3-8-1991

The Resolution of Lexical Ambiguity in Machine Translation

Dan W. Higinbotham

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/dlls/vol17/iss1/7

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen amatangelo@byu.edu.
In July of 1949, Warren Weaver sent a memorandum to about 30 friends, suggesting the possibility of using computers to translate text from one language into another. This group and their contacts began research in machine translation, and much of the work in natural language processing (and artificial intelligence) has its roots in the early efforts of this group. Weaver’s memo began by quoting from a letter he had written:

Recognizing fully, even though necessarily vaguely, the semantic difficulties of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate ...

(Weaver 1967, p. 190)

It was recognized from the very beginning that lexical ambiguity would be one of the major stumbling blocks in processing natural language.

1.0 FIFTY YEARS OF AMBIGUITY

This paper will survey some of the methods used in the last 50 years for automatically resolving lexical ambiguity, and will report on a reasonably effective algorithm combining some of the most successful of those methods.

1.1 STATISTICAL METHODS

One of Weaver’s own suggestions was based on the idea that a word may be ambiguous in isolation, but given sufficient context, should be unambiguous. He suggested that a window of N words on either side of a word should be sufficient, for some value of N; it would be an experimental issue to find out how much context was necessary.

1.1.1 LOCAL CONTEXT

The year after the appearance of Weaver’s memo, Abraham Kaplan wrote a paper called “An experimental study of ambiguity and context” (Kaplan 1950). His basic purpose was to discover how much context was necessary for humans to disambiguate ambiguous words. He took examples of ambiguous words from books on mathematics. The words averaged 5.6 dictionary senses. Examples were presented to human subjects with varying degrees of context, namely one or two words before or after the ambiguous word (or both before and after), or within the whole sentence; each subject was asked to decide which sense of the word was being used in that context. After tabulating the results, Kaplan concluded

... A context consisting of one or two words on each side of the key word has an effectiveness not markedly different from that of the whole sentence.

... Under optimal conditions ... ambiguity is reduced from ... about 5 senses to about 1 or 2.

(Kaplan 1955, pp. 46-47)

Optimal conditions were obtained (1) when the translator was trained in the subject of the text (mathematics); (2) when the context included at least one word on each side of the ambiguous word; and (3) when the context words were content words (nouns, verbs, adjectives, and adverbs) rather than function words (prepositions, articles, etc.).

Kaplan’s results were encouraging to many, because the study was taken as evidence that a local context (of two to four words) was sufficient to resolve lexical ambiguity; the study had shown that such a context was nearly as useful as the whole sentence for the purpose of resolving ambiguity.
Although Kaplan's results were taken to mean that not more than 2 words on each side of the word in question, the method of storing all such contexts was never applied directly. Even using a context of one word on each side of an ambiguous word, a vocabulary of M words would require a matrix of size $M^3$, where the value at any position $(X,Y,Z)$ in the matrix would be the most frequently correct sense number of word $Y$ in the local context "X Y Z". For example, in the phrase 'put into a bank account', ignoring the function words 'into' and 'a', it is almost certain that the meaning of 'bank' ($Y$) in the presence of 'put' ($X$) and 'account' ($Z$) is the financial meaning. But, assuming that values in the matrix could be automatically determined based on frequencies in a text tagged with correct sense numbers, and assuming a moderate vocabulary of 10,000 words, this would require a matrix of a trillion entries, and, at a millisecond per entry, would take over thirty years to fill.

1.1.2 FREQUENCY

Early translation programs had simply left the problem up to the reader, by presenting translations of each of the possible senses of a word, separated by slashes. Another method was simply to leave the ambiguous source word in the translated document for a post-editor to fix.

Another common approach was simply to translate the sense of the word used most frequently. This could be determined automatically.

A trans-semantic frequency count is a listing of the words of the source language, together with the various possible renderings of each in the target language, and the frequency of occurrence of each of the latter. Such a listing would resemble a normal translation dictionary, with the addition of information, probably in the form of percentages, giving the frequency of occurrence of each meaning in the target language.

(Pimsleur 1957, p. 11)

Frequency studies could also be based on various types of text.

Alternative frequencies should also be given for various subject areas, scientific, military, etc.

