Statistical versus symbolic parsing for captioned-information retrieval

Neil C. Rowe

Code CS/Rp, Department of Computer Science
Naval Postgraduate School
Monterey, CA USA 93943
rowe@cs.nps.navy.mil

1. Summary

We discuss implementation issues of MARIE-1, a mostly symbolic parser fully implemented, and MARIE-2, a more statistical parser partially implemented. They address a corpus of 100,000 picture captions. We argue that the mixed approach of MARIE-2 should be better for this corpus because its algorithms (not data) are simpler.

Subject areas: parser implementation, binary word correlation probabilities.

2. Introduction

Our MARIE project has been investigating information retrieval of multimedia data using a new idea: putting primary emphasis on caption processing. Although content analysis methods such as substring searching for text media and shape matching for picture media can obviate captions, content analysis usually requires unacceptably-large amounts of time at retrieval time. Captions can be cachings of the results of content analysis, but they can also include auxiliary information like the date or customer for a photograph. Since captions can be considerably smaller than the media data they describe, checking captions before retrieving media data can save time if it can rule out many bad matches quickly, the idea of "information filters" (Belkin and Croft, 1992).

However, caption processing does not necessarily give faster multimedia retrieval. The terms of the caption are often synonyms or subterms of those supplied by a user during retrieval, so a complete thesaurus of synonyms and a complete type hierarchy of terms should be used during information retrieval (Smith et al, 1989). Furthermore, to obtain high query recall and precision, natural-language processing of the captions must be done to determine the word senses and how the words relate, to get beyond the well-known limits of

keyword matching (Krovetz and Croft, 1992). This additional processing could be slow, so the MARIE project is concerned with methods of improving its efficiency in caption-based retrieval. This paper reports on an important direction that we have explored recently: mixing traditional symbolic parsing with probabilistic ranking based on a restricted kind of statistical information.

While the MARIE project is intended for multimedia information retrieval in general, we have used as testbed the Photo Lab of the Naval Air Warfare Center (NAWC-WD), China Lake California USA. This is a library of approximately 100,000 pictures and 37,000 captions for those pictures. The pictures cover all activities of the center, including pictures of equipment, tests of equipment, administrative documentation, site visits, and public relations. With so many pictures, many of which looking virtually identical, captions are indispensable to find anything. But the existing computerized keyword system for finding pictures from their captions is unhelpful, and is mostly ignored by personnel. (Rowe and Guglielmo, 1993) reports on MARIE-1, a prototype implementation in Prolog that we developed for them, a system that appears much more in the direction of what users want.

But MARIE-1 took a man-year to construct and only handled 220 pictures (averaging 20 words per caption) from the database. To handle the full database, efficiency and implementation-difficulty concerns have become paramount. MARIE-2, currently under development, will address these problems by exploiting a large statistical-correlation database, allowing for simpler parse rules and fewer semantic routines. This should make it run more efficiently while being much easier to apply to the full captions database. This will provide an interesting test of statistical parsing ideas from an engineering standpoint.

3. Example captions

To illustrate the problems posed by the corpus, we present some example captions. All are single-case.

> *an/apq-89 xan-1 radar set in nose of t-2 buckeye modified aircraft bu# 7074, for flight evaluation test. 3/4 overall view of aircraft on runway.*

This is typical of many captions: two noun phrases, each terminated with a period, where the first describes the photographic subject and the second describes the picture itself. Also typical are the complex nominal-compound strings, "an/apq-89 xan-1 radar set" and "t-2 buckeye modified aircraft bu# 7074". Domain knowledge, or statistics as we shall argue, is necessary to recognize "an/apq-89" as a radar type, "xan-1" a version number for that radar, "t-2" an aircraft type, "buckeye" a slang additional name for a T-2, "modified" a conventional adjective, and "bu# 7074" as an aircraft code ID.

program walleye, an/awg-16 fire control pod on a-4c bu# 147781 aircraft, china lake on tail, fit test. 3/4 front overall view and closeup 1/4 front view of pod.

This illustrates some common domain-dependent noun-phrase syntax. "A-4c bu# 147781" is a common pattern of <equipment-type> <prefix-code> <code-number>, a pattern frequent enough to deserve its own grammar rule. Similarly "an/awg-16 fire control pod" is the common pattern of <equipment-name> <equipment-purpose> <equipment-type>, and "3/4 front overall view" is of the form <view-qualifier> <view-qualifier> <view-type>.

graphics presentation tid progress 76. sea site update, wasp head director and hawk screech/sun visor radars. top portion only, excellent.

