Engineering "Word Experts" for Word Disambiguation

Daniel Berleant

Dept. of Computer Systems Engineering
University of Arkansas
Fayetteville, AR 72701
email: djb@engr.uark.edu
phone: (501) 575-5590
fax: (501) 575-5339

(*Received 2 September 1995; Revised 19 January 1996*)

Contents

This paper is dedicated to Prof. Robert F. Simmons, 1925–1994.

Abstract

Every word in the lexicon of a natural language is used distinctly from all the other words. A word expert is a small expert system-like module for processing a particular word based on other words in its vicinity. A word expert exploits the idiosyncratic nature of a word by using a set of context testing decision rules that test the identity and placement of context words to infer the word's role in the passage.

The main application of word experts is disambiguating words. Work on word experts has never fully recognized previous related work, and a comprehensive review of that

work would therefore contribute to the field. This paper both provides such a review, and describes guidelines and considerations useful in the design and construction of word expert based systems.

1 Introduction

A word expert is a set of decision rules — an expert system-like module — for analyzing passages containing a particular target word. It tests the context in which a target word appears, disambiguating or otherwise analyzing the target on the basis of other nearby words with little or no syntactic pre-processing. Thus, word expert based disambiguation relies on the value of nearby context words for disambiguation. This value has been directly addressed for single context words (Yarowsky 1993) and for multiple context words (Gale *et al.* 1993), as well as indirectly addressed by the disambiguation results described in much of the published work on word experts. The context of an ambiguous word contains a number of different sources of evidence relevant to its disambiguation, but in one experiment, using just the best source of evidence led to results that were better than combining the various sources of evidence, presumably since combining evidence for this task is difficult to do well (Yarowsky 1994a). Thus while word expert construction, like related statistical methods, has often relied on corpus-based frequency counts, word expert construction is significantly eased by not having to combine different sources of evidence.

Lexical disambiguation is important and is a focus of word expert work. Table 1 lists some varieties and applications of lexical disambiguation.

Disambiguating numerous different words requires a lexicon containing numerous different word experts. Developing a lexicon of word experts is an engineering task whose goal is a set of modules, approximately one per word in the lexicon. Because new word experts are easily added, a lexicon of word experts is easily extended, may be developed in parallel by several individuals, and in general has desirable characteristics of modularity, decomposability, and incremental scalability. Because the total amount of labor involved in manually creating a large lexicon of word experts is high, automation or at least partial automation is an important goal.

2 Roots of Word Experts

The roots of word experts go back long before the advent of the computer. Basic information about individual words was tabulated at least as early as 2800 years ago (Wieger 1965). Concordances constitute words in context, from which context testing decision rules for use in word experts may be derived. Sizable concordances have existed since at least as early as *circa* 1100 (ben Jehiel 1969). Frequencies of occurrence for the different disambiguations of ambiguous words can be a significant source of baseline evidence for disambiguation. Lorge (1937) (1938) (1949) directed the compiling of such occurrence frequencies just prior to the computer era, hiring lexicographers to painstakingly sift through a sizable corpus of written text and

Table 1.
Some disambiguation tasks. Application domains include computer assisted language learning (cf. the journal Computer Assisted Language Learning) (Berleant et al. 1994); machine translation; automated content analysis (e.g. Stone (1966) (1969)), and information retrieval.

Disambiguation task	**Example**
Sense selection	Does *can* refer to "able" or "container"?
Word translation	The English *can* translates more than one way into most other languages.
Accent restoration	In re-accenting de-accented text, if *resume* means "continue" leave it unaccented; if it means "vita" then accent it.
Part of speech tagging	Is *graduate* used as a noun or a verb?
Automated spelling correction	Should *wurd* be changed to "ward" or "word" or . . . ?
Case checking and recovery	In case recovery from upper-cased text should *LISP* be left upper case (a computer language) or made lower case (a speech impediment)?
Homograph pronunciation selection	Should *bass* be pronounced as in the fish or the instrument?
Homophone de-transliteration	In recovering the original Chinese character for the transliteration *chiou* there are several possibilities.
Abbreviation expansion	Restoring vowel diacritics in texts written in semitic alphabets.

tabulate the frequencies of word senses distinguished in the 1933 edition of the *Oxford English Dictionary.*

3 *Circa* 1970: Word Expert Research Begins

The advent of the computer brought great promise for automated processing of natural language, leading to the application of computers to word experts, first proposed by Stone[1] et al. (1966) on pp. 152-167, later implemented (Stone 1969), and subsequently automatically acquired (Weiss 1973)[2]. Stone's word experts were derived from passages each containing the word to be disambiguated, several words before it, and several words after it. Weiss's word experts were derived from sentences containing the word to be disambiguated. Decision rules tested the context of an ambiguous target word for the presence and location relative to the ambiguous target word of various other words, then concluded the meaning of the target.

Stone's application was automated content analysis (1966); Weiss's was information retrieval (1973). Both content analysis and information retrieval can benefit from semantic disambiguation because they are ideally based on concepts, not the often conceptually ambiguous words used to specify concepts.

3.1 Design and Examples

Both Stone and Weiss designed word experts as sets of decision rules — miniature expert systems. A rule's left hand side specified requirements on the context words surrounding the target word to be disambiguated. If these context requirements were satisfied then the right hand side concluded the meaning of the target. Weiss provided two different ways for a rule's left hand side to express requirements on context.

1. A specific context word is required to be a specific positive or negative distance from the target. An example would be,
   `IF:` the word after *provide* is "for"
   `THEN:` *provide* means "furnish support."
   This decision rule applies to the example:
   S1. The animal shelter will *provide* for abandoned pets.
2. A specific context word is required to be a nonspecific distance from the target. An example would be,
   `IF:` "music" appears near *band*
   `THEN:` *band* means "group of musicians."
   This decision rule applies to the example:
   S2. The music of that *band* is unusual.

Stone also provided for two additional kinds of rules:

[1] Philip Stone is with Harvard University.
[2] Stephen Weiss is with U. of N. Carolina, Chapel Hill.

3. A nonspecific member of a category of context words is required to be a specific distance from the target. An example would be,
   `IF:` the word before *stop* is a transit method
   `THEN:` *stop* means "a place for passengers to get on or off."
   This decision rule applies to the example:
   S3. The trolley *stop* has two benches.
4. A nonspecific member of a category of context words is required to be a nonspecific distance from the target. An example (after Stone (1969) p. 211) would be,
   `IF:` a "do" verb appears within four words preceding *matter*
   `THEN:` *matter* means "having significance."
   This decision rule applies to the example:
   S4. It does not really *matter* much.

Stone and his then student Kelly later provided some less frequently useful extensions to these rule types, such as testing for absence rather than presence of context words. The final form of their word expert architecture is described in Kelly and Stone (1975) on pp. 23–29; see also Stone (in preparation). However, the main difference between their architecture and that of Weiss is in their reliance on testing for nonspecific context words. This is a significant difference because of the resultant frequent necessity to determine the categories of context words in order to determine whether a rule's `IF` condition is satisfied. Unfortunately, determining a context word category involves more than simply looking it up in a table, because the category can be ambiguous. Thus two word experts WE_1 and WE_2 can potentially deadlock, with WE_1 waiting for WE_2 to disambiguate the category of its target word so that word's category can be tested by WE_1, and likewise with WE_2 waiting for WE_1 to disambiguate the category of its own target word, which is required by a rule in WE_2. Kelly and Stone ((1975) p. 31 item c) discuss this problem further.

