US005541836A

United States Patent [19]

Church et al.

[11] Patent Number: 5,541,836

[45] Date of Patent: Jul. 30, 1996

[54] **WORD DISAMBIGUATION APPARATUS AND METHODS**

[75] Inventors: **Kenneth W. Church**, Chatham; **William A. Gale**, Maplewood; **David E. Yarowsky**, Summit, all of N.J.

[73] Assignee: **AT&T Corp.**, Murray Hill, N.J.

[21] Appl. No.: **814,850**

[22] Filed: **Dec. 30, 1991**

[51] **Int. Cl.**[6] **G06F 19/00**

[52] **U.S. Cl.** **364/419.07**

[58] **Field of Search** 364/419, 419.07, 364/419.08, 419.12, 419.13

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,661,924	4/1987	Okamoto et al.	364/DIG. 2
4,868,750	9/1989	Kucera et al.	364/419
4,914,590	4/1990	Loatman et al.	364/419
4,930,077	5/1990	Fan	364/419
4,942,526	7/1990	Okojima et al.	364/419
5,056,021	10/1991	Ausborn	364/419
5,088,038	2/1992	Tanaka et al.	364/419.08
5,109,509	4/1992	Katayama et al.	395/600
5,128,865	7/1992	Sadler	364/419
5,146,405	9/1992	Church	364/419
5,170,349	12/1992	Yagisawa et al.	364/419
5,237,503	8/1993	Bedecarrax et al.	364/419.08
5,243,520	9/1993	Jacobs et al.	364/419.08
5,317,510	3/1994	Yoshimura et al.	364/419.08

OTHER PUBLICATIONS

IBM Technical Disclosure Bulletin vol. 33 No. 1B Jun. 1990 Armonk US pp. 54–55 "Method for Inferring Lexical Associations from Textual Co–Occurrences".

E. Black, "An Experiment in Computational Discrimination of English Word Senses", *IBM Journal of Research and Development*, vol. 32, No. 2, Mar. 1988, Armonk, NY, USA.

"Word Sense Disambiguation Using an Untagged Corpus", *IBM Technical Disclosure Bulletin*, vol. 35, No. 6, Nov. 1992.

J. D. Benson and B. Brainerd, "Chesterton's Parodies of Swinburne and Yeats: A Lexical Approach", *Literary and Linguistic Computing*, vol. 3, No. 4, 1988.

R. Krovetz, W. Bruce Croft, "Word Sense Disambiguation Using Machine–Readable Dictionaries", Proceedings of the 12th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, Jun. 1989, Cambridge.

Primary Examiner—Donald E. McElheny, Jr.

Attorney, Agent, or Firm—Gordon E. Nelson; Jeffrey M. Weinick

[57] **ABSTRACT**

Apparatus and methods for determining whether a word/sense pair is proper for a context. Wide contexts (100 words) are employed for both training and testing, and testing is done by adding the weights of vocabulary words from the context. The weights are determined by Bayesian techniques which interpolate between the probability of occurrence of a vocabulary word in a conditional sample of the training text and the probability of its occurrence in the entire training text. A further improvement in testing takes advantage of the fact that a word is generally used in only a single sense in a single discourse. Also disclosed are automated training techniques including training on bilingual bodies of text and training using categories from Roget's Thesaurus.

