Ambiguity Resolution for Machine Translation of Telegraphic Messages[1]

Young-Suk Lee
Lincoln Laboratory
MIT
Lexington, MA 02173
USA
ysl@sst.ll.mit.edu

Clifford Weinstein
Lincoln Laboratory
MIT
Lexington, MA 02173
USA
cjw@sst.ll.mit.edu

Stephanie Seneff
SLS, LCS
MIT
Cambridge, MA 02139
USA
seneff@lcs.mit.edu

Dinesh Tummala
Lincoln Laboratory
MIT
Lexington, MA 02173
USA
tummala@sst.ll.mit.edu

Abstract

Telegraphic messages with numerous instances of omission pose a new challenge to parsing in that a sentence with omission causes a higher degree of ambiguity than a sentence without omission. Misparsing induced by omissions has a far-reaching consequence in machine translation. Namely, a misparse of the input often leads to a translation into the target language which has incoherent meaning in the given context. This is more frequently the case if the structures of the source and target languages are quite different, as in English and Korean. Thus, the question of how we parse telegraphic messages accurately and efficiently becomes a critical issue in machine translation. In this paper we describe a technical solution for the issue, and present the performance evaluation of a machine translation system on telegraphic messages before and after adopting the proposed solution. The solution lies in a grammar design in which lexicalized grammar rules defined in terms of semantic categories and syntactic rules defined in terms of part-of-speech are utilized together. The proposed grammar achieves a higher parsing coverage without increasing the amount of ambiguity/misparsing when compared with a purely lexicalized semantic grammar, and achieves a lower degree of ambiguity/misparses without decreasing the parsing coverage when compared with a purely syntactic grammar.

1 Introduction

Achieving the goal of producing high quality machine translation output is hindered by lexical and syntactic ambiguity of the input sentences. Lexical ambiguity may be greatly reduced by limiting the domain to be translated. However, the same is not generally true for syntactic ambiguity. In particular, telegraphic messages, such as military operations reports, pose a new challenge to parsing in that frequently occurring ellipses in the corpus induce a higher degree of syntactic ambiguity than for text written in "grammatical" English. Misparsing triggered by the ambiguity of the input sentence often leads to a mistranslation in a machine translation system. Therefore, the issue becomes how to parse telegraphic messages accurately and efficiently to produce high quality translation output.

In general the syntactic ambiguity of an input text may be greatly reduced by introducing semantic categories in the grammar to capture the co-occurrence restrictions of the input string. In addition, ambiguity introduced by omission can be reduced by lexicalizing grammar rules to delimit the lexical items which typically occur in phrases with omission in the given domain. A drawback of this approach, however, is that the grammar coverage is quite low. On the other hand, grammar coverage may be maximized when we rely on syntactic rules defined in terms of part-of-speech at the cost of a high degree of ambiguity. Thus, the goal of maximizing the parsing coverage while minimizing the ambiguity may be achieved by adequately combining lexicalized rules with semantic categories, and non-lexicalized rules with syntactic categories. The question is how much semantic and syntactic information is necessary to achieve such a goal.

In this paper we propose that an adequate amount of lexical information to reduce the ambiguity in general originates from verbs, which provide information on subcategorization, and prepositions, which are critical for PP-attachment ambiguity resolution. For the given domain, lexicalizing domain-specific expressions which typically occur in phrases with omission is adequate for ambiguity resolution. Our experimental results show that the mix of syntactic and semantic grammar as proposed here has advantages over either a syntactic grammar or a lexicalized semantic grammar. Compared with a syntactic grammar, the proposed grammar achieves a much lower degree of ambiguity without decreasing the grammar coverage. Compared with a lexicalized semantic grammar, the proposed grammar achieves a higher rate of parsing coverage without increasing the ambiguity. Furthermore, the generality introduced by the syntactic rules facilitates the porting of the system to other domains as well as enabling the system to handle unknown words efficiently.

This paper is organized as follows. In section 2 we discuss the motivation for lexicalizing grammar rules with semantic categories in the context of translating telegraphic messages, and its drawbacks with respect to parsing coverage. In section 3 we propose a grammar writing technique which minimizes the ambiguity of the input and maximizes the parsing coverage. In section 4 we give our experimental results of the technique on the basis of two sets of unseen test data. In section 5 we discuss system engineering issues to accommodate the proposed technique, i.e., integration of part-of-speech tagger and the adaptation of the understanding system. Finally section 6 provides a summary of the paper.