This illustrates the need for domain-dependent lexicon information. Here "wasp", "hawk", and "sun visor" should not be interpreted in their common English word senses, but as special equipment terms. Furthermore, "progress 76" means "progress in 1976", and "excellent" refers to the quality of the picture. And the "head director" is not a person but a guidance system, and the "sea site" is not in the sea but a dry lakebed flooded with water to a few inches. Such unusual word senses strongly call for inference from domain-dependent statistics. They are also a good argument for natural-language processing for information retrieval instead of keyword matching.

aerial low oblique, looking s from inyodern rd at main gate down china lake bl to bowman rd. on l, b to t, water reservoirs, trf crcl, pw cmpnd, vieweg school, capehart b housing, burroughs hs, cimarron gardens, east r/c old duplex stor. lot. on r, b to t, trngl, bar s motel, arrowsmith, comarco, hosp and on to bowman rd.

This illustrates the problems with the misspellings and nonstandard abbreviations in the captions. "Trf crcl" is supposed to be "traffic circle", "trngl" is triangle, "capehart b" is "capehart base", but "b to t" is "bottom to top". "Vieweg" which looks like a misspelling of "viewed" is actually the correct name of a former base commander, but "inyodern" which looks correct actually is a misspelling of "Inyokern", a nearby town. Such abbreviations and misspellings can only be found by reference to known domain words and using heuristics.

per-heps, parachute extraction rocket-helicopter escape propulsion system, test setup, 700# f in launcher showing 50# deadweight, nylon strap, and parachute cannister.

This illustrates the difficulties of interpreting the numerous acronyms in the captions. Here the first word of the above is an immediately-explained acronym; a careful search for such constructs helps considerably, as often an acronym is explained in at least one

caption. But even explained acronyms cause difficulties. We can generally take the subject of the appositive phrase after the acronym as the type of the acronym, "system" in this case, but how the other words relate to it is complicated and less determined by conventional English syntax than the need to obtain a cute acronym.

4. Our approach to statistical parsing

MARIE-1 uses the standard approach of intelligent natural-language processing for information retrieval (Grosz et al, 1987; Rau, 1988; Sembok and van Rijsbergen, 1990) of hand-coding of lexical and semantic information for the words in a narrow domain. We used the DBG software from Language Systems, Inc. (in Woodland Hills, CA) to help construct the parser for MARIE-1. Nonetheless, considerable additional work was needed to adapt DBG to our domain. Even though we focused on a random sample of only 220 captions, they averaged 50 words in length and required a lexicon and type hierarchy of 1000 additional words beyond the 1000 we could use from the prototype DBG application for cockpit speech. A large number of additional semantic rules had to be written for the many long and complicated noun-noun sequences that had no counterpart in cockpit speech. These required difficult debugging because DBG's multiple-pass semantic processing is tricky to figure out, and the inability of DBG to backtrack and find a second interpretation meant that we could only find a maximum of one bug per run. But hardest of all to use were DBG's syntactic features. These required a grammar with fixed assigned probabilities on each rule, which necessitated a delicate balancing act that considered the entire corpus, to choose what was often a highly sensitive number. The lack of context sensitivity meant that this number had to programmed artificially for each rule to obtain adequate performance (for which some researchers have claimed success), instead of being taken from applicable statistics on the corpus, which makes more sense. But this "programming" was more trial-and-error than anything.

MARIE-1's approach would be unworkable for the 29,538 distinct words in the full 100,000-caption NAWC database. Statistical parsing has emerged in the last few years as an alternative. It assigns probabilities of co-occurrence to sets of words, and uses these probabilities to guess the most likely interpretation of a sentence. The probabilities can be derived from statistics on a corpus, a representative set of example sentences, and they can capture fine semantic distinctions that would otherwise require additional lexicon information.

