that were carefully chosen as having concrete objects. For test set one, without any encoding of unknown verb arguments, the initial result was an accuracy rate of 57.8% . After adding the unknown nouns as new lexical items and providing them with links to the nominal hierarchy, the accuracy rate rose to 99.45%. The single error in the above experiment is due to an encoding error. The high accuracy rate is reasonable since our lexicon has complete coverage for the concrete senses of *break*, each of which can be clearly distinguished by selectional restrictions.

Test set two contained 116 sentences including sentences with non-concrete objects, metaphorical usages, etc. When the system was run on the second test set, before encoding the unknown verb arguments, the accuracy rate was 31%. After adding the unknown nouns as new lexical items with links in the nominal hierarchy, the rate rose to 75%. Then the extended selection process module was activated, and an additional 13.8% of the sentences containing unexpected verb usages had their translations correctly hypothesized, giving a total accuracy rate of 88.8%. The extended selection process first hypothesizes the most probable source verb sense, then selects the best possible target verb based on the similarity measure.

From these tests, we can see the benefit of associating the individual lexical items with the interlingua conceptual hierarchy which provides a method of quantitatively measuring the similarities among different verb senses. With the extended selection process module, many extended usages were correctly analyzed. The test result is summarized in Figure 9.

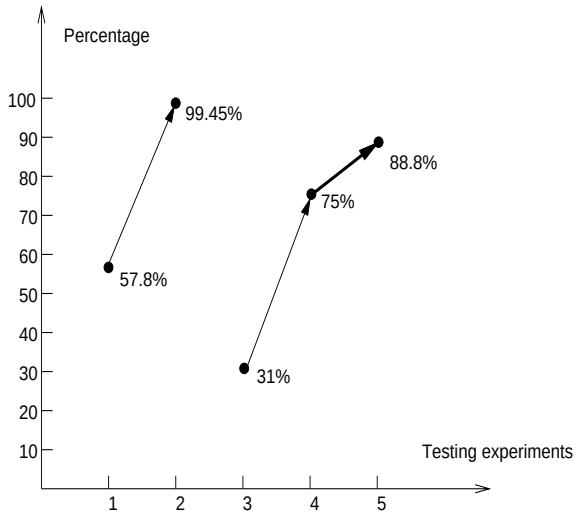

Figure 9. Experimental results

9. Conclusion

Using examples from the translation of English to Chinese, we have shown that lexical divergences among different languages make it difficult to exhaustively list

all possible source/target verb pairs. Selectional restrictions on verb arguments can at best define default situations for verb events, and are often overridden by contextual information. As an alternative we have suggested semantically rich conceptual representations for the verbs that capture these lexical divergences, and have demonstrated that these representations can provide the information necessary for not only correctly selecting target verb senses for well-known usages, but also correctly hypothesizing source and target verb senses for unexpected usages. A cornerstone of this approach is the structuring of the conceptual representations for both languages into an interlingua conceptual hierarchy which makes possible a simple quantitative measure for conceptual similarity, allowing inexact matches to be made. This measure, used in tandem with the standard satisfaction of selectional restrictions, is the basis of the selection of target verb senses, and the hypothesis of possible target verb senses for unexpected usages.

This work is very preliminary, and there are still many areas that have not been touched on. The techniques presented in this paper cannot be extended to larger classes of examples without much more complete conceptual lattices. The problem of verifying the conceptual lattices for each language must be addressed, and the use of automatic or semi-automatic acquisition of lexical knowledge could be very useful for this purpose. We are looking into the suitability of using existing resources such as WordNet, EMICS [4] and the Chinese morpheme database [26]. Identifying language-specific classification schemas is a major research project in itself, let alone the question of whether or not they can be merged into a single, interlingual, conceptual lattice. An alternative to trying to construct such a lattice would be finding methods of automatically matching the lattices for the individual languages. In addition we would like to pursue the influence local context, and in particular the choice of the instrument, has on the selection of the *action* component of the Chinese verb compounds.