2 Translation of Telegraphic Messages

Telegraphic messages contain many instances of phrases with omission, cf. (Grishman, 1989), as in (1). This introduces a greater degree of syntactic ambiguities than for texts without any omitted element, thereby posing a new challenge to parsing.

(1)
TU-95 destroyed 220 nm. (≈ An aircraft TU-95 was destroyed at 220 nautical miles)

Syntactic ambiguity and the resultant misparse induced by such an omission often leads to a mistranslation in a machine translation system, such as the one described in (Weinstein et al., 1996), which is depicted in Figure 1.

The system depicted in Figure 1 has a language understanding module TINA, (Seneff, 1992), and a language generation module

[1]This work was sponsored by the Defense Advanced Research Projects Agency. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

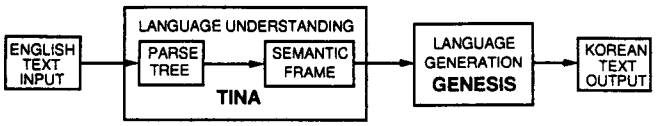

Figure 1: An Interlingua-Based English-to-Korean Machine Translation System

GENESIS, (Glass, Polifroni and Seneff, 1994), at the core. The semantic frame is an intermediate meaning representation which is directly derived from the parse tree and becomes the input to the generation system. The hierarchical structure of the parse tree is preserved in the semantic frame, and therefore a misparse of the input sentence leads to a mistranslation. Suppose that the sentence (1) is misparsed as an active rather than a passive sentence due to the omission of the verb *was*, and that the prepositional phrase *220 nm* is misparsed as the direct object of the verb *destroy*. These instances of misunderstanding are reflected in the semantic frame. Since the semantic frame becomes the input to the generation system, the generation system produces the non-sensical Korean translation output, as in (2), as opposed to the sensible one, as in (3).[3]

(2) TU-95-ka 220 hayli-**lul** pakoy-**hayssta**
TU-95-NOM 220 nautical mile-OBJ destroyed

(3) TU-95-ka 220 hayli-**eyse** pakoy-**toyessta**
TU-95-NOM 220 nautical mile-LOC was destroyed

Given that the generation of the semantic frame from the parse tree, and the generation of the translation output from the semantic frame, are quite straightforward in such a system, and that the flexibility of the semantic frame representation is well suited for multilingual machine translation, it would be more desirable to find a way of reducing the ambiguity of the input text to produce high quality translation output, rather than adjusting the translation process. In the sections below we discuss one such method in terms of grammar design and some of its side effects.x

2.1 Lexicalization of Grammar Rules with Semantic Categories

In the domain of naval operational report messages (MUC-II messages hereafter),[4] (Sundheim, 1989), we find two types of ellipsis. First, top level categories such as subjects and the copula verb *be* are often omitted, as in (4).

(4)
Considered hostile act (= **This was** considered **to be a** hostile act).

Second, many function words like prepositions and articles are omitted. Instances of preposition omission are given in (5), where *z* stands for Greenwich Mean Time (GMT).

(5)
a. Haylor hit by a torpedo and put out of action *8 hours* (= **for** 8 hours)
b. All hostile recon aircraft outbound *1300 z* (= **at** 1300 z)

If we try to parse sentences containing such omissions with the grammar where the rules are defined in terms of syntactic categories (i.e. part-of-speech), the syntactic ambiguity multiplies. To accommodate sentences like (5)a-b, the grammar needs to allow all instances of noun phrases (NP hereafter) to be ambiguous between an NP and a prepositional phrase (PP hereafter) where the preposition is omitted. Allowing an input where the copula verb *be* is omitted in the grammar causes the past tense form of a verb to be interpreted either as the main verb with the appropriate form of *be* omitted, as in (6)a, or as a reduced relative clause modifying the preceding noun, as in (6)b.

(6)
Aircraft launched at 1300 z ...
a. Aircraft were launched at 1300 z ...
b. Aircraft which were launched at 1300 z ...