Statistical parsing is especially well suited for information retrieval because the goal of the latter is to find data that will probably satisfy a user, but satisfaction is never guaranteed. Also, good information retrieval does not require the full natural-language understanding

that hand-tailored semantic routines provide: Understanding of the words matched is not generally helpful beyond their synonym, hierarchical type, and hierarchical part information. For instance, the query "missile mounted on aircraft" should match all three of:

--"sidewinder on f-18"
--"sidewinder attached to wing pylon"
--pylon mounted aim-9m sidewinders"

since "sidewinder" and "aim-9m" are types of missiles, "f-18" as a kind of aircraft, and "on" and "attached" mean the same thing as "mounted". NAWC-WD captions are often imprecise with verbs, so detailed semantic analysis of them is usually fruitless. Parsing is still essential to connect related words in a caption, as to recognize the similar deep structure of the three examples above. But a parser for information retrieval can have fewer grammatical categories and fewer rules than one for full natural-language understanding.

Creating the full synonym list, type hierarchy, and part hierarchy for applications of the size of the NAWC-WD database (42,000 words including words closely related to those in the captions) is considerable work. Fortunately, a large part of this job for any English application has been already accomplished in the Wordnet system (Miller et al, 1990), a large thesaurus system that includes this kind of information, plus rough word frequencies and morphological processing. From Wordnet we obtained information for 6,843 words mentioned in the NAWC-WD captions (for 24,094 word-sense entries), together with 15,417 "alias" facts relating other word senses to 24,094 as synonyms. (The alias facts shortened the lexicon by about 85%.) This left 22,695 words in the captions that did not have available Wordnet data, for which we used a variety of methods to create lexicon entries. The full breakdown of the lexicon was:

Number of distinct words: 29,538
Recognized by Wordnet: 6,843
--Morphological variants of above words: 2,134
--Related superconcepts, wholes, aliases, and phrases recognized by Wordnet: 12,294
--Numbers: 3,718
--Person names: 2,160
--Place names: 246
--Company names: 149
--Words with unambiguous defined-code prefixes: 2,987
--Miscellaneous identifiable special formats: 6,033
--Identifiable misspellings: 826
--Identifiable abbreviations: 928
--Current domain-dependent explicitly-defined words (mostly from MARIE-1): 770
--Current remaining unidentified words (90% equipment names): 4,215

The special-format rules do things like interpret "BU# 462945" as an aircraft identification number and "02/21/93" as a date. Misspellings and abbreviations were obtained mostly automatically, with human checking, from rule-based

systems described in (Rowe and Laitinen, 1994). The effort for lexicon-building, although it is not yet complete, was relatively modest (0.25 of a man-year) thanks to Wordnet, which suggests good portability. Some of this success can be attributed to the restrictions of caption semantics.

We converted all this information to a Quintus Prolog format compatible with the rest of MARIE-2, and used this in parsing and interpretation. The basic meaning assigned to a noun or verb is that it is a subtype of the concept designed by its name in the type hierarchy, with additional pieces of meaning added by its relationships (like modification) to other words in the sentence. For instance, for "big missile on stand", a representative meaning list currently obtained is:

```
[a_kind_of(v3,projectile-1),
property(v3,big-1),
locationover(v3,v5),
a_kind_of(v5,base-2)]
```

where v3 and v5 are variables and the numbers after the hyphen indicate the word sense number.

5. Statistical parsing techniques

This approach can be fast since we just substitute standard synonyms for the words in a sentence, append the type and relationship specifications for all the nouns, verbs, adjectives, and adverbs, and resolve references using the parse tree, to obtain a "meaning list" or semantic graph, following the paradigm of (Covington, 1994) for the nonstatistical aspects. But this can still be slow because it would seem we need to find all the reasonable interpretations of a sentence in order to rank them. To simplify matters, we restricted the grammar to binary parse rules (context-free rules with one or two symbols for the replacement). The likelihood of an interpretation can be found by assigning probabilities to word senses and rules. If we could assume near-independence of the probabilities of each part of the sentence, we could multiply them to get the probability of the whole sentence (Fujisaki et al, 1991). This is mathematically equivalent to taking the sum of the logarithms of the probabilities, and hence a branch-and-bound search could be done to quickly find the N best parses of the a sentence.

But words of sentences are obviously not often independent or near-independent. Statistical parsing often exploits the probabilities of strings of successive words in a sentence (Jones and Eisner, 1992). However, with binary parse rules, a simpler and more semantic idea is to consider only the probability of co-occurrence of the two subparses. For example, the probability of parsing "f-18 landing" by the rule "NP -> NP PARTICIPLEPHRASE" should include the likelihood of an F-18 in particular doing a landing and the likelihood of this syntactic structure. The co-occurrence probability for "f-18" and "land" is especially helpful because it is unexpectedly large, since there are only a few things in the world that land.