Kelly and Stone provide detailed explanations of their word experts for *last* and *to,* and a twelve page description (p. 99 ff.) of how their system disambiguates the italicized words in the sentence

S5. But *rather — just like* my *relative,* he *grew rather upset.*

They provide a sizeable manually generated lexicon of word experts for sense disambiguation, over 1000 of them, many of which handle both base and suffix forms and many of which include performance data, taking a total of 145 pages.

Both Stone and Weiss found their word experts to be generally effective. Stone (1969) pp. 217–218 found a precision of 97.5% for the word *like* and 89.6% for the word *that* (see also Table 3 of this paper). Weiss (1973) found an average precision of 96% and an average recall of 90% for the words *degree, type, volume, board,* and *charge* (see also Table 5 of this paper). Note that strict comparisons between Stone's and Weiss's performance figures are impossible because they are based on different ambiguous words and because of differences in criteria for what constitutes a sense of an ambiguous word (Kelly and Stone (1975) distinguish only a "bare minimum" of senses (p. 13) many of which corresponded to part of speech

distinctions (pp. 3, 11, 16)). The number of potential disambiguations of a word depends on the lexicographer and the application, and the more disambiguations that are distinguished or the less clearcut the distinctions among them, obviously the more difficult disambiguation will tend to be.

4 *Circa* 1980: Complex Communicating Word Experts

Whereas workers such as Kelly and Stone often used word experts of about a half dozen lines, Rieger and Small considered a word expert of a half dozen *pages* deprived. While their word expert for *throw* had six pages of code, they felt it should have been "10 times that size" (Small and Rieger 1982). Their word experts not only disambiguated but performed natural language understanding roles as well.

Chuck Rieger[3] began using word experts for natural language understanding in the late '70s (1978) (1978a) (1979). He was soon joined by his then student Steven Small[4] (Rieger and Small (1979), (1981)) who ultimately led the research (Small (1979) (1980) (1981) (1983)) (Small and Rieger 1982). Geert Adriaens[5] (1986) (1987) extended the work to Dutch. Adriaens and Small (1988) provide a retrospective. The research heavily influenced ongoing work by Adriaens' group (Adriaens 1989) (Devos *et al.* 1988) (Devos and Adriaens 1994) and Udo Hahn[6] (e.g. (1989) (1994)). A description of the Rieger-Small approach is next, emphasizing the word sense disambiguation aspect of their work. Section 4.2 synopsizes the literature.

4.1 Disambiguation and the Rieger-Small Approach

The first paper (Rieger 1978), largely speculative, describes a game suggesting the potential usefulness of interactions among word experts.

Game. *Each player is given a different word from a sentence which is kept secret, and takes on the role of word expert for that word. Players try to disambiguate their word and understand its role in the sentence by asking questions of other players. The group thus informally simulates interacting word experts.*

Participants demonstrate "surprising" accuracy in disambiguating their respective words and even in interpreting the entire sentence ((1978) p. 134). Such a game can shed light on the kinds of communication that are most useful among word experts, and suggests protocol analysis and human simulation as potentially useful tools in word expert construction.

In the Rieger-Small approach several word experts are invoked on a sentence, one for each word in it. The word experts are designed as coroutines which communicate among themselves, sending and receiving messages. This increases the

[3] Charles (usually publishing as "Chuck") Rieger did his work while at the University of Maryland, College Park.
[4] Small is with the University of Pittsburgh.
[5] Adriaens is with the University of Leuven, Belgium.
[6] Hahn is with the University of Freiburg, Germany.

potential disambiguating power of a word expert because it can use as evidence not only the identity of nearby words but additional facts supplied by their respective word experts. Consider for example a word expert attempting to disambiguate the translation of the Chinese word 一个 (pronounced "yi-ge") into English as "a" or "an" given a sentence

S6: . . . 一个 海洋. . .

It would help that expert greatly to find out from the 海洋 ("hai-yang") expert whether the current occurrence of 海洋 is best translated as "ocean" or "sea," so that S6 could be correctly translated as ". . . an ocean . . ." or ". . . a sea . . . ."

The Rieger-Small school of word expert based analysis aimed to do not only word disambiguation, but sentence understanding as well. Consider for example two sentences provided by Small (1981):

S7. The man eating tiger growled.

S8. The man eating spaghetti growled.

The sense of the word *eating* is similar in both cases, yet who or what growled and what is or may be eaten differs greatly. Sentence understanding is harder than word sense disambiguation, so more sophisticated word experts are implied. Rieger and Small needed to:

- infer not only the meaning of a verb but also its subject, and its object if present.
- be able to analyze any word, even closed class words, which are more difficult to handle (see Section 7.2 item 2. of this paper).
- use a blackboard to accomplish the "messy details" (Cottrell 1989) of control and communication among word experts.

The complexity of these tasks led to complexity in their word experts. The size and complexity of their word experts contrasts with the small size of Stone's and Weiss's word experts. While Stone and Weiss built effective word experts consisting of just a few decision rules each, building word experts in the style of word expert parsing "involves . . . intricate labor" (Small 1980). Their word experts are decision tree-like structures whose leaves each represent a sense of the word. The details and names of this structure change from paper to paper, reflecting their uncertainty about exactly how to structure the word experts.

The reason for these differences between Rieger's and Small's word experts on the one hand, and Stone's and Weiss's on the other, is the output requirements for word expert analysis: disambiguation alone for Stone and Weiss versus full scale natural language understanding for Rieger and Small. With a goal like NLU it is not surprising that their approach was strained to its limits or perhaps past them.

4.2 Synopsis of the Literature

After Rieger's first speculative publication (1978) he next proposes a blackboard to assist in the goal of story understanding (1978a). Rieger (1979) summarizes the main points. Communication among word experts implemented as coroutines was also added in 1979 (Rieger and Small 1979). The implementation was not

yet working on complete sentences (p. 728) although that *caveat* is later removed (Rieger and Small 1981). Small (1979) describes how the system understands the sentence

S9. The deep philosopher throws the peach pit into the deep pit.

Sentence S9 contains the words *deep* and *pit* twice each, with different meanings each time. Small's dissertation (1980) provides an account of understanding a sentence containing another sense of the root *throw,*

S10. The case was thrown out by federal court.

In 1981 Small argued for the validity of word expert parsing from a psycholinguistic viewpoint (1981); updated arguments appear in Adriaens and Small (1988). It certainly seems that people do not need a word and its disambiguating evidence to be part of the same syntactic parse structure. This is obvious in the case of relatively distant pragmatic evidence, and is also true for evidence that is quite local to the ambiguity in time and space (Berleant 1982).