(Pimsleur 1957, p. 11)

Researchers at the University of Washington in 1958 tried categorizing science into about seventy subfields and tagging word senses according to which subfield they most properly belonged to. They reported that so few of the word senses could be marked in this way, that disambiguation was possible in only a small number of cases (Madhu and Lytle 1965, p. 9).

1.1.3 COVER WORDS

A variation in the use of most common translation was the use of "cover-words", which were words "of relatively high semantic frequency which can be used in place of words of lower semantic frequency, with little possibility of misinforming the reader." (Pimsleur 1957, p. 13) A target language "cover-word" would be a word of relatively broad coverage, whose available meanings could "cover" most of the meanings of other more specific translations of a given source word. This would allow the human reader of the translated text to do some of the disambiguation work, based on context, using human intelligence.

1.1.4 CATEGORY COUNTING

Walker and Amsler suggested using frequencies of occurrence of subject codes in the Longman Dictionary of Contemporary English (LDOCE). In the typesetting tape, certain senses of some words are given a four-character domain code. Given a segment of text with ambiguous words, the domain codes on each of the senses of each of the words in the text were assembled and counted. Basically, the sense of a word was chosen whose subject code had the highest frequency in the text. Unfortunately, Walker and Amsler reported that in an eight million word corpus, only 23% of the words were in the LDOCE (Walker 1986), and even of the words that do occur in the dictionary, most of the senses are not marked with subject codes.

1.2 WORD EXPERTS

Small's Word Expert Parser depended on idiosyncratic procedures associated with each ambiguous word in the lexicon. Each procedure was a discrimination net, with a decision tree based on the local context, and sometimes on human input. The
procedures soon became long and cumbersome; the one for the word 'throw', for example, was six pages long (Hirst 1987, p. 83). Unfortunately, developing the procedures was so time consuming that questions of a global nature were just relegated to interaction with an on-line user, thus begging all the more important questions of automatic disambiguation (Adriens and Small 1988, p.18).

1.3 APPROACHES BASED ON A THESAURUS

The organization of Roget’s Thesaurus, which lists content words under more than 1000 concept headings, presented the possibility of a more scientifically based method.

1.3.1 SEQUENCES OF THESAURUS CATEGORIES

Roderick Gould of Harvard had the idea of storing frequencies of sequences of semantic categories, where the semantic category of each word in the sequence would be the correct Roget classification. These would be based on an automatic frequency analysis of a large sample of source language text, in which each content word had been manually tagged with the appropriate Roget concept category number (Gould 1957, pp. 15-27). He was not able to try the idea at the time, but it should be noted that a matrix for sequences of two categories would require a million entries; Kaplan’s data suggests that at least sequences of three categories would be required, and a matrix storing three category sequences would need a billion entries.

1.3.2 COOCCURRING SEMANTIC CLASSES

The Cambridge Language Research Group thought of using the thesaurus in a more practical and immediate way. For example, in the phrase ‘flowering plant’, both ‘flowering’ and ‘plant’ are listed in the thesaurus under the heading ‘vegetable’, so we pick the vegetable sense of both words (Masterman 1957, p. 36). It was soon discovered, however, that in Roget’s Thesaurus, some words were not listed at all, and other words were not listed in all of their senses. The Cambridge group attempted to develop a new thesaurus, in which only synonymous words would be listed under the same heading. This attempt failed to be useful for lexical disambiguation, since words which are in no way synonymous to the correct sense of the ambiguous word can often trigger the selection of the correct sense (Sparck-Jones 1965, p. 97).

1.3.2 THESAURAL CHAINS

A model of chaining in the Thesaurus was developed by Robert Bryan. He defined chains of entries according to word-groups and categories (Sedelow 1986). The basic idea was that if the same two words occurred in more than one category, there was a strong conceptual link between the two words, and also between the two categories. Chains of entries in the thesaurus could be formed so that every link in the chain either connected two senses of the same word, or two categories. If every link was conceptually strong, the entries (which are essentially word-senses) were considered to be semantically related. The thesaurus could be partitioned so that two entries belonged to the same subset if and only if they could be connected by a strong chain. Disambiguation would therefore be possible if a sense of the ambiguous word belonged to the same subset as a sense of some other word in the context.