Estimates of co-occurrence probabilities can inherit in the type hierarchy (Rowe, 1985). So if we have insufficient statistics in our corpus about how often an F-18 lands, we may enough on how often an aircraft lands; and assuming that F-18s are typical of aircraft in this respect, we can estimate how often F-18s land. The second word can be generalized too, so we can use statistics on "f-18" and "moving", or both the words can be simultaneously generalized, so we can use statistics on "aircraft" and "moving". The idea is to find some statistics that can be reliably used to estimate the co-occurrence probability of the words. Each parse rule can have separate statistics, so the alternative parse of "f-18 landing" by "NP -> ADJECTIVE GERUND" would be evaluated by separate statistics.

To keep this number of possible co-occurrence probabilities manageable, it is important to restrict them to two-probability. When parse rules recognize multiword sequences as grammatical units, those sequences can be reduced to "headwords". For instance, "the big f-18 from china lake landing at armitage field" can also be parsed by "NP -> NP PARTICIPLEPHRASE" and the same co-occurrence probability used, since "f-18" is the principal noun and hence headword of the noun phrase "the big f-18 from china lake", and "landing" is the participle and hence headword of the participial phrase "landing at armitage field". We can get a measure of the interaction of larger numbers of words by multiplying the probabilities for all such binary nodes of the parse tree. This is not an independence assumption anymore because an important word can appear as headword of many different syntactic units, and thus affect the overall rating of a parse in many places.

A big advantage for us of statistical parsing is in identification of unknown words. As we noted earlier, our corpus has many equipment terms, geographical names, and names of people that are not covered by Wordnet. But for information retrieval, detailed understanding of these terms is usually not required beyond recognizing their category, and this can be inferred by co-occurrence probabilities. For instance, in "personnel mounting ghw-12 on an f-18", "ghw-12" must be a piece of equipment because of the high likelihoods of co-occurrence of equipment terms with "mount" and equipment terms with "on".

6. More about the statistical database

We will obtain the necessary counts from running the parser on the 100,000 captions. Using branch-and-bound search, the parser will find what it considers the most likely parse; if this is incorrect, a human monitor will say so and force it to consider the second most likely parse, and so on. Counts are incremented for each binary node in the parse tree, and also for all superconcepts of the words involved. As counts accumulate, the system should gradually become more likely to guess the correct

parse on its first try.

The statistical database for binary co-occurrence statistics will need careful design because the data will be sparse and there will be many small entries. For instance, for the NAWC-WD captions there are about 20,000 synonym sets about which we have lexicon information. This means 200 million possible co-occurrence pairs, but the total of all their counts can only be 610,182, the total number of word instances in all captions. Our counts database uses four search trees indexed on the first word, the part of speech plus word sense of the first word, the second word, and the part of speech plus word sense of the second word. Storing counts rather than probabilities saves storage and reduces work on update. Various compression techniques can further reduce storage, but especially the elimination of data that can be closely approximated from other counts using sampling theory (Rowe, 1985). For instance, if "f-18" occurs 10 times in the corpus, all kinds of aircraft occur 1000 times, and there are 230 occurrences of aircraft landing, estimate the number of "f-18 landing"s in the corpus as 230 * 10 / 1000 = 2.3; if the actual count is within a standard deviation of the value, do not store it in the database. The standard deviation when n is the size of the subpopulation, N is the size of the population, and A the count for the population, is $\sqrt{A(N-A)(N-n)/nN^2(N-1)}$ (Cochran, 1977). Such calculations require also "unary" counts stored with each word or standard phrase, but there are far fewer of these. (While unary counts also directly affect the likelihood of a particular sentence, that effect can be ignored since it is constant over all sentence interpretations.)

We need not store statistics for every word in the statistical database. Many words and phrases used in are corpus are codes that appear rarely, like airplane ID numbers and dates. For such concepts, we only keep statistics on the superconcept, "ID number" and "date" for these examples. Which concepts are to be handled this way is domain-dependent, but generally simple.