37 Claims, 2 Drawing Sheets

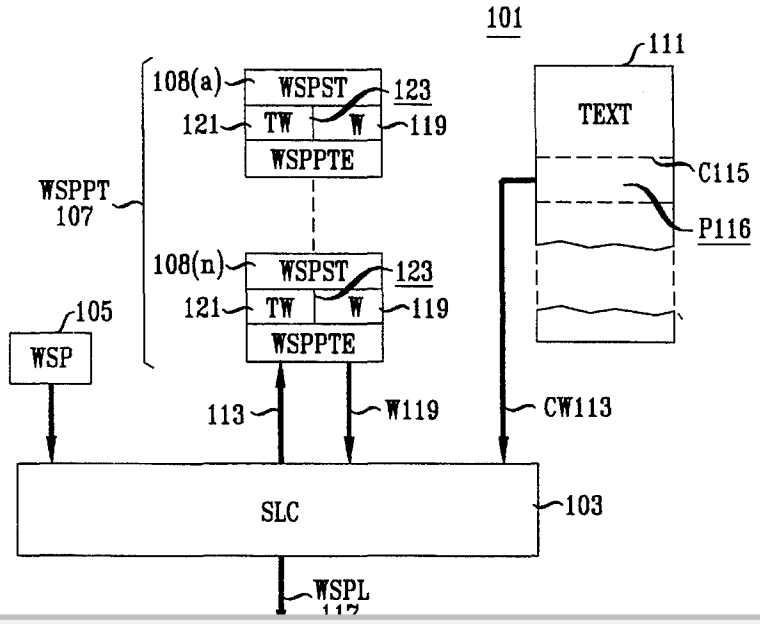


FIG. 1

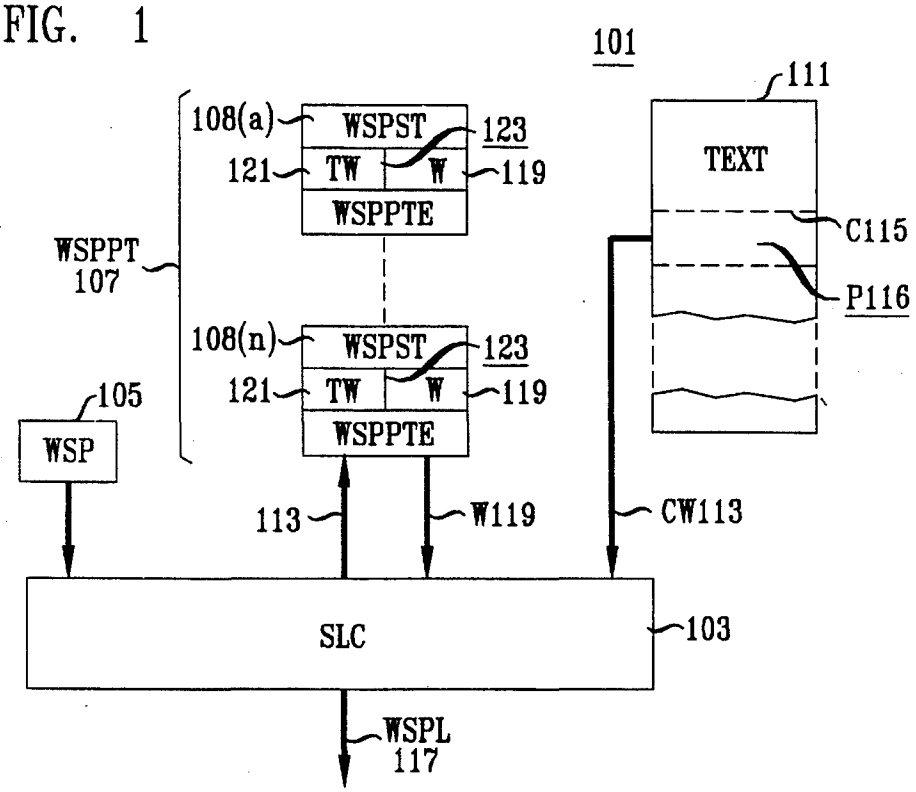


FIG. 2

201

Table 6: Selected Portions of Two Models

tax sense of *duty*				obligation sense of *duty*			
weight*freq	weight	freq	word	weight*freq	weight	freq	word
285	5.7	50	countervailing	64	3.2	20	petitions
111.8	4.3	26	duties	59.28	0.26	228	to
99.9	2.7	37	u.s	56.28	0.42	134	|
73.1	1.7	43	trade	51	3	17	petition
70.2	1.8	39	states	47.6	2.8	17	pursuant
69.3	3.3	21	duty	46.28	0.52	89	mr
68.4	3.6	19	softwood	37.8	2.7	14	honour
68.4	1.9	36	united	37.8	1.4	27	order
58.8	8.4	7	rescinds	36	2	18	present
54	3	18	lumber	33.6	2.8	12	proceedings
50.4	4.2	12	shingles	31.5	3.5	9	prescription
50.4	4.2	12	shakes	31.32	0.87	36	house
46.8	3.6	13	35	29.7	3.3	9	reject
46.2	2.1	22	against	29.4	4.2	7	boundaries
41.8	1.1	38	canadian	28.7	4.1	7	electoral