Once Roget’s Thesaurus was partitioned, however, it was discovered that of the 199,427 entries in Thesaurus, 133,672 of the entries had no strong links with any other entry. The remaining 65,755 entries were partitioned into 5966 sets. One of these sets contained 22,480 entries, or about one-third of the remaining entries. Although some of the entries in this group were instances of the words ‘cozy’, ‘intimate’, ‘snug’, ‘familiar’, ‘close’, ‘near’, ‘tight’, ‘thick’, and ‘compact’, the group also contained instances of words such as ‘vile’ and ‘humble’. Of the other sets, 3373 were formed from only four entries. The other groups ranged in size from six entries to 229 (Talburt 1990). Obviously, such sets are of little help resolving ambiguity in real text.

1.3.3 THESAURAL CATEGORY COUNTING

An approach similar to Walker’s idea of counting categories, as discussed above, was suggested for thesaural categories by John Brady (Brady 1990). Given a text with ambiguous words, the codes for each of the words in the text would be
assembled and counted (using the 1042 groupings in Roget’s Thesaurus). For each word, the sense would be chosen that corresponded to the thesaurus code that occurred most frequently in the text. If this was insufficient to disambiguate a particular word, higher levels of classification in the Thesaurus would be used in the same manner (Sedelow 1990). This method still needs to be tested on a large corpus.

1.4 PREFERENCE SEMANTICS AND COLLATIVE SEMANTICS

1.4.1 PREFERENCES

The most important idea of Preference Semantics as applied to the resolution of lexical ambiguity, is that predicators have preference for certain semantic classes of arguments, and modifiers have preference for certain semantic classes of head. The verb ‘drink’, for example, prefers an animate subject, and the adjective ‘blue’ prefers a concrete head. Therefore incorrect senses of an ambiguous noun may be eliminated if they do not satisfy the preference of a modifying adjective or a governing verb. Formulas were developed to describe word senses, based on about 80 semantic primitives (depending on the version of the theory). The formulas were so hard to write that a full-blown system that could test the accuracy of the approach has never been completed.

1.4.2 COLLATIVE SEMANTICS

In Collative Semantics, a recent extension of Preference Semantics, word senses are the semantic primitives, and each sense of a word is defined by a frame structure. Each frame has arc information, which relates the frame to other frames, creating a hierarchical hyponymy structure, and node information, in which there are features and values, preferences for arguments, and assertions.

Each frame includes an arc which specifies the name of the next higher frame in the hierarchy. By following links, one can find all of the superordinates of a frame in the hierarchy. Preferences can be expressed by giving the name of a frame in the hierarchy. If the preferred frame is not among the superordinates of the frame associated with a sense of an ambiguous word, that sense can be eliminated. Preferences can also be expressed by listing a set of preferred features and values. In this case, the word sense whose frame best matches the preferred list is the one chosen. Again, the frames are difficult to code, so no system has been developed large enough to test the accuracy of ambiguity resolution on real text (Fass 1988).

1.4.3 RELATIONAL TRIPLES

Yet another group using the idea of preferences is the Distributed Language Processing group (DLT) in the Netherlands. The major disambiguation method is based on a hierarchy of word senses, and a set of wordsense:relator:wordsense triplets. These triplets represent commonly expected relationships. The relators are an abstracted form of prepositions or thematic roles. The syntax builds dependency trees, and each pair of words and the connection between them is mapped to a set of possible triplets. The possible triplets are given scores, according to the distance in the hierarchy between the tree word senses and the triplet word senses. For each word, the scores of the triplets in which it is involved are summed, and the word sense with the highest score is chosen (Papegaaij 1986). Initial results were considerably worse than if the most frequent sense of each ambiguous word had been chosen.

1.5 MARKER PASSING

Charniak developed a system of marker passing based on frame structures for knowledge representation. When a word in a sentence is encountered each of its sense frames is marked, even if it is not the sense eventually selected. Then, all frames referenced in these marked frames are also marked, and so forth. Usually, some kind of strength is associated with each mark, and the strength of the mark diminishes each time until it falls below some pre-established bound, after which marking is discontinued. If some frame in the system gets marks from two different origins, the two sense frames of the two words which originated the marker passing are chosen as the correct senses for those two words (Charniak 1983). Again, coding the knowledge frames is so difficult that no tests have been reported that show how viable this method may be.

