ANALYZING TELEGRAPHIC MESSAGES

Ralph Grishman
John Sterling

New York University

Most people have little difficulty reading telegraphic-style messages such as

SHIPMENT GOLD BULLION ARRIVING STAGECOACH JAN. 7 3 PM

even though lots of material has been omitted which would be required in "standard English", such as articles, prepositions, and verbs. Our concern in this paper is how to process such messages by computer. Even though people don't send many telegrams anymore, this problem is still of importance because many military messages are written in this telegraphic style:

2 FLARES SIGHTED 230704Z6 SOUTH APPROX 5 MI SPA ESTABLISHED

(here *230704Z6* is the time, and *SPA* is the Submarine Probability Area).

Alternative Strategies

The particular class of messages which we have studied are a set of Navy tactical messages called RAINFORM (ship) sighting messages [8]. Several other researchers have previously constructed systems to analyze these messages. In the NOMAD system [1] the knowledge was principally realized as procedures associated with individual words. This made it difficult to extend the system, as Granger has noted [1]. Some of the shortcomings of the internal knowledge representation were remedied in a later system named VOX [5] which used a *conceptual grammar*, mixing syntactic and semantic constraints. However, the power of the grammar was still quite limited when compared to grammars traditionally used in computational linguistics applications.

In the development of our system, in contrast, we have taken as our starting point a relatively broad coverage grammar of standard English. More generally, it has been our goal to use, to the extent possible, system components which would be appropriate to a general-purpose English language analyzer. We see several benefits to such an approach:

- Using general-purpose components minimizes the labor in porting the system to a new domain.
- Using a standard English grammar makes it easier to analyze the complex constructions (involving subordinating and coordinating conjunctions, for example) which occur with some frequency in these messages.
- Starting from a standard grammar clarifies the ways in which these messages differ from standard English.

This approach is in keeping with earlier work at NYU, on medical records and equipment failure reports [4,3], and more recent work at UNISYS, primarily on equipment failure reports [6,2].

In the next section, we briefly describe the overall structure of the message understanding system. In the two sections which follow, we focus on the two core problems of analyzing such telegraphic text: first, the problem of analyzing the structure of the text ("parsing"); second, the problem of recovering the arguments which are omitted in the telegraphic text.

System structure

The text processing system is organized as a pipeline consisting of the following modules:

1. A parser using an augmented context-free grammar consisting of context-free rules plus procedural restrictions. The grammar is modeled after the Linguistic String Project English Grammar [7]; the parser is based on a chart parsing algorithm.

2. A syntactic regularizer whose primary function is to convert all clauses into a standard operator-argument form. The regularizer is organized as a set of Montague-style translation rules associated with the individual productions of the parsing grammar.

3. A semantic analyzer which checks semantic class requirements for arguments of verbs, and which translates clauses and nominalizations into domain predicates.

4. Simplification rules, which perform certain simplifications on the output of the semantic analyzer (for example, *conduct an attack* $\Rightarrow$ *attack*).

5. Reference resolution, which resolves anaphoric references.

6. Discourse analysis, which identifies implicit relations between events in the text.

The control structure is not strictly sequential. In particular, the parser, regularizer, and the checking functions of the semantic analyzer are run in parallel. Also, reference resolution and discourse analysis may be interleaved using a priority-based scheme (discussed below).

The entire system has been run successfully on 25 messages drawn from the set of RAINFORM sighting messages in [8]. These messages are, on average, roughly 25 words long.

Analyzing sentence structure

An noted above, we began our work on message analysis with a relatively broad coverage grammar of standard English. Furthermore, we generally followed the approach of Sager and Marsh [4,3] in treating the deviations not as instances of ill-formedness but rather as constructions specific to such telegraphic sublanguages. In our analysis of the RAINFORMs, we found two types of omissions. The first, which had been previously characterized by Sager and Marsh (in their analysis of medical reports and equipment failure messages), involved the omission of top-level sentence elements, such as sentence subjects ("[We] conducted attack at close range.") and the verb "be" ("Results [are] unknown at this time."). The second class can be generally characterized as function words which mark particular cases and types of complements. These include prepositions such as "of" and "at" ("Hydrophone effects [at] bearing [of] 173degt [were] classified [as] surface combatant ..."), "as", and "to" in infinitival strings ("Intend [to] make sweep of area ...").