204 205 207 209

202 203

FIG. 3

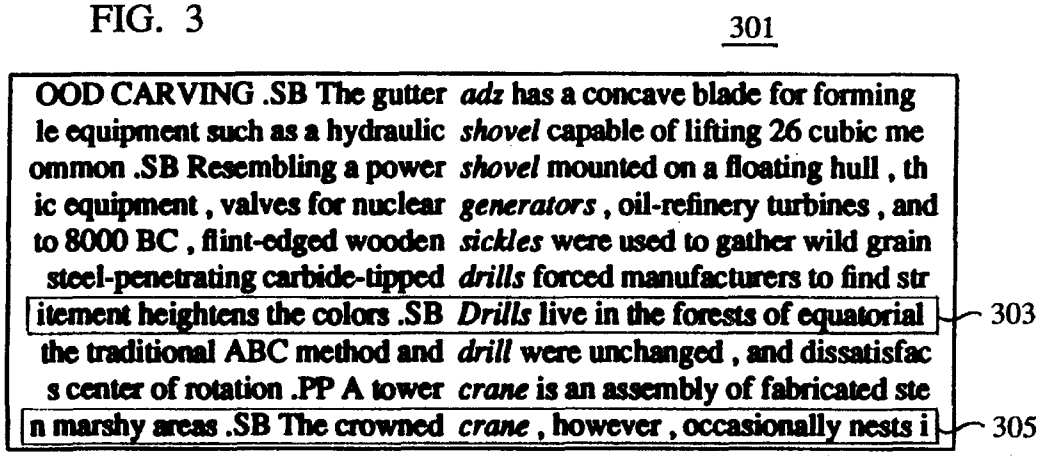
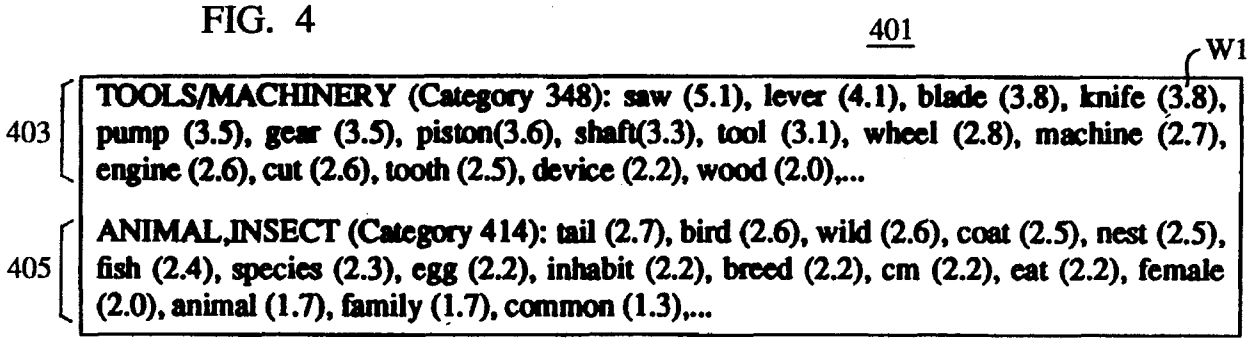


FIG. 4



WORD DISAMBIGUATION APPARATUS AND METHODS

BACKGROUND OF THE INVENTION

1. Field of the Invention

The invention relates to computerized text analysis generally and more specifically to the problem of determining whether a given word-sense pair is proper for a given context.

2. Description of the Prior Art

Machine translation of natural language texts has long been a goal of researchers in computer science and linguistics. A major barrier to high-quality machine translation has been the difficulty of disambiguating words. Word disambiguation is necessary because many words in any natural language have more than one sense. For example, the English noun sentence has two senses in common usage: one relating to grammar, where a sentence is a part of a text or speech, and one relating to punishment, where a sentence is a punishment imposed for a crime. Human beings use the context in which the word appears and their general knowledge of the world to determine which sense is meant, and consequently do not even have trouble with texts such as:

> The teacher gave the student the sentence of writing the sentence "I will not throw spit wads" 100 times.

Computers, however, have no general knowledge of the world, and consequently, have had a great deal of trouble translating sentences such as the above into languages such as French, where the word used to translate sentence when it is employed in the grammatical sense is phrase and the word used to translate sentence when it is employed in the sense of punishment is peine.