Several groups have tried to build knowledge bases from machine-readable dictionaries by filling slots in word sense frames with words that appear in
their definitions. Several years ago, I tried marker passing directly within dictionary definitions, skipping the step of building a framed knowledge base. I parsed the definitions from the Longman Dictionary of Contemporary English, assigning grammatical categories to the content words in every definition, giving special status to the head word of each definition. Given an ambiguous word, and a list of context words, each word in the definitions of the senses of the ambiguous words and context words were given a certain level of activation. Then words in the definitions of those words were activated somewhat less, giving different activation strengths to head words. Whenever a word was ultimately activated by both a sense of the ambiguous word and one of the senses of a context word, a value based on the activation strengths from both sources was added to the score of the appropriate sense of the ambiguous word. The sense of the ambiguous word with the highest score was chosen. Various formulas for diminishing activation strength and determining depth of activation were used. When this method was used to resolve ambiguities in real text, the results showed that selection was nearly random. Upon closer examination, it was discovered that unforeseen spurious connections were actually in the majority in most cases. Perhaps if each word in each definition had been marked for its correct sense, resolution would have been better. Nevertheless, this experience casts doubts that any kind of spreading activation in a knowledge base will ever be very successful.

1.6 CONNECTIONIST APPROACHES

The idea of spreading activation naturally appeals to those who are interested in modelling the neural structure of the brain. Automatic learning algorithms have been developed for some kinds of neural networks so that a sequence of input and expected output patterns can be presented to the learning routine, and it will automatically tune the network so that the it will calculate the function implicit in the data. Some networks can generalize and develop internal representations for prototypical patterns; others can fill in missing portions of familiar patterns when only presented with a part of a pattern. Naturally these kinds of automatic learning and generalization capabilities are of great interest to the problem of ambiguity resolution, since it seems that humans learn words and semantic relationships by a process of examining large quantities of data and extracting generalizations. Unfortunately, research into the abilities of neural networks for natural language processing is still in its infancy.

Garrison Cottrell has discussed a localist approach, in which each concept is represented by a separate node in a neural network (Cottrell 1988). There are nodes for each word (level 1) and for each sense of each word (level 2), connected to a mysterious third level of nodes that represent the interconnections of knowledge. Activation is supposed to spread among semantically related nodes and somehow cause the node corresponding to the correct sense of the ambiguous word to end up with the highest activation. Unfortunately, the structure of the third level is too vague to be useable.

Kawamoto uses a distributed model in which a concept is represented not by a single node, but by a certain pattern of activation over a subset of the nodes of the network. He trained the network with patterns for two senses of each of twelve ambiguous words. Each pattern consisted of a pattern of 216 nodes, which included representations of the written form of the word, the phonemic form, the grammatical category, and some ad hoc semantic features. After sufficient training, it was possible to present partial patterns, and the network would fill in the rest of the appropriate pattern, as long as the part of the pattern presented was not ambiguous (it could not belong to more than one of the known patterns). After a pattern had settled into a stable state, connection weights were temporarily modified so that the network would decay out of the stable state. When an ambiguous new partial pattern was presented to the network in this decayed state, the network would settle on the pattern most similar to the previous stable state and also consistent with the new input pattern, thus showing a contextual effect by the previous word. Unfortunately, one word of context is far from sufficient, and it is not at all obvious how to scale up this toy network to a real system.
1.7 SUMMARY

Many of the ideas for resolving lexical ambiguity have been so difficult to implement that no full systems have ever been built to give them a proper test. Some have been abandoned midstream when initial results were negative. Others simply never worked at all. Part of the problem may have been a lack of clear perspective about what kinds of lexical ambiguity problems occur in actual text. The next section will discuss examples of various kinds of ambiguity in real raw text.

2.0 AMBIGUITIES IN REAL TEXT

This section presents the results of an analysis of the types of lexical ambiguities that occur in actual text. The ambiguities found were manually classified according to which of several general approaches might successfully resolve them.