Small and Rieger (1982) describe the architecture of their word experts (the "Sense Discrimination Language") and of the interactions among word experts (the "Lexical Interaction Language"). These architectures are exemplified by a word expert for the word *deep,* and no less than a thirty-two page account of how the system understands the sentence

S11. The man eating peaches throws out a pit.

Small (1983) contains a similarly detailed trace of how the system understands the sentence

S12. The man throws in the towel.

Meanings of *throw* and its inflections were examined closely by Small and his associates. See also Cottrell and Small (1983) and Small et al. (1982).

Adriaens's group extended the approach to Dutch in 1986 (Adriaens (1986) (1987)). Their extension also includes parallel execution of the experts, which required "drastic revisions" (Devos *et al.* 1988) to Rieger and Small's ostensibly parallelizable system design. Their implementation was in Flat Concurrent Prolog, rather than LISP which Rieger and Small used.

Hirst (1987) and Cottrell (1989) provide some commentary on the approach. A retrospective by Adriaens and Small appeared later (1988). Ongoing follow-on work by Adriaens and his associates (Devos *et al.* 1988) (Adriaens 1989) (Devos and Adriaens 1994) emphasizes parallelism and NLU but says little about disambiguation. Ongoing follow-on work by Hahn (1989) (1994) deals with NLU (including disambiguation of anaphoric references) using generalized experts for large word categories.

5 Acquisition of Word Experts from People

Yaacov Choueka[7] and Serge Lusignan (1985) investigated disambiguation of representative common ambiguous French words (Table 2). They considered context

[7] Choueka is with Bar-Ilan University, Israel.

Table 2.
Choueka and Lusignan (1985) chose 31 representative target words from the most frequent 500 in a turn of the century French text. A target was deemed ambiguous if its part of speech, sense or both was ambiguous. Of all occurrences of the 31 words, 88% used the most common disambiguation.

# of disambiguations	% of set
2	75%
3	22%
4	3%

Preponderance of most common disambiguation	% of set
$\geq 90\%$	55%
$70\% - 89\%$	26%
$< 70\%$	19%

words no further than two words from the ambiguous target. An interesting aspect of their work is their structured approach to knowledge acquisition from human informants.

All the rules in a word expert are similar in that they all specify the same type of context to match against. The possible types are: context preceding the target ("pre-context"), context succeeding it ("post-context"), or context both preceding and following it ("symmetric context"). A rule LHS specifies word(s) and their location(s) that a context must satisfy for the rule to conclude the target's disambiguation.

Determining whether the disambiguation rules in a given word expert should best use the pre-, post-, or symmetric context type was treated as a knowledge acquisition task. College educated native French speakers without special training were queried. These informants chose pre-contexts for 68% of the targets, post-contexts for 22%, and symmetric contexts for the remaining 10%. In making those choices, "intuition is an excellent guide" and the choices were "almost always . . . correct" (Choueka (1985) p. 152).

Once the context type was determined, all the contexts appearing in the corpus of that type and for that target were printed out up to a distance of only one word from the target. Informants decided for each whether or not that "1-context" is sufficient to disambiguate the target. If so, the 1-context became the LHS of a rule and the disambiguation became the RHS. If not, the same procedure is repeated except that the context length is increased to incorporate context words up to a distance of two (a "2-context"). Of 2,841 different 1-contexts, 77% were deemed sufficient for disambiguation. Allowing informants to use a 2-context for a rule LHS when they deemed a 1-context insufficient led to disambiguation rules covering 91% of the cases. The error rate depends on how it is measured; figures ranging from 1.3% to 2.1% overall for different measurement methods are given. However,

```
Let C be an accented corpus
Let C' be a copy of C
De-accent C'
Re-accent C' (e.g. using word experts)
Evaluate performance by comparing C' with C
```

Fig. 1. Outline of algorithm for automated evaluation of word experts.

perhaps reflecting the relatively greater complexity of the task, the error rate for rules based on 2-contexts was about five times higher than that for rules based on 1-contexts, and the error rate for targets words with three disambiguations was about four times higher than that for binary ambiguities.

6 *Circa* 1990: Automatically Acquirable Word Experts

David Yarowsky[8] and his associates have demonstrated automatically acquired word experts for two forms of disambiguation. One is determining the pronunciations of homograph occurrences (Sproat et al. (1992) Section 4). Homographs are single spellings corresponding to two or more unrelated words which may have different pronunciations, such as *bow* (bend) vs. *bow* (front of a boat), and *bass* (fish) vs. *bass* (sound). Disambiguating homograph pronunciations is useful in text-to-speech systems.

The other disambiguation form investigated by Yarowsky is recovering proper accenting for de-accented Spanish and French text. For example *terminara* and *terminará,* different tenses of a Spanish verb meaning "terminate, end, finish," appear identical when de-accented presenting a disambiguation problem for an accent restoration system. One positive feature of the accent restoration task is that performance evaluation can be automated (see Figure 1).

Yarowsky's word experts are each composed of an ordered list of rules, as in the earlier work of Weiss but in contrast to the simple decision trees of Stone and Kelly and the complex decision tree-like structures of Rieger and Small. The word experts are generated mostly automatically from the corpus (see Section 7.3 of this paper). The rules composing an expert are constrained in their expressive power. Rules for specific distances try to match either the word to the right of the target, the word to the left, the words directly to the right and left, the two words to the right, or the two words to the left. Rules for nonspecific distances try to match any word within a $\pm k$ window centered on the target (he does not specify what value of k was used). As with Stone, rules specify either a particular word to match or a word category if the system provides them (his system provides word categories for Spanish but not French).

Some of Yarowsky's results for accent restoration are summarized in Table 3.

[8] Yarowsky is with the University of Pennsylvania, Philadelphia.

Table 3.
Some performance results; figures reflect %-age of correct disambiguations. The naïve method is to always pick the most common disambiguation. Spanish results were obtained from a 49 million word corpus using a system containing automatic morphological analysis, a part-of-speech lexicon, and some predefined semantic categories. French results were obtained from 20 million words using a system without such additions. Results are Yarowsky's unless otherwise noted. **Notes:** *(a) figure for French was not given; (b) these figures are derived from the table "Performance on Individual Ambiguities" for a "random set of the most problematic cases — words exhibiting the largest absolute number of non-majority accent patterns" (Yarowsky 1994); (c) figures are means over the problematic words weighted by their frequencies; (d) figures are unweighted means over the problematic words; (e) figures are derived from Kelly and Stone (1975) p. 47; (f) figures derived from Yarowsky (1993) Table 3 (see also this paper, Section 7.2 item 2.).*

(See notes)	Portion of corpus tested (and language of corpus)	Word Expert method	Naïve method
(a)	All words (Spanish)	99.6%	98.7%
(a)	Ambiguous words (Spanish)	98%	93%
(b)(c)	Problematic word occurrences (Spanish)	96.5%	60.1%
(b)(c)	Problematic word occurrences (French)	96.5%	62.0%
(b)(d)	Problematic words (French)	96.4%	60.2%
(b)(d)	Problematic words (Spanish)	92.7%	62.8%
(e)	Sense selection (English)	92.1%	68.5%
(f)	Binary sense selection (English)	92-97%	69%