Modifying the grammar to allow for these omissions was quite straightforward: several definitions were added for sentence fragments, and prepositions, "as", and "to" were allowed to be empty. What made the task less than trivial was *controlling* these omissions. Adding the definitions for sentence fragments alone (following Sager and Marsh) increased syntactic ambiguity, but a sequential analysis (first syntactic analysis, then semantic filtering) was still feasible. However, when the grammar was extended to include function word omission and run-on sentences, the degree of syntactic ambiguity became much greater. If you consider that, in the grammar, each noun can be a sentence fragment or a prepositional phrase (with a deleted preposition), and add the fact that run-on sentences with no punctuation are frequent:

> Sighted periscope an asroc [anti-submarine rocket] fired proceeded on to station visual contact lost, constellation helo hovering in vicinity.

you can imagine the explosion in parses which would occur. Such telegraphic input is understandable, however, only because of the strong semantic clues which are available. We take advantage of these semantic constraints by applying basic semantic checks on the semantic classes of arguments and modifiers each time a noun phrase or a clause is completed during parsing.

In addition, we associate a score with each partial and complete parse, and use a best-first search for parsing. Roughly speaking, we associate a lower score with analyses which imply the existence of a larger number of omitted elements. The scoring mechanism serves to focus the search and thus greatly reduce the parsing time. In addition, it provides a means for preferring one analysis over another in some cases of syntactic ambiguity. For example, the "sentence"

Two cats drinking milk two cats eating fish.

would get, in addition to the analysis as a run-on sentence, *Two cats [are] drinking milk [.] Two cats [are] eating fish.*, the analysis as a single sentence with missing main verb "be", *Two cats [who are] drinking milk [are] two cats [who are] eating fish.*. We have experimented with several scoring schemes; our current scheme exacts a constant penalty for each omitted preposition, "to", and "as", and for each clause (including reduced relative clauses) and sentence fragment in the analysis. This scheme produces the correct analysis for the example just above.

One further modification is required to handle zeroed prepositions. The semantic checks mentioned earlier operate from a set of case frames, one or more for each verb. Each case frame specifies a list of arguments and modifiers, and for each argument or modifier the case marker (such as *subject* or *object* or a list of prepositions) and the semantic class of the argument/modifier. An omitted preposition is marked in the analysis by the symbol *prep* and the semantic checking routine has been modified to accept *prep* in place of a particular preposition (but not to match positional markers such as *subject* or *object*).

Recovering omitted and anaphoric arguments

The second major task in analyzing the telegraphic messages is recoving the missing arguments. In the case frames, certain arguments are marked as *essential*; if they are omitted from the text, reference resolution attempts to fill them in. It does so using essentially the same mechanism employed for anaphora resolution. This commonality of mechanism has been previously noted by UNISYS [6,2].

The basic anaphora resolution mechanism is quite simple, and is based on a hierarchy of semantic classes for the objects and events in the domain. If an argument is omitted, the case frame indicates the semantic class of the argument which was expected. If an argument is present and corresponds to a semantic class more specific than that required by the case frame, we take the semantic class of the argument. Reference resolution searches for the most recently mentioned entity or event of the same semantic class. For example, in analyzing

Fired 2 missiles on Barsuk. Results of attack unknown.

we would recognize firing as a type of attack and thus link *attack* in the second sentence to the event related by the first sentence.

This mechanism is in fact too simple. Component (part/whole) relationships are sometimes needed in order to link anaphor and antecedent. Thus, to resolve *My attacks* in the message

Exchange missile fire with Kynda. ... My attacks successful.

we must recognize that *exchange* involves two activities, my firing at Kynda and Kynda's firing at me. We can then resolve *My attacks* with the first of these activities and thus determine that it was my attacks on Kynda which were successful.