7. More about the restriction to binary probabilities

It may seem inadequate to restrict co-occurrence probabilities to pairs of headwords. We argue that while this is inadequate for general natural-language processing, information retrieval in general and captions in particular are minimally affected. That is because the sublanguages of such applications are highly case-oriented, and cases are binary. So structures like subject-verbal-object can be reduced to verbal-subject and verbal-object cases; adjective1-adjective2-noun can be reduced to two adjective-noun case relationships, either separately with each adjective or by reducing adjective1-adjective2 to a composite concept if adjective2 can be taken as a noun. Prepositional phrases would seem to be trouble, however, because the preposi-

tion simultaneously interacts with both its subject and its object. We handle them by subclassifying prepositions as location, time, social, abstract, or miscellaneous, reflecting the main features that affect compatibility with subjects and objects. Then, say, a parse of "aircraft at nawc" retrieves only a high count if the preposition is a location and not time preposition, and so permits compatibility under those syntactically restricted circumstances of "nawc" and "aircraft".

Another objection raised to binary probabilities is the variety of nonlocal semantic relationships the can occur in discourse. Captions at NAWC-WD are usually multi-sentence, and anaphora do occur which can usually be resolved by simple methods. More difficult is the problem of resolving multiple possible word senses. For "sidewinder on ground", are we talking about the snake or the missile? (NAWC-WD captions are all lower case.) The proper interpretation depends on whether the previous sentence was "flora and fauna of the china lake area" (NAWC-WD has many such pictures for public relations) or "loading sequence for a missile". We have three answers to this challenge. First, most of the words that have many multiple meanings in English are abstract or metaphorical, and not appropriate for use on captions. Second, when ambiguous words do occur, the odds are good that some immediate syntactic relationship will provide the necessary clues to resolving ambiguity; for instance, both "sidewinder mounted" and "sidewinder coiled" are unambiguous when using co-occurrence counts. Third, even if multiple interpretations cannot be ruled out for a word, an information retrieval system can just try each, and take the union of the results (i.e. as a logical disjunction); generally only one interpretation will every match a query. Note that if count statistics are derived from the same corpus that is subsequently used for retrieval, as MARIE-2 intends to do, the probabilities obtained from our parse will be a rough estimate of the yield (selectivity) of each interpretation.

8. References

Belkin, N. J. and Croft, W. B. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM, 35*, 12 (December 1992), 29-38.

Cochran, W. G. *Sampling Techniques, third edition.* New York: Wiley, 1977.

Covington, M. *Natural language processing for Prolog programmers.* Englewood Cliffs, NJ: Prentice-Hall, 1994.

Fujisaki, T., Jelinek, F., Cocke, J., Black, E., and Nishino, T. A probabilistic parsing method for sentence disambiguation. In *Current issues in parsing technology*, ed. Tomita, M., Boston: Kluwer, 1991.

Grosz, B., Appelt, D., Martin, P. and Pereira, F. TEAM: An experiment in

the design of transportable natural language interfaces. *Artificial Intelligence, 32* (1987), 173-243.

Jones, M. and Eisner, J. A probablistic parser applied to software testing documents. Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, July 1992, 323-328.

Krovez, R. and Croft, W. B. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems, 10*, 2 (April 1992), 115-141.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. Five papers on Wordnet. *International Journal of Lexicography*, 3, 4 (Winter 1990).

Rau, L. Knowledge organization and access in a conceptual information system. *Information Processing and Management*, 23, 4 (1988), 269-284.

Rowe, N. Antisampling for estimation: an overview. *IEEE Transactions on Software Engineering, SE-11*, 10 (October 1985), 1081-1091.

Rowe, N. Inferring depictions in natural-language captions for efficient access to picture data. *Information Processing and Management, 30*, 3 (1994), 379-388.

Rowe, N. and Guglielmo, E. Exploiting captions in retrieval of multimedia data. *Information Processing and Management, 29*, 4 (1993), 453-461.

Rowe, N. and Laitinen, K. Semiautomatic deabbreviation of technical text. Technical report, Computer Science, Naval Postgraduate School, April 1994.

Sembok, T. and van Rijsbergen, C. SILOL: A simple logical-linguistic document retrieval system. *Information Processing and Management, 26*, 1 (1990), 111-134.

Smeaton, A. F. Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal, 35*, 3 (1992), 268-278.

Acknowledgement: This work was sponsored by DARPA as part of the I3 Project under AO 8939.