7 Designing and Building Word Experts

What guidance can be provided to those wishing to write word experts? A background in traditional syntactic parsing is unnecessary. Rather, builders should consider carefully the ways in which a particular word is used, and exercise judgement and cleverness in constructing its expert, because the word expert approach emphasizes the idiosyncratic nature of a word. Nevertheless, general heuristics exist:

1. It is often useful for a word expert to have a default decision rule, fired as a last resort when no other rule matches the passage. A default rule would output the most common disambiguation. Relative frequencies of meanings, e.g. Lorge (1937) (1938) (1949) may be useful in this. Meanings ordered by frequency appear in some dictionaries carrying the Barnhart name. For exam-

ple, Barnhart (1967) is explicit about its reliance on Lorge's work; Barnhart (1993) simply states that meanings are ordered by frequency. An advantage of a domain like accent restoration is that baseline disambiguation frequencies can be found by mechanically counting occurrences in a large enough corpus.

2. The rule representation language should allow a rule LHS to specify either a specific or a nonspecific distance. While a specific distance for a context word can provide precise control over whether a decision rule will match a given text passage, a nonspecific distance results in a more general decision rule (cf. Section 3.1). Both rule types can be written in terms of ranges: an example of a nonspecific distance is "from 20 words before to 20 words after," whereas an example of a specific distance written as a range is "from 2 to 2 words after."

7.1 Word Experts Can be Easy to Construct

With only word translation as a goal, this author's experience is that students in a second undergraduate computer programming course can construct a small word expert as a regular homework assignment, and a master's degree student can write dozens of word experts without difficulty as part of a three credit Master's Project (Prasad Sunkara tested nine for recall and precision, then revised them. Upon retesting the nine all scored 100% on both recall and precision for the (few) sentences used in testing them.)

Stone ((1969) footnote 1) relied on undergraduates in his work although it is not clear what they did.

Choueka and Lusignan used educated individuals without specialized training in the area to generate rules for their word experts.

Thus, despite the difficulty of writing word experts in the Rieger-Small style, it appears that the word expert approach is a quite accessible technology when intended for word disambiguation.

7.2 Design Tradeoffs in Word Experts

In contrast to design heuristics are design tradeoffs. A tradeoff in word expert design typically balances, along some dimension, the sophistication of the word experts against the effort involved in creating them. The optimal level of sophistication along a given dimension will vary with the application, and depend on the answers to questions such as these:

- **Performance:** how well can the word expert perform with a less sophisticated design, and how much can be gained by increasing its sophistication along some dimension?
- **Cost effectiveness:** would the benefits of the more sophisticated design outweigh the additional cost?
- **Comparative cost:** is increasing the sophistication of the design along one dimension more cost-effective than increasing the sophistication along some other dimension?

Here are some major dimensions along which sophistication can vary and tradeoffs be made in word expert system design.

1. Size of the word experts. Stone (1969) p. 213 found that the word *think,* whose ambiguities had been troublesome in their content analysis work, was disambiguated correctly in 1574 of 1575 occurrences using a word expert of only four decision rules. Yet Small advocates a word expert of 60 pages of code for the word *throw* (Section 4, this paper).
2. Degree to which the category of the target word determines the design of its expert. Word expert architectures might or might not take advantage of this information. Of course, this category could be ambiguous.
   One categorization to consider is a simple distinction between open and closed class words. This would be useful in that one might want to avoid writing word experts for closed class words because the task is more difficult. Kelly and Stone ((1975) p. 12) were unable to write effective experts for almost all of them. Cerri (1989) describes some problems and solutions in machine analysis of closed class words.
   Another categorization to consider is distinguishing among parts of speech, since different parts of speech tend to respond differently to the same word expert architecture (Yarowsky 1993). In investigating binary sense ambiguities both of which were of the same part of speech, Yarowsky found that nouns and adjectives tend to respond to local disambiguating evidence better than verbs ((1993) Table 3), and that nouns tend to respond to distant evidence significantly better than ambiguous verbs and adjectives ((1993) Figure 1). Binary ambiguous nouns, verbs and adjectives immediately preceded or followed by a noun, verb, adjective or adverb context word known to the system were disambiguated correctly on the basis of that context word 96–97% of the time. When the nearest such context word was not adjacent, correct disambiguation decreased to 92–94%.
   A reasonable approach would be to model all or part of a new word expert after previously written word expert code for a word deemed similar. This amounts to *ad hoc,* fine-grained, implicit categorization.
3. Whether or not a rule LHS must designate specific context words to match, or instead can designate categories of words to match. Allowing a decision rule to test for a context word based on its membership in a predefined category (cf. rule types 3. and 4. in Section 3.1) can shorten a word expert since the members of the category need not each be specified individually. A reference to a category containing N words is logically equivalent to a reference to a disjunction of N individual words, where N is high for a category like NOUN. However creating a category system has a cost. Category ambiguity (where different disambiguations of a word belong in different categories) becomes more troublesome as the complexity of the category system increases. Category ambiguities can also lead to interaction problems among word experts (Section 3.1). Defining and implementing an appropriate system of word cat-

egories can be a significant task. However there is an economy of scale in that once created, a category system can be used by many different word experts. Easier than creating a taxonomy from scratch is to use one already available in the literature. Parts of speech constitute a well known taxonomy for which automatic taggers are available, e.g. see (Natural Language Software Registry) in the references of this paper. Stone (1969) pp. 208–209 describes a taxonomy created specifically for word expert construction. Kelly and Stone ((1975) p. 16) update it slightly and point out that while subdividing existing categories and adding new ones are feasible, rearranging categories is "hideously complex" because word experts that use the previous categories would then need modification. Consequently their taxonomy, while "serviceable" (p. 16), would "merit the most effort" (p. 22) if their word expert system was redesigned. Various other taxonomies such as semantic heirarchies and networks exist (Allen 1995).
Categorizing a context word as a "content" or "function" word is simple, and useful in estimating its disambiguating effectiveness; function words are much less effective (Yarowsky (1993) Section 6.2).

4. Whether or not different word experts working on different words in the same passage can communicate. Conclusions about one word may be useful in making conclusions about other nearby words (see Sections 3.1 and 4, Mathew (1993), and sentence S6.), but implementing communication among word experts can be quite complex as illustrated by the works of Rieger, Small, Adriaens, Hahn, and their associates. The system designer must decide whether the increment in performance outweighs the high added costs for the application under consideration.
Kelly and Stone (1975) observed that if a disambiguation rule LHS requires an ambiguous context word in a particular one of its senses, if the context word is found near the target as specified by the rule it will "almost invariably" have that sense (p. 36). Yarowsky (1993) formally investigated such issues, concluding that there is a strong tendency toward "one sense per collocation."