The study was based on four small texts. The first was a newspaper article about some whales which were trapped in Arctic ice (Provo Herald 1988); the second was made up of selections from an article on AI (Dreyfus 1985); the third was from Joseph Smith’s Testimony (Smith 1978); and the fourth was from a LISP manual (Gold Hill Computers 1983). Together, the texts included 3848 words. Of these, 1537 were adjectives, verbs or nouns (about 40%). An English-to-Japanese translation dictionary was used to make a list of all the words in the texts, along with their grammatical categories and various possible translations within each category. Each ambiguous word in the source texts was then annotated by its grammatical category in that context, along with the possible translations for that word within that category. It was found that 883 of the 1537 adjectives, verbs and nouns had intra-category ambiguities (about 57%). Each word was then annotated with the subset of possible approaches which could hope to successfully resolve the ambiguity in its particular context.

The following subsections present each of the approaches with examples, and the results of the study.

2.1 SYNTAX

Complement types and agreement for verbs, countability for nouns, and sentence position for adjectives can be used to resolve ambiguity. For example, in the phrase

As Husserl saw ...

'saw' could either be the past tense of 'see', or 'saw' as in 'cutting wood', but 'cutting' sense fails to agree with the third person subject.

2.2 IDIOMS

Some normally ambiguous words can be resolved because they occur as part of a fixed idiom. For example, one would simply translate the idiom

In the first place ...

rather than worrying about the ambiguity of the word 'place' outside of this context.

2.3 FREQUENCY

Often, simply choosing the most frequent sense yields the correct meaning. For example, in

When the waitress came to the table ...

the most frequent meaning is that of furniture 'table' rather than 'table' of figures.

2.4 TECHNICAL GLOSSARIES

In a phrase in an article talking about knowledge representation, such as

... the script accounts for the possibilities in the restaurant game ...

the word 'script' is likely to mean 'procedure'. The most likely meaning in a computer science text might be 'font', and in a text about the theatre, 'the written form of a play'. Technical glossaries can be used to translate certain words with technical senses when translating texts within a given domain.

2.5 DISCOURSE MEMORY

Often, the context surrounding the first use of a word is more specific than the context surrounding its later uses. Once the intended sense of the word is established by the original context, the same sense is assumed for further instances of the word. For example, 'ice' is listed in the dictionary as meaning
either 'frozen water' or 'sherbet'. If the context

... thick Arctic ocean ice ...

resolves the meaning of 'ice' early in the text, the more difficult phrase later in the text,

... they quickly cleared the ice ...

can be resolved by using the same sense.

Words were marked as resolvable by this method if the correct sense of the word was the same as the correct sense of the previous instance of the word in the text.

2.6 COVER WORDS

Although a source word may be translated by several different target words, one of the possible translations may be more general in meaning than the others, and actually include or "cover" those more specific meanings. For example, the verb 'decide' in English can be translated into Japanese by 'kettei suru', meaning 'decide definitely on'; or by 'kesshin suru', meaning 'decide in one's heart'; or by 'kimeru', meaning 'decide upon'. The last meaning is general enough to cover the other two and the context is usually sufficient for a Japanese reader to understand. In practice, this method would be utilized by entering only the cover meaning in the machine translation lexicon.

2.7 PREFERENCES

A verb or preposition often prefers certain classes of arguments and an adjective often prefers certain classes of heads to modify. For example, in

The ... whales became trapped in the ice two weeks ago while migrating south.

the two listed Japanese translations for 'migrate' were 'idoo suru', meaning to 'move or locomote', and 'ijuu suru', meaning to 'immigrate or emigrate'. The second prefers a human subject, so the first translation is chosen.

2.8 TRIGGER WORDS

Often nearby words give a clue to the proper sense of a word. For example, in the fragment

... California gray whales, whose species is endangered, became trapped in the ice ...

'species' can be translated by 'shu', meaning 'type of living thing'; or by 'shurui', meaning 'type or kind'. In this case, the proximity of the word 'whales' makes it clear that the first is the best translation. The word 'ice' can be translated by 'koori', meaning 'frozen water'; or by 'shaabetto', meaning sherbet. Whales are clearly related to water, and therefore to 'frozen water', but there is no such immediate connection between whales and sherbet.

2.9 LOGICAL REASONING

Some kinds of ambiguities could not be resolved by any of the above methods or any combinations of them. In the phrase

But Minsky seems oblivious to the hand-waving optimism of his proposal that programmers rush in where philosophers such as Heidegger fear to tread ...