The ability to determine a probable sense of a word from the context in which the word is used important in other areas of text analysis as well. For example, optical character recognition systems and speech recognition systems often can only resolve a printed or spoken word into a small set of possibilities; one way of making a choice among the words in the small set is to determine which word has a sense which best fits the context. Other examples in this area are determining whether characters such as accents or umlauts should be present on a word or whether the word should be capitalized. Additionally, there are text editing tools such as spelling checkers or interactive thesauri which present the user with a set of suggested alternatives for a word. These tools, too, are improved if the set of alternatives is limited to words whose senses fit the context.

Another area of text analysis that will benefit from good techniques for determining the probable sense of a word from its context is data base searching. Word searches in data bases work by simply matching a search term with an occurrence of the term in the data base, without regard to the sense in which the term is used in the data base. The only way to restrict a search to a given sense of a term is to provide other search terms which the searcher expects to find in conjunction with the first search term. Such a search strategy will, however, miss occurrences of the first term where the first term has the proper sense but is not found in conjunction with the other search terms. Given a useful way of determining what sense of a word best fits a context, it will be possible to search by specifying not only the search term, but also the sense in which it is being used.

Past researchers have used three different general approaches to the word disambiguation problem sketched

1. Qualitative Methods, e.g., Hirst (1987)
2. Dictionary-based Methods, e.g., Lesk (1986)
3. Corpus-based Methods, e.g., Kelly and Stone (1975)

In each case, the work has been limited by a knowledge acquisition bottleneck. For example, there has been a tradition in parts of the AI community of building large experts by hand, e.g., Granger (1977), Rieger (1977), Small and Rieger (1982), Hirst (1987). Unfortunately, this approach is not very easy to scale up, as many researchers have observed:

> "The expert for THROW is currently six pages long, . . . but it should be 10 times that size." (Small and Reiger, 198X)

Since this approach is so difficult to scale up, much of the work has had to focus on "toy" domains (e.g., Winograd's Blocks World) or sublanguages (e.g., Isabelle (1984), Hirschman (1986)). Currently, it is not possible to find a semantic network with the kind of broad coverage that would be required for unrestricted text.

Others such as Lesk (1986), Walker (1987), Ide (1990, Waterloo Meeting) have turned to machine-readable dictionarys (MRD) such as Oxford's Advanced Learner's Dictionary of Current English (OALDCE) in the hope that MRDs might provide a way out of the knowledge acquisition bottleneck. These researchers seek to develop a program that could read an arbitrary text and tag each word in the text with a pointer to a particular sense number in a particular dictionary. Thus, for example, if Lesk's program was given the phrase pine cone, it ought to tag pine with a pointer to the first sense under pine in OALDCE (a kind of evergreen tree), and it ought to tag cone with a pointer to the third sense under cone in OALDCE (fruit of certain evergreen trees). Lesk's program accomplishes this task by looking for overlaps between the words in the definition and words in the text "near" the ambiguous word.

Unfortunately, the approach doesn't seem to work as well as one might hope. Lesk (1986) reports accuracies of 50–70% on short samples of *Pride and Prejudice*. Part of the problem may be that dictionary definitions are too short to mention all of the collocations (words that are often found in the context of a particular sense of a polysemous word). In addition, dictionaries have much less coverage than one might have expected. Walker (1987) reports that perhaps half of the words occurring in a new text cannot be related to a dictionary entry.

Thus, like the AI approach, the dictionary-based approach is also limited by the knowledge acquisition bottleneck; dictionaries simply don't record enough of the relevant information, and much of the information that is stored in the dictionary is not in a format that computers can easily digest, at least at present.

A third line of research makes use of hand-annotated corpora. Most of these studies are limited by the availability of hand-annotated text. Since it is unlikely that such text will be available in large quantities for most of the polysemous words in the vocabulary, there are serious questions about how such an approach could be scaled up to handle unrestricted text. Kelly and Stone (1975) built 1815 disambiguation models by hand, selecting words with a frequency of at least 20 in a half million word corpus. They started from key word in context (KWIC) concordances for each word, and used these to establish the senses they perceived as useful for content analysis. The models consisted of an ordered set of rules, each giving a sufficient condition for deciding on one classification, or for jumping to another rule

another word. The conditions of a given rule could refer to the context within four words of the target word. They could test the morphology of the target word, an exact context word, or the part of speech or semantic class of any of the context words. The sixteen semantic classes were assigned by hand.

Most subsequent work has sought automatic methods because it is quite labor intensive to construct these rules by hand. Weiss (1973) first built rule sets by hand for five words, then developed automatic procedures for building similar rule sets, which he applied to an additional three words. Unfortunately, the system was tested on the training set, so it is difficult to know how well it actually worked.