5. Whether or not world knowledge is available to the word experts. World knowledge can be useful in disambiguation, e.g. Hahn (1989). For example consider the famous sentence
S13. Time *flies* like an arrow.
World knowledge allows us to rule out the meaning "kind of insect" for *flies* because there does not happen to be a kind of fly called a "time fly," and flies and arrows have no particular affinity. However the potential benefits of world knowledge must be weighed against the cost of including it. Disambiguation can work to a significant extent without it, so including it can merely provide improvement. Since representing and using world knowledge is a major problem in artificial intelligence, it is better to avoid it when building word expert based systems unless the level of performance required by the application can-

not otherwise be met, as is the case for full natural language understanding (Section 4) — another major problem in artificial intelligence.[9]

6. Expressiveness and flexibility of decision rule LHS's. At the simple end of the spectrum are rules that check the context of an ambiguous target for a specific context word at a specific distance, and rules that check for a specific context word at a non-specific distance of $\pm k$ where k is defined globally for all rules (cf. Section 3.1). Choueka and Lusignan (1985) used $k = 1$ for many rules and $k = 2$ when $k = 1$ was deemed insufficient. Weiss ((1973) p. 38) used $k = 5$. Stone (1969) specified $+k$ and $-k$ independently for each rule; examples range from $k = -4$ to $k = +1$. Gale et al. ((1993) pp. 419, 427–429) conclude $k > 20$ is warranted, used $k = 50$, and found correlations between context words and target meaning up to a distance as high as 10,000 (ten thousand) for the Canadian Hansards corpus.[10]

   More complex are boolean functions of words — their presence, absence, conjunctions, disjunctions, and combinations of those. (A word category, for example, is expressible as a conjunction.) Still more complex are weighted functions of potential context words. Weights can be determined by context word identities and/or by their distances from the target. A complex but intriguing approach to matching is to use sophisticated case based reasoning (Kolodner 1993) to retrieve and use a stored case most similar to the current context (Schaller on Kitano (1993)).

   Word expert designers should pick the simplest matching strategy that will work for the application, to ease building and using a word expert based system.

7. Whether a given expert should only apply to one word or instead to a set of morphologically related variants. If only one word, then a lexicon would need a distinct word expert for each word it contains (the "full listing" approach of Adriaens and Small (1988)). Kelly and Stone (1975) have each word expert handle both the base and common suffix forms of a word, but indirectly support full listing by writing "it is often convenient to fragment a disambiguation into subproblems corresponding to different endings or groups of endings" (p. 16). On the other hand, some of Yarowsky's experts work well with many words from a particular category. For example, the same expert can correctly determine the proper accenting of the suffix for many Spanish verbs ending in *-ara* or *-ará*. Rieger and Small's system does morphological analysis prior to calling word experts, thus eschewing the full listing approach. Full listing can in principle lead to better disambiguation performance because an expert about a given word may not apply completely to even very closely related variants. However full listing might seem inefficient if similar

[9] On the other hand, a word expert based system could be used as a vehicle for research on representing and using world knowledge.

[10] Ambiguities of a syntactic character (part of speech, for instance) tend to respond to nearby words (witness the success of n-gram part-of-speech taggers), whereas semantic ambiguities are more likely to respond to words that are farther away (Yarowsky (1994) (1994a)).

Table 4.
Full listing means that every word handled by the word expert lexicon has its own expert. The alternative is for morphologically related words to be processed by the same expert.

Advantages of Full Listing
Different forms of some roots are best handled separately.
Some words cannot be morphologically analyzed before disambiguation.[11]
Morphological analysis takes significant effort.
Disadvantages of Full Listing
An unlisted word cannot be analyzed.

experts for related words are written, stored, and used independently. Fortunately this seeming inefficiency could be alleviated by having different experts call the same code routines as appropriate, thus avoiding inefficient code duplication. Full listing offers fewer obstacles to getting a word expert based system up and running since morphological analysis is not then a prerequisite. Adding morphological analysis later is a natural extension that would address disambiguation of some words not present in the word expert lexicon. Table 4 summarizes advantages and disadvantages of the full listing strategy.

7.3 Automated Acquisition of Word Experts

Partial or full automation in word expert acquisition is an attractive goal because of the large amount of human labor required to build a lexicon containing numerous word experts. Automatically acquiring word experts means automatically acquiring knowledge about how the context words of an ambiguous target word affect its disambiguation. On-line data has been used for automated acquisition; partial automation by computer directed questioning of human informants is an untried alternative although Choueka and Lusignan's acquisition method (1985) (Section 5 of this paper) could be easily so adapted.

7.3.1 On-line Data to Support Automated Acquisition

Computer readable resources are increasingly available from such sources as the CLR (Consortium for Lexical Research), the OTA (Oxford Text Archive), the LDC (Linguistic Data Consortium), and others locatable by browsing the World Wide Web.[12] The two main varieties of resource are reference works such as dictionaries and thesauri, and text corpora (bodies of relatively ordinary text). Text corpora

[11] For example the Dutch word *kwartslagen* can be from either *kwarts+lagen* ("quartz layers") or *kwart+slagen* ("quarter beats") (Adriaens and Small1988).
[12] `http://xxx.lanl.gov/cmp-lg/` is one possible starting point.

are useful because they contain many examples of target words in disambiguating contexts.

Reference works and corpora can be in one, two, or more languages. Bilingual corpora are called "parallel texts." Parallel texts are typically available coarsely aligned on document or file boundaries. Section headings, sentence lengths (Gale and Church 1993), and other clues may be used to compute other alignment points; improved alignment techniques are an active area of computational linguistics research. A "bi-text" or "bitext" is a parallel text aligned closely so that it associates the corresponding translation units, typically phrases (Harris 1988). A bi-text, due to its pairs of corresponding translation units each containing words, their translations, and the immediate context words (which usually determine those translations), can be used as data to generate word experts for translation (Section 7.3.5). Section 6.1.2 of Sadler's (1989) description of the well known DLT machine translation system comments further on the value of the bi-text concept.

7.3.2 Previous Work

Weiss and Yarowsky are the only word expert researchers to address automated acquisition of full word experts. Brown et al. (1991) however describe a method for automatically deriving a single optimal context checking rule for disambiguating a word instance between two senses (or two sense categories if the word has more than two senses) for translation purposes, using heavily tagged and aligned input examples.

Weiss and Yarowsky both describe acquisition algorithms for full word experts. Their algorithms each start with a somewhat *ad hoc* rule generation step, followed by an also somewhat *ad hoc* rule deletion step. We summarize Weiss's method first, followed by Yarowsky's.

7.3.3 Automated Acquisition: Weiss

Weiss's approach to automated acquisition (1973) extracts a word expert for a given ambiguous target word from a set of sentences, which each contain the target and are hand tagged with its meaning. A sentence is converted into a set of rules for disambiguating the target by using the identities and positions of its context words relative to the target to generate rules whose left hand sides match those identities and positions and whose right hand sides conclude that the disambiguation of the target is its meaning in that sentence.

For example, consider the ambiguous target *wild* in training sentence S14 which has been tagged with the meaning "untamed":

S14. The rhinocerous is a *wild* animal.