'oblivious' can be translated as 'kizukanai', meaning 'unaware'; or as 'wasureppoi', meaning 'apt to forget'. In this case, it seems more likely that Minsky was unaware of his "hand-waving optimism" than that he had once been aware of it, but had forgotten about it. This kind of reasoning is not something that could easily be captured by any of the previous methods for resolving ambiguities.

2.10 RESULTS OF THE STUDY

The following table reports the statistics collected. Lines 1 through 5 give general statistics on the number of words, number of words in the categories surveyed, and the number of polysemous words (with intra-category ambiguities); also the number resolvable by syntactic considerations alone, or by assuming the words occurred in idioms that had been entered in the dictionary.

The second part is based on the number of ambiguities left after the syntactic and idiom methods had been used. The statistics in lines 7 through 12 are given as percentages of the number of remaining ambiguities (shown in line 6) that can be resolved by the given method alone. In line 13, the percentages represent the number of ambiguities
left unresolved by all of methods used in lines 7 through 12 (and combinations of those methods).

<table>
<thead>
<tr>
<th>Text 1</th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>442</td>
<td>1776</td>
<td>1212</td>
<td>439</td>
<td>3848</td>
</tr>
<tr>
<td>1. Total words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>731</td>
<td>411</td>
<td>201</td>
<td>1537</td>
</tr>
<tr>
<td>2. Adj., verbs, nouns</td>
<td></td>
<td></td>
<td></td>
<td>1537</td>
</tr>
<tr>
<td>84</td>
<td>424</td>
<td>248</td>
<td>127</td>
<td>883</td>
</tr>
<tr>
<td>3. Polysemous words</td>
<td></td>
<td></td>
<td></td>
<td>883</td>
</tr>
<tr>
<td>20</td>
<td>93</td>
<td>61</td>
<td>20</td>
<td>194</td>
</tr>
<tr>
<td>4. Resolvable by syntax</td>
<td></td>
<td></td>
<td></td>
<td>194</td>
</tr>
<tr>
<td>12</td>
<td>36</td>
<td>11</td>
<td>10</td>
<td>69</td>
</tr>
<tr>
<td>5. Part of an idiom</td>
<td></td>
<td></td>
<td></td>
<td>69</td>
</tr>
<tr>
<td>52</td>
<td>297</td>
<td>176</td>
<td>97</td>
<td>622</td>
</tr>
<tr>
<td>6. Unresolved by above</td>
<td></td>
<td></td>
<td></td>
<td>622</td>
</tr>
<tr>
<td>46%</td>
<td>39%</td>
<td>26%</td>
<td>39%</td>
<td>36%</td>
</tr>
<tr>
<td>7. Frequency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>15%</td>
<td>3%</td>
<td>23%</td>
<td>11%</td>
</tr>
<tr>
<td>8. Technical glossaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21%</td>
<td>40%</td>
<td>37%</td>
<td>79%</td>
<td>44%</td>
</tr>
<tr>
<td>9. Discourse memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27%</td>
<td>18%</td>
<td>27%</td>
<td>19%</td>
<td>21%</td>
</tr>
<tr>
<td>10. Cover words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27%</td>
<td>11%</td>
<td>15%</td>
<td>3%</td>
<td>12%</td>
</tr>
<tr>
<td>11. Preferences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>9%</td>
<td>13%</td>
<td>3%</td>
<td>10%</td>
</tr>
<tr>
<td>12. Trigger Words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td>2%</td>
<td>5%</td>
<td>1%</td>
<td>3%</td>
</tr>
<tr>
<td>13. Logical reasoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The percentages do not add up to 100% because some ambiguities are resolvable by more than one of the methods. Of the ambiguities not resolved by syntax and idioms, only 2% required some combination of the methods in lines 7 through 12. Each of the methods in lines 7 through 12 resolved between 3% and 9% that could not be resolved by any of the others. It seems clear that some combination of the methods is necessary, and could potentially yield above 95% accuracy with resorting to logical reasoning.

The most difficult to actually deal with are trigger words. It is difficult to determine the correct sense of an ambiguous word based on trigger words, because they can be at arbitrary distances without any syntactic relationship. Sentences like

The ink ran out, so the pen is empty.
The pig ran out, so the pen is empty.

were the kind that led to skepticism about machine translation 25 years ago. The next section discusses one way to approach the problem.