Black (1987, 1988) studied five 4-way polysemous words using about 2000 hand tagged concordance lines for each word. Using 1500 training examples for each word, his program constructed decision trees based on the presence or absence of 81 "contextual categories" within the context of the ambiguous word. He used three different types of contextual categories: (1) subject categories from LDOCE, the Longman Dictionary of Contemporary English (Longman, 1978), (2) the 41 vocabulary items occurring most frequently within two words of the ambiguous word, and (3) the 40 vocabulary items excluding function words occurring most frequently in the concordance line. Black found that the dictionary categories produced the weakest performance (47 percent correct), while the other two were quite close at 72 and 75 percent correct, respectively.

There has recently been a flurry of interest in approaches based on hand-annotated corpora. Hearst (1991) is a very recent example of an approach somewhat like Black (1987, 1988), Weiss (1973) and Kelly and Stone (1975), in this respect, though she makes use of considerably more syntactic information than the others did. Her performance also seems to be somewhat better than the others', though it is difficult to compare performance across systems.

As may be seen from the foregoing, the lack of suitable techniques for determining which word-sense pair best fits a given context has been a serious hindrance in many areas of text analysis. It is an object of the apparatus and methods disclosed herein to provide such techniques.

SUMMARY OF THE INVENTION

In one aspect, the invention is a method of automatically determining that a word/sense pair has a sense which suits a given position in a text. The method includes the steps of:

determining a sequence of words in the text which includes the given position and is substantially longer than a single line of the text; and

determining whether the word/sense pair has the suitable sense by automatically analyzing the sequence.

In another aspect, the invention is a method of automatically determining a probability that a word/sense pair has a sense which suits a given position in a text. The method includes the steps of:

determining a sequence of words in the text which includes the given position; and

automatically employing a Bayesian discrimination technique involving the words in the sequence and the sense of the word-sense pair to determine the probability that the word/sense pair has a sense which suits the given position.

In still another aspect, the invention is a method of automatically determining whether a given occurrence of a word

making a first determination of the sense of the given occurrence of the word; and

making a final determination of the sense of the given occurrence of the word by comparing the first determination with a determination of the sense of a neighboring occurrence of the word.

The foregoing and other objects, aspects, and advantages of the invention will be apparent to one of ordinary skill in the art who peruses the following Drawing and Detailed Description, wherein:

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a block diagram of apparatus for determining the probability that a word/sense pair is proper for a context;

FIG. **2** is a table of data from which table **107** of FIG. **1** may be constructed;

FIG. **3** is an example of part of a conditional sample; and

FIG. **4** is an example of weights computed using Roget's categories.

The reference numbers employed in the Drawing and the Detailed Description have three or more digits. The two least significant digits are a number within a figure; the remaining digits are the figure number. Thus, the element with the reference number "**305**" is first shown in FIG. **3**.

DETAILED DESCRIPTION

The following Detailed Description will first provide an overview of the theoretical approach to the disambiguation problem in a preferred embodiment, will then describe apparatus for solving the disambiguation problem, and will finally discuss how the apparatus for solving the disambiguation problem is trained.

BAYESIAN DISAMBIGUATION TECHNIQUES

The word-sense disambiguation problem is a discrimination problem, not very different from problems such as author identification and information retrieval. In author identification and information retrieval, it is customary to split the problem up into a testing phase and a training phase. During the training phase, we are given two (or more) sets of documents and are asked to construct a discriminator which can distinguish between the two (or more) classes of documents. These discriminators are then applied to new documents during the testing phase. In the author identification task, for example, the training set consists of several documents written by each of the two (or more) authors. The resulting discriminator is then tested on documents whose authorship is disputed. In the information retrieval application, the training set consists of a set of one or more relevant documents and a set of zero or more irrelevant documents. The resulting discriminator is then applied to all documents in the library in order to separate the more relevant ones from the less relevant ones. In the sense disambiguation case, the 100-word context surrounding instances of a polysemous word (e.g., duty) are treated very much like a document.

It is natural to take a Bayesian approach to these discrimination problems. Mosteller and Wallace (1964, section 3.1) used the following formula to combine new evidence (e.g., the term by document matrix) with prior evidence (e.g., the historical record) in their classic authorship study of the

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.