The following rules are generated:

1. *Wild* means "untamed" if the next word is "animal."
2. *Wild* means "untamed" if the previous word is "a."
3. *Wild* means "untamed" if "is" appears two words before it.

4. *Wild* means "untamed" if "rhinocerous" is within five words before.

Note from the example that common words like *the, is,* and *a* are used to generate specific distance rules but not nonspecific distance rules, and that rules for specific distances of only one or two are generated, but rules for nonspecific distances of up to five are generated. Unfortunately, rule 2. will disambiguate incorrectly given a sentence like

S15. He picked a wild card.

This is partially explained by more recent results that establish the low value of function words in disambiguating senses with the same part of speech (Yarowsky 1993).

The specific distance rules are placed in a specific distance rule list, and the nonspecific distance rules in a lower priority nonspecific distance rule list. The word expert's rule set, as generated from the first training sentence, is used to try to disambiguate the same target word in a new tagged training sentence. If some rule provides a correct disambiguation, the word expert remains unchanged. If no rule matches then the target word in the new training sentence cannot be disambiguated using the existing rules, and the sentence is used to generate more specific and nonspecific distance rules. Finally, if the existing rules disambiguate the new target occurrence incorrectly, the particular rule responsible for the incorrect disambiguation is transferred from the rule set to a spurious rule set (in one version of the algorithm). This spurious rule set (if present) is scanned whenever a new sentence is used to generate new rules, and any rule on the spurious rule list is not added since it previously led to an incorrect disambiguation. The training continues for additional sentences tagged with the proper disambiguation of the target. Weiss provides no termination criteria.

The spurious rule list technique means a rule which is nearly always useful could be permanently eliminated by one unusual or anomalous sentence. While Weiss found this problem minor, this is at least partly due to the small corpora he used and the consequent low number of exceptions. With larger corpora such as are available now, contamination with exceptions would be a more serious problem, and contrary to Table 5 the spurious rule list version of the algorithm might well be outperformed by the version without a spurious rule list.

7.3.4 Automated Acquisition: Yarowsky

While Weiss addressed resolving ambiguous *meanings,* Yarowsky (1994) addressed resolving ambiguous *accentings* of Spanish and French words from which accents are missing (this paper, Section 6). These disambiguation tasks are similar in that in almost all cases differences in accenting correspond to differences in meaning. However the accenting domain has an advantage in the availability of on-line training and test data. While Weiss's automatic acquisition method requires a training corpus of input sentences that are tediously hand tagged with the meaning of the target word, Yarowsky's method uses a training corpus of already accented text (Figure 1).

Table 5.
Weiss's results for three word expert system architectures, averaged over the words degree, type and volume. Notes: (a) rules were ordered using a predefined heirarchy of rule categories (%-ages calculated from Figure 2 of Weiss (1973)); (b) results taken from Figure 5 of (Weiss 1973); (c) any rule that led to an incorrect result during training was moved to a spurious rule list.

(See note)	Type of word expert system	Recall	Precision
(a)	Rules hand generated	93%	95%
(b)	Rules automatically generated	96%	96%
(b)(c)	Rules automatically generated and filtered	97%	99%

Yarowsky's word expert architecture is similar to Weiss's in that a word expert is a serially searched list of rules. Both Yarowsky and Weiss generate their rules by extracting in straightforward manner, from a given passage tagged with the disambiguation, all rules that are both expressible by their respective rule description method and match the passage.

Yarowsky's word expert architecture differs from Weiss's in the method of ordering the rules. Yarowsky's method measures the reliability of the disambiguating rules, then orders the rules from most to least reliable. When the expert is invoked, the rules are serially checked and the first (and thus most reliable) rule to match is used.

Measuring the reliability of a rule requires a corpus that is tagged with disambiguations: for all passages in the corpus that the rule matches, count how many it disambiguates correctly and how many it doesn't. From those numbers estimate the likelihood that the rule disambiguates correctly. Thus if the RHS of one rule disambiguates the word expert's target correctly 99% of the times its LHS matches a passage, this rule should have a higher precedence than a rule that disambiguates correctly only 98% of the time.[13] The rules are then tested on new material not used for training, and any rule whose disambiguation is usually wrong is deleted since this suggests it may be unreliable no matter what its previous reliability estimate.

In assessing a rule's reliability, should all the passages in the corpus that the rule matches be used, or instead should those passages that match but would actually be processed by a higher priority rule be removed from consideration? For example, one rule might try to match both the preceding and following words, and a second rule might try to match the same preceding word but ignore the following word, so that any passages that would be matched by both rules will be caught by the

[13] The situation is trickier when a rule's probability of disambiguating correctly must be determined on the basis of only a few samples. For example a rule that only matches two passages and works correctly on both should not be estimated to work 100% of the time. Yarowsky used a somewhat *ad hoc* approach to such cases.

highest ranking of the two and the other would never process it. The question then is whether that passage should be used in assessing the disambiguating power of the rule that would never be used on the passage.

Posed this way the question invites the answer that no. Nevertheless, Yarowsky (1994) found that assessing every rule on every passage in the corpus it matches is computationally much cheaper, and the results compare "surprisingle well" to the seemingly better method. Indeed, when passages caught by higher-ranking rules are deleted from the corpus before assessing lower-ranking rules, there might not be enough matching passages left to generate a statistically reliable assessment of some lower-ranking rules, so the seemingly poorer method also has a statistical advantage in addition to the computational advantage.

7.3.5 Automated Acquisition for Word Translation

The following method uses an on-line corpus of parallel texts and, for the languages of the parallel texts, a simple on-line bilingual dictionary listing the possible translations of words in one language into the other. It is similar to an algorithm described by Gale et al. (1993) except in producing word experts rather than Bayesian formulas that incorrectly assume each context word is a source of evidence independent of other context words.

1. Align the parallel corpus so that wherever a word for which a word expert is desired occurs in the source language version, and has a one-word translation in the destination language version, a correspondence is recorded connecting the word and its translation. This requires that the corpus be aligned on sentence boundaries (Gale and Church (1993) give an algorithm for this). Then associate words which have different translations depending on the context with their actual translations in the corpus. Algorithms for finding word correspondence are also described by Brown et al. (1993). Kay and Röscheisen (1993) give an alternative, integrated treatment of both sentence alignment and word correspondence. The actual translations of words provide disambiguation tags and their contexts constitute examples from which disambiguation rules may be derived.
2. Learn word experts from the tagged examples, so that later when a target word appears in a new context, its translation can be inferred. Examples of learning algorithms are described by Weiss (1973) and Yarowsky (1994), and many others might be adapted from the artificial intelligence literature such as case based reasoning (Kolodner 1993), Schaller on Kitano (1993) and advanced decision tree induction algorithms such as derivatives of ID3.

8 Discussion

Lexical disambiguation is an important field with diverse and overlapping approaches. Syntactic pre-processing has traditionally been a critical step in natural language analysis and has been well studied. Word expert style processing has been

combined with rudimentary syntactic pre-processing (Hearst 1991) (see also earlier work of the Rieger-Small approach).