3.0 STATISTICAL COOCCURRENCE

Modern technology has finally made it possible to begin to apply Weaver's idea of local context. Although it is still difficult to store all possible contexts for a word, along with the appropriate sense for each, using statistical cooccurrence can give better results than simply translating the most

frequent sense.

This can be done with the help of a base text which has been translated into the desired target language. The parallel texts must be divided into small segments of local context (usually 1 to 7 sentences), and the divisions correlated so that the nth segment of the target text is the translation of the nth segment of the source text. Since the base text will be used to determine the proper translation of source words based on local context, the number of senses in which a source word is used can be equated to the number of different ways in which it is translated in the target text. Statistically significant cooccurrences can then be used to resolve ambiguity based on local context. Since content words (rather than function words) are most useful for contextualization, function words should be removed from both the base text in both the source and target languages, and the remaining words reduced to base form (this can be done using a tool such as Morfogen; see Pentheroudakis 1991, this volume). Now, using an inverted index and a bilingual dictionary appropriate to the texts, it is possible to determine statistically significant cooccurrences. Specifically, it is possible to find all pairs of source words A and B, such that in n% or more of the cases that A and B occur in the same segment, A is translated as target word T. In order to be statistically significant, some minimum number of cooccurrences of A and B should be required. This information can then be used for translating other texts of a similar subject area. Whenever word A occurs in a local context which includes B, A will be translated by the target word T (unless a higher priority method takes precedence; see the next section).

For example, since the Book of Mormon is already divided into segments (i.e. verses), the English and Spanish versions offer some examples. Function words were removed from both versions, and the remaining content words reduced to base form using Morfogen (all except Moroni's Introduction and 12 chapters of the Book of Alma, which were accidentally left out). The first 80% of the text was used as base text, and the remaining 20% was used as test text.

It was discovered that in the base text, the word 'kindred' was translated as 'parientes' meaning
'relatives' 13 times, as 'familia' meaning 'family' 18 times, and as 'tribu' meaning 'tribe' 8 times. A tabulation was done of English context words which cooccurred with 'kindred' 5 or more times, such that the translation of 'kindred' in the context of the context word was the same 80% or more of the time. It was discovered that 'kindred' was translated 'parientes' in 4 of 5 cooccurrences with the word 'friend'. 'Friend' could therefore be considered a trigger word for the 'parientes' translation of 'kindred.' In the test portion of the text, 'kindred' appears 5 times, 4 times translated as 'familia' and once as 'parientes.' Using only the most frequent sense, 80% accuracy could be achieved. However, the one verse for which 'kindred' was translated as 'parientes' was

And it was the daughter of Jared who put it into his heart to search up these things of old; and Jared put it into the heart of Akish; wherefore, Akish administered it unto his kindred and friends, leading them away by fair promises to do whatsoever thing he desired.

(Ether 8:17)

In this case the trigger word 'friend' also appeared, but none of the trigger words for 'familia' or 'tribu' appeared. Trigger words for neither 'parientes' nor 'tribu' appeared in the other 4 verses, so using statistical cooccurrence therefore led to correct resolution of 100% of the cases of ambiguity of 'kindred' in the test portion of the text.

It was also found that 'judgment-seat' was translated as either 'tribunal' or 'judicial.' Of 50 occurrences in the base text, 42 were translated 'judicial'. 'Judgment-seat' appeared 7 times in the test text, and was translated all 7 times as 'tribunal.' Therefore using the most frequent sense from the base text achieved 0% accuracy in the test text for 'judgment-seat.' The word 'judge' was found as a trigger word for 'judicial', and 'Christ' was a trigger for 'tribunal.' 'Christ' appeared in 6 of the 7 verses in the test text, but 'judge' also cooccurred in 2 of those 6. Using the cooccurrence method correctly resolved 4 of the 7 instances, or 56%.