Acknowledgments

The work reported here would have been impossible without the assistance of many people both at the National University of Singapore and the University of Pennsylvania. We would in particular like to thank Hsu Loke Soo, Tan Chew Lim, and Lee Cher Leng at the National University of Singapore for their help and support. We would also like to thank Aravind Joshi, Bonnie Webber and the members of the CliFF group at Penn for the valuable discussions and helpful comments, especially David Yarowsky and Libby Levison. Thanks also go to the following persons: Dong ZhengDong at Institute of System Science, National University of Singapore for his kind permission to use his TranStar system in the project, Bonnie Dorr at the University of Maryland for the use of her UNITRAN system and for her valuable comments, and Mark Kantrowitz at CMU for the use of his FrameWork language in our system. Finally, we would like to thank our anonymous reviewers for their extremely constructive comments.

References

1. Yuen Ren Chao. *A grammar of spoken Chinese*. University of California Press, 1968.
2. K. Church. Some statistical opportunities in speech and language. In *Proceedings of 23rd Symposium on the Interface, Computing Science and Statistics*, Seattle, Washington, April 1991.
3. Z. D. Dong. Language, translation, machine translation. In *Proceedings of International Forum on Today's Translation*, Hong Kong, 1987.
4. Z. D. Dong. A categorization system of chinese movement concepts. In *Proceedings of International Conference on Chinese Computing*, Singapore, 1994.
5. Bonnie Jean Dorr. *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge, Masssachusetts, 1993.
6. Dania Egedi. Learning through metaphor. In D. Nerin-Aime, editor, *Knowledge Modeling and Expertise Transfer*. ISO Press, Washington, DC, 1991.
7. Zhichun Feng and XingJian Zhou, editors. *New Chinese Multi-purpose Dictionary*. International culture publisher, 1989.
8. R. Grishman and J. Sterling. Preference semantics for message understanding. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Cape Cod, MA, 1989.
9. Henry Henne. *A handbook on Chinese language structure*. Columbia University Press, 1977.
10. Jerry Hobbs. Overview of the tacitus project. *Computational Linguistics*, 12(3), 1986.
11. Ray Jackendoff. *Semantic Structures*. MIT Press, 1990.
12. Sadao Kurohashi and Makoto Nagao. Dynamic programming method for analyzing conjunctive structures in japanese. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France, 1992.
13. Ronald W. Langacker. An overview of cognitive grammar. In Brygida Rudzka-Ostyn, editor, *Topics in Cognitive Grammar*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1988.
14. Beth Levin. Approaches to lexical semantic representation. In Beth Levin, editor, *Readings for Lexical Semantics*. Northwestern University, 1987.
15. Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, 1993.
16. Longman. *Longman Dictionary of Contemporary English*. Longman, 1978.
17. Sergei Nirenburg, James Carbonell, Masaru Tomita, and Kenneth Goodman. *Machine Translation: A Knowledge-Based Approach*. Morgan Kaufmann Publishers, 1992.
18. Martha Palmer. Customizing verb definitions for specific semantic domains. *machine Translation*, 5(30), 1990.
19. Martha Palmer and Alain Polguère. A preliminary lexical and conceptual analysis of break: a computational perspective. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*. Cambridge University Press, 1995.
20. James Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4), 1991.
21. Philip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, Department of Information and Computer Science, University of Pennsylvania, 1993.
22. Y. Wilks. An intelligent analyzer and understander of english. In Karen Sparck-Jones Barbara J. Grosz and Bonnie Lynn Webber, editors, *Readings in Natural Language Processing*. Morgan Kaufmann, 1986.
23. Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of ACL94*, New Mexico State University, USA, May 1994.
24. Daoping Wu. *On Serial verb Construction*. PhD thesis, Department of Information and Computer Science, University of Maryland, 1991.
25. Zhibiao Wu. *Verb semantics and lexical selection*. PhD thesis, Department of Information System and Computer Science, National University of Singapore, 1994.

26. Chunfa Yuan, Changning Huang, Yingxi Peng, and Zhili Guo. Construction and application of the chinese mprpheme database. In *Proceedings of International Conference on Chinese Computing*, Singapore, 1994.
27. Uri Zernik. The self-extending phrasal lexicon. *Computeational Linguistics*, 13(3-4), 1987.