N-gram methods for part of speech tagging are related to the word expert approach, since they disambiguate the part of speech of words by examining local context without syntactic pre-processing. We have not reviewed part of speech tagging here as it is already a well defined area, with actual systems available such as XPOST (Xerox) and a system (University of Birminham) that receives email, tags it, and returns it (those systems are listed in the references herein). However its success is consistent with word expert approaches to lexical disambiguation. Indeed, lexical disambiguation in general often includes a significant component of part of speech disambiguation, and this as true of word expert approaches as it is of others.

Word expert based analysis contrasts with the statistical approaches, which also use word co-occurrence evidence (Allen 1995). Such statistical techniques try to measure the disambiguation evidence provided by a context word from the tendency of that context word to co-occur with a particular disambiguation of the target. After a training stage, when the target word appears in a new passage it is disambiguated using words in its context by combining the evidences provided by those context words. From a practical standpoint this requires assuming those evidences are independent. Yet the independence assumption is a significant weakness since it is usually false.[14] Word expert approaches avoid that problem by relying on one best rule or one best path through a decision tree to disambiguate a given occurrence.

Yarowsky conducted an empirical comparison of word experts vs. the (Bayesian) combination of evidence method using 20 homographs, each with two possible disambiguations. Word experts provided better performance, in that of the 2% of the disambiguations for which the methods disagreed, the word expert method was right and the Bayesian statistical method wrong 65% of the time (Section 4.7 of Yarowsky (1994)). This result was statistically significant for the homographs and sample passages used in the test.

9 Conclusion

Word expert system development has some useful characteristics — modularity, incremental scalability, decomposability in development, and accessibility to relatively untrained persons:

- Word expert knowledge possesses a natural modularity, because different experts process different words. However, if one expert is to handle several morphological variants of a word there may be unintended interactions within a given word expert, and if each variant has its own expert there may be duplication of effort.

[14] In this domain, sources of evidence that support each others' conclusions will usually not be independent, but sources of evidence that support incompatible conclusions may be expected to be independent. Hence combining the single best source of evidence for *each* incompatible conclusion using an independence assumption may be even better, although this has apparently not been tried.

- Thus, word expert based systems are incrementally scalable, by gradually adding new experts. However, the scalability of an individual word expert is no greater than that of any other small expert system.
- Word expert lexicon development is decomposable in that it can be done in parallel by different individuals, who construct experts for different words. For example, Kelly and Stone (1975) on p. 22 report a total of 25 people contributed to their lexicon with 8 contributing heavily.
- Word expert development contrasts with much natural language processing work in that it can and has been done by relatively untrained individuals (Section 7.1).

Automating knowledge acquisition is an important issue when human labor is a bottleneck in knowledge acquisition, as in constructing a sizable word expert lexicon. Appropriate on-line resources can enable automating word expert creation (Section 7.3). While shown for accent restoration (Yarowsky 1994) and sense discrimination (Weiss 1973), this has not yet been demonstrated directly on word translation.

The word expert approach appears easily ported to new domains. Yarowsky ((1994) p. 18) adapted his system from its original domain of disambiguating homograph pronunciations to accent restoration in Spanish with "virtually no modifications in the code," and adapted the Spanish accent restorer to French "in a matter of days."

The various favorable features of word expert based approaches suggest an exciting potential to contribute to society's use of the natural language processing technologies.

Acknowledgements

I thank a number of individuals for comments, suggestions and other contributions, although their endorsement is not implied. They are Joe Adams, Geert Adriaens, Salman Ali, Josh Backon, Hal Berghel, Arnold Berleant, Trey Grubbs, Rob Hartsuiker, Syed Jamil, Cris Jansson, Travis Morris, Alan Munn, Lingyun Shi, Robert F. Simmons (1925–1994), Philip Stone, Prasad Sunkara, and Scott Watts.

References

Adriaens, G., The Parallel Expert Parser: A Meaning-Oriented, Lexically-Guided, Parallel-Interactive Model of Natural Language Understanding, International Workshop on Parsing Technologies, Carnegie-Mellon University, 1989, pp. 309–319.

Adriaens, G., WEP (word expert parsing) revised and applied to Dutch. In: ECAI '86: Proc. of 7th European Conference on Artificial Intelligence, Brighton, U.K., 21-25 July 1986, pp. 222-235. Reprinted in Du Boulay, B., D. Hogg, and L. Steels, eds., Advances in Artificial Intelligence II, Elsevier Science Publishers, 1987, pp. 403–416.

Adriaens, G., Word Expert Parsing: A Natural Language Analysis Program Revised and Applied to Dutch, *Leuvensche Bijdragen* **75** (1) 73–154, 1986.

Adriaens, G. and Small, S. L. Word expert parsing revisited in a cognitive science perspective. In: S.L. Small, G.W. Cottrell, M.K. Tanenhaus (Eds.), *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence,* San Mateo, CA: Morgan Kaufmann, 1988, pp.13-43.

Allen, J., Statistical Word Sense Disambiguation, chapter 10 in *Natural Language Understanding,* 2nd ed., Benjamin/Cummings Publishing Co., 1995.

Armstrong, S., *Using Large Corpora,* MIT Press, 1994.

Barnhart, C. L., ed., *Thorndike-Barnhart Comprehensive Desk Dictionary,* Doubleday and Co., 1967 (pp. XI, 5).

ben Jehiel, N., *Sefer Arukh Ha-Shalem,* Shiloh, 1969.

Berleant, J. D., Subliminal Stimuli and the Interpretation of Ambiguous Sentences, B.S. thesis, MIT, 1982.

Berleant, D., S. Lovelady, and K. Viswanathan. A Foreign Vocabulary Learning Aid for the Networked World of Tomorrow: The LEARN Project. SIGICE Bulletin **19** 3, Feb. 1994, pp. 22-29.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, Word-Sense Disambiguation Using Statistical Methods, *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics* (1991) 264–270.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics* **19** (2) (June 1993) 263–311. Also in Armstrong (1994).

Cerri, S. A., ALICE: Acquisition of Linguistic Items in the Context of Examples, in *Instructional Science* **18** (63-92), 1989.

Choueka, Y. and S. Lusignan, Disambiguation by Short Contexts, *Computers and the Humanities* **19** (1985) 147–157.

Computer Assisted Language Learning (journal), Swets & Zeitlinger.

Consortium for Lexical Research (CLR), World Wide Web location `http://crl.nmsu.edu/clr/CLR.html` .

Cottrell, G. W., *A Connectionist Approach to Word Sense Disambiguation,* Morgan Kaufmann, CA, 1989.

Cottrell, G. and S. Small, A Connectionist Scheme for Modelling Word Sense Disambiguation, *Cognition and Brain Theory 6* (1983) 89-120.

Devos, M. and G. Adriaens, The Parallel Expert Parser, in G. Adriaens and U. Hahn, eds., *Parallel Natural Language Processing,* Ablex, New Jersey, 1994.

Devos, M., Adriaens, G., Willems, Y.D., The parallel expert parser (PEP): a thoroughly revised descendent of the word expert parser (WEP). Budapest. Proc. of the 12th Intl. Conf. on Computational Linguistics (COLING '88) Vol. 1 pp. 142–147, Budapest, 22-27 August, 1988.