A similar method was used on parallel texts of approximately a quarter million words of English and French text, which contained a variety of government and non-government documents (this text was obtained from Dr. Alan Melby, from some texts used to test the DLT algorithm discussed in section 1.4.3). The text was divided into over 6000 parallel sections, each of which contained from one to seven sentences. Again, the parallel texts were divided into a sample corpus consisting of the first 80% of the parallel texts, and a test corpus. The sample corpus was searched for pairs of English words that occurred together five or more times within five words of each other; if the pair of English words mapped to the same pair of French translations in 85% of their cooccurrences, the cooccurrence was deemed to be significant, and each source word was considered a trigger word for the given sense of the other. 54 words had such cooccurrence data for more than one sense. 67.3% of the instances of ambiguity of these words in the test corpus could be resolved correctly simply by picking the sense that had been most frequent in the sample corpus. Using trigger words within five words of the ambiguous word in the test text, and defaulting to the most frequent sense in the sample corpus if there was a tie, 76.4% of the ambiguities were resolved correctly.

It is clear that using statistical cooccurrence information is an imperfect means of resolving lexical ambiguity, but that it does give better results than simply using the most frequent sense of the ambiguous word. Its greatest strength comes when used in combination with other already proven methods.

4.0 RESOLVING LEXICAL AMBIGUITY

This section presents an algorithm that includes the methods discussed in section 2 and the statistical cooccurrence method of section 3. It is assumed that the lexicon is coded so that source words are mapped to target cover words when possible to reduce ambiguity.

4.1 THE ALGORITHM

1) If the ambiguous word occurs in an idiom, translate the entire phrase with its idiomatic meaning; otherwise
2) Attempt to reduce the number of possible senses by using syntactic constraints; if only one sense is left, use the appropriate translation; otherwise
3) Reduce the number of possible senses by taking into account argument preferences of verbs and prepositions, and modifier preferences of adjectives and prepositions; if only one sense is left, use the appropriate translation; if no senses are left, ignore the preferences and continue; otherwise

4) If the word has already occurred in the text, use the same translation used in its previous occurrence; otherwise

5) If any of the senses are marked with the same technical area as the text being translated, eliminate other senses; if only one is left, use the appropriate translation; otherwise

6) Use cooccurrence statistics based on parallel texts to identify trigger words. If trigger words indicate one of the remaining senses is most likely, choose it; otherwise

7) Use the statistically most frequent sense.

4.2 THE ORDER AND INTERACTION OF STEPS IN THE ALGORITHM

It is impossible to present the full rationale behind the ordering and interaction of the steps in this paper (but see Higinbotham 1990). It is only possible to give some examples showing the importance of the ordering as given.

Idioms should be given precedence over syntactic conditions.

The scissors cut paper and the hack saw wood.

Even though "the hack" could be third person singular, implying the 'see' meaning, "hack saw" should be one unit.

Syntax should take precedence over preference. In

The man saw a two-by-four.

a piece of wood more closely satisfies the preference of 'to saw', but syntax forces the 'see’ reading.

Preference information can override a word sense used previously in a text, as with 'draft' in

The sergeant drafted a memo to his commander, saying, "Today, we drafted ten more men."

Discourse memory takes precedence over technical tagging. In a military text,

Two water tanks were punctured. Each tank lost over 20 gallons.

the word 'tank' in the second sentence is still being used in the non-M1 sense.

Technical tags can overcome spurious triggering. In a sports column,

The rain had little effect on the pitcher.

still most likely refers to a baseball 'pitcher', even though water pitchers contain water, and rain is water.

Trigger words, by definition, take precedence over the most frequent sense.

4.3 IMPLEMENTATION OF THE ALGORITHM

The ECS Machine Translation Toolkit incorporates all of the steps of the algorithm for resolving lexical ambiguity, in the order shown, including a preliminary undocumented form of the statistical cooccurrence step. Initial tests of the algorithm show a high degree of accuracy.

5.0 FUTURE RESEARCH

Statistical cooccurrence methods deserve considerably more attention. The statistical cooccurrence method was tested in isolation, and could have benefited from accurate tagging of the parallel texts for grammatical category, perhaps ala Church 1989. Ways of scoring competing trigger words still need to be investigated.

It may also be possible to compile statistical cooccurrence information into a neural network, allowing for automatic generalization so that even context words which do not cooccur with a given ambiguous word in the parallel texts may provide some triggering capability.
RESOLUTION OF LEXICAL AMBIGUITY IN MACHINE TRANSLATION

BIBLIOGRAPHY