Most of the anaphoric references in these messages can be correctly resolved using this combination of type and component relationships. In some cases, however, we need to make use of richer contextual information, about the relationship of the events in the message to one another. For example, in

Three missiles fired at Kobchic. One missile hit.

reference resolution first uses the general rule that the omitted subject in a sentence fragment is "us" (the ship sending the message), in effect expanding the first sentence to "Three missiles fired [by us] at Kobchic." It is then faced with the problem, in the second sentence, of whom the missiles hit, us or Kobchic, since both antecedents are salient at this point. To resolve this problem we use a set of discourse coherence rules, which capture the cause/effect and precondition/action relationships between the events in the domain. Reference resolution generates the alternate readings, and then discourse analysis scores a reading which matches the coherence rules higher than one which does not. In this case we have a rule that relates firing at a ship with hitting that ship, so the system prefers the analysis where Kobchic was hit.

Both component information and contextual relationships are needed to process

Visual sighting of periscope followed by attack... .

First we fill in "us" as the implicit subject of "sighting". There is no antecedent for *attack*, so we proceed to fill in the essential arguments of *attack*. The object of *attack* must be a ship. The two salient entities at this point are "us" and the periscope. Reference resolution finds a link, through the part-whole hierarchy, between periscope and submarine, a type of ship, so it creates a submarine entity. It then proposes "us" and this submarine as the possible objects of attack. In this domain, we are hunting for enemy ships, so sighting a vessel is typically followed by attacking it. We have included a coherence rule to that effect, so that the "attack on sub" reading is preferred. In other environments, we might flag this passage as ambiguous.

Summary

We have shown how highly telegraphic messages can be analyzed through straightforward extensions of the mechanisms employed for the syntactic and semantic analysis of standard English text.

We have extended previous work on the grammatical analysis of telegraphic messages by allowing for the omission of function words as well as major sentence constituents. This substantially increases syntactic ambiguity, but we have found that this ambiguity can be controlled by applying semantic constraints during parsing and by using a "best-first" parser in which lower scores are associated with analyses which assume omitted function words.

To recover missing arguments from telegraphic text, we have adopted a strategy in which such omitted arguments are treated as anaphoric elements. In order to resolve anaphoric ambiguities, we have extended the anaphora resolution procedure to take account of the implicit causal and enablement relations in the text. We generate alternative resolutions of anaphoric reference and then select the text analysis with the highest "coherence": the analysis for which we can identify the greater number of intersentential relations.

Acknowledgements This research was supported by the Defense Advanced Research Projects Agency under contract N00014-85-K-0163 from the Office of Naval Research. Most of the modifications to the parser required for these messages were programmed and tested by Mahesh Chitrao.

References

[1] Richard H. Granger. The NOMAD system: expectation-based detection and correction of errors during understanding of syntactically and semantically ill-formed text. *American Journal of Computational Linguistics*, 9(3-4):188–196, July-December 1983.

[2] Marcia C. Linebarger, Deborah A. Dahl, Lynette Hirschman, and Rebecca J. Passonneau. Sentence fragments regular structures. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, NY, June 1988.

[3] Elaine Marsh. Utilizing domain-specific information for processing compact text. In *Proceedings of the Conference on Applied Natural Language Processing*, pages 99–103, Santa Monica, CA, February 1983.

[4] Elaine Marsh and Naomi Sager. Analysis and processing of compact texts. In J. Horecky, editor, *COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics*, pages 201–206, North-Holland, Amsterdam, 1982.

[5] A. Meyers. VOX—an extensible natural language processor. In *Proceedings of IJCAI-85*, pages 821–825, Los Angeles, CA, 1985.

[6] Martha S. Palmer, Deborah A. Dahl, Rebecca J. [Schiffman] Passonneau, Lynette Hirschman, Marcia Linebarger, and John Dowding. Recovering implicit information. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Columbia University, New York, August 1986.

[7] Naomi Sager. *Natural Language Information Processing: A Computer Grammar of English and Its Applications.* Addison-Wesley, 1981.

[8] B. M. Sundheim and R. A. Dillard. *Navy Tactical Messages: Examples for Text-Understanding Technology.* Technical Document 1060, Naval Ocean Systems Center, February 1987.