Gale, W. A. and K. W. Church, A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics* **19** (1) (March 1993) 75–102. Also in Armstrong (1994).

Gale, W. A., K. W. Church, and D. Yarowsky, A Method for Disambiguating Word Senses in a Large Corpus, *Computers and the Humanities* **26** (1993) 415–439.

Hahn, U. Making understanders out of parsers: Semantically driven parsing as a key concept for realistic text understanding applications, *International Journal of Intelligent Systems* **4** (3) (1989) 345–393.

Hahn, U., An actor model of distributed natural language parsing. In G. Adriaens and U. Hahn, eds., *Parallel Natural Language Processing,* Ablex, New Jersey, 1994.

Harris, B., Bi-text, a New Concept in Translation Theory, *Language Monthly* **54** (1988) 8-10.

Hearst, M., Noun Homograph Disambiguation Using Local Context in Large Text Corpora, in *Using Corpora,* University of Waterloo, Waterloo, 1991.

Hirst, G., *Semantic Interpretation and the Resolution of Ambiguity,* Cambridge University Press, 1987.

Kay, M. and M. Röscheisen, Text-Translation Alignment, *Computational Linguistics* **19** (1) (March 1993) 121–142. Also in Armstrong (1994).

Kelly, E. and P. Stone, *Computer Recognition of English Word Senses,* North-Holland, 1975.

Kolodner, J. L., *Case-Based Reasoning,* Morgan Kaufmann Publishers, 1993.

Linguistic Data Consortium (LDC), World Wide Web location `ftp://ftp.cis.upenn.edu/pub/ldc_www/hpage.html` .

Lorge, I., The English Semantic Count, *Teacher's College Record* **39** (1) 65–77, 1937.

Lorge, I., *A Semantic Count of English Words,* 3 volumes, Institute of Educational Research, Teachers College, Columbia University, New York, 1938 and 1979.

Lorge, I., *The Semantic Count of the 570 Commonest English Words,* Institute of Educational Research, Teachers College, Columbia University, New York, 1949 and 1979.

Mathew, J., J. M. Conrad, and D. Berleant, Word Sense Disambiguation by Constraint Satisfaction: A Feasibility Study, *Proceedings of the Arkansas Computer Conference*, Little Rock, AR, 1993.

The Natural Language Software Registry, on the world wide web at `http://cl-www.dfki.uni-sb.de/cl/registry/draft.html` .

The Oxford English Dictionary Vols. I-XII and Supplement (corrected re-issue), Oxford University Press, 1933.

Oxford Text Archive, Word Wide Web location `ftp://ota.ox.ac.uk/pub/ota/public/dicts/info` .

Rieger, C., Computational Linguistics, in W. Dingwall, ed., *A Survey of Linguistic Science* (Second edition only) 97–134, Greylock, 1978.

Rieger, C., GRIND-1: First Report on the Magic Grinder Story Comprehension Project, *Discourse Processes* **1** (3) (1978a) 267–303.

Rieger, C., Five Aspects of a Full-Scale Story Comprehension Model, in N. Findler, ed., *Associative Networks: Representation and Use of Knowledge by Computers,* Academic Press, NY, 1979.

Rieger, C. and S. Small, Word Expert Parsing, in proceedings, *International Joint Conference on Artificial Intelligence (IJCAI-79)* (1979) 723-728.

Rieger, C. and S. Small, Toward a Theory of Distributed Word Expert Natural Language Parsing, *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-11** (1) (January 1981) 43–51.

Sadler, V., *Working with Analogical Semantics: Disambiguation Techniques in DLT,* Foris Publications, Dordrecht, The Netherlands, 1989.

Schaller, N., Massively Parallel Approach Aids Speech Translation, IEEE's *Computer* **26** (11) (Nov. 1993) p. 78.

Small, S., Parsing as Cooperative Distributed Inference: Understanding through Memory Interactions, in M. King (ed.), *Parsing Natural Language,* London, Academic Press, 1983.

Small, S., Viewing Word Expert Parsing as Linguistic Theory, in proceedings, *International Joint Conference on Artificial Intelligence (IJCAI-81)* (1981) 70-76.

Small, S., Word Expert Parsing: A Theory of Distributed Word-Based Natural Language Understanding, dissertation, technical report TR-954, Department of Computer Science, University of Maryland, College Park, Maryland, 1980.

Small, S., Word Expert Parsing, Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics, published by the Association for Computational Linguistics, 1979.

Small, S., G. Cottrell, and L. Shastri, Toward Connectionist Parsing, in proceedings, *National Conference on Artificial Intelligence (AAAI-82),* American Association for Artificial Intelligence, 1982, 247-250.

Small, S. and C. Rieger, Parsing and Comprehending with Word Experts (A Theory and its Realization), in W. Lehnert and M. Ringle, eds., *Strategies for Natural Language Processing* (pp. 89–147), Lawrence Erlbaum Associates, New Jersey, 1982.

Sproat, R., J. Hirschberg, and D. Yarowsky, A Corpus-Based Synthesizer, in *Proceedings, International Conference on Spoken Language Processing,* 1992.

Stone, P.J., Thematic Text Analysis: New Agendas for Analyzing Text Content, in Carl Roberts, ed., *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts,* Lawrence Erlbaum, Hillsdale, New Jersey.

Stone, P. J., Improved Quality of Content Analysis Categories: Computerized-Disambiguation Rules for High-Frequency English Words, in G. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisley, and P. J. Stone, eds., *The Analysis of Communication Content,* John Wiley and Sons, 1969.

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, with associates, *The General Inquirer: A Computer Approach to Content Analysis* (M.I.T. Press) (1966).

Thorndike, E. L. and C. L. Barnhart, eds., *Thorndike-Barnhart Student Dictionary (Updated Edition),* HarperCollins Publishers, 1993 (p. 14).

University of Birminham, Corpus Linguistics research group's email tagging service at `tagger@clg.bham.ac.uk` (more information at `http://clg1.bham.ac.uk/tagger.html`).

Weiss, S. F., Learning to Disambiguate, *Information Storage and Retrieval* **9** 33-41, 1973.

Wieger, L., *Chinese Characters: Their Origin, Etymology, History, Classification and Signification: A Thorough Study from Chinese Documents,* Dover Publications, 1965.

Xerox Part-of-Speech Tagger (XPOST), `http://cl-www.dfki.uni-sb.de:80/cl/registry/parsers/xerox.html .`

Yarowsky, D., Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94),* World Wide Web location `http://xxx.lanl.gov/abs/cmp-lg/9406034,` 1994.

Yarowsky, D., A Comparison of Corpus-Based Techniques for Restoring Accents in Spanish and French Text, Proceedings, Second Annual Workshop on Very Large Text Corpora, 1994a.

Yarowsky, D., One Sense per Collocation, Proceedings of the 1993 ARPA Human Language Technology Workshop, pp. 266–271, Morgan Kaufmann, CA.