Such instances of ambiguity are usually resolved on the basis of the semantic information. However, relying on a semantic module for ambiguity resolution implies that the parser needs to produce all possible parses of the input text and carry them along, thereby requiring a more complex understanding process.

One way of reducing the ambiguity at an early stage of processing without relying on a semantic module is to incorporate domain/semantic knowledge into the grammar as follows:

- Lexicalize grammar rules to delimit the lexical items which typically occur in phrases with omission;
- Introduce semantic categories to capture the co-occurrence restrictions of lexical items.

Some example grammar rules instantiating these ideas are given in (7).

(7)
a. .locative_PP
{at in near off on ...} NP
headless_PP
b. .headless_PP
[*at*] np_distance
[*at*] np_bearing
c. .np_distance
numeric nautical_mile
numeric yard
d. .temporal_PP
{during after prior_to ...} NP
time_expression
e. .time_expression
[*at*] numeric gmt
f. .gmt
z

(7)a states that a locative prepositional phrase consists of a subset of prepositions and a noun phrase. In addition, there is a subcategory *headless_PP* which consists of a subset of noun phrases which typically occur in a locative prepositional phrase with the preposition omitted. The head nouns which typically occur in prepositional phrases with the preposition omission are *nautical miles* and *yard*. The rest of the rules can be read in a similar manner. And it is clear how such lexicalized rules with the semantic categories reduce the syntactic ambiguity of the input text.

2.2 Drawbacks

Whereas the language processing is very efficient when a system relies on a lexicalized semantic grammar, there are some drawbacks as well.

- Since the grammar is domain and word specific, it is not easily ported to new constructions and new domains.
- Since the vocabulary items are entered in the grammar as part of lexicalized grammar rules, if an input sentence contains words unknown to the grammar, parsing fails.

These drawbacks are reflected in the performance evaluation of our machine translation system. After the system was developed on all the training data of the MUC-II corpus (640 sentences, 12 words/sentence average), the system was evaluated on the held-out test set of 111 sentences (hereafter TEST set). The results are shown in Table 1. The system was also evaluated on the data which were collected from an in-house experiment. For this experiment, the subjects were asked to study a number of MUC-II sentences, and create about 20 MUC-II-like sentences. These

[3] In the examples, *NOM* stands for the nominative case marker, *OBJ* the object case marker, and *LOC* the locative postposition.

[4] MUC-II stands for the Second Message Understanding Conference. MUC-II messages were originally collected and prepared by NRaD(1989) to support DARPA-sponsored research in message understanding.

Total No. of sentences	111
No. of sentences with no unknown words	66/111 (59.5%)
No. of parsed sentences	23/66 (34.8%)
No. of misparsed sentences	2/23 (8.7%)

Table 1: TEST Data Evaluation Results on the Lexicalized Semantic Grammar

Total No. of sentences	281
No. of sentences with no unknown words	239/281 (85.1%)
No. of parsed sentences	103/239 (43.1%)
No. of misparsed sentences	15/103 (14.6%)

Table 2: TEST' Data Evaluation Results on the Lexicalized Semantic Grammar

MUC-II-like sentences form data set TEST'. The results of the system evaluation on the data set TEST' are given in Table 2.

Table 1 shows that the grammar coverage for unseen data is about 35%, excluding the failures due to unknown words. Table 2 indicates that even for sentences constructed to be similar to the training data, the grammar coverage is about 43%, again excluding the parsing failures due to unknown words. The misparse[5] rate with respect to the total parsed sentences ranges between 8.7% and 14.6%, which is considered to be highly accurate.

3 Incorporation of Syntactic Knowledge

Considering the low parsing coverage of a semantic grammar which relies on domain specific knowledge, and the fact that the successful parsing of the input sentence is a prerequisite for producing translation output, it is critical to improve the parsing coverage. Such a goal may be achieved by incorporating syntactic rules into the grammar while retaining lexical/semantic information to minimize the ambiguity of the input text. The question is: how much semantic and syntactic information is necessary? We propose a solution, as in (8):

(8)
(a) Rules involving verbs and prepositions need to be lexicalized to resolve the prepositional phrase attachment ambiguity, cf. (Brill and Resnik, 1993).
(b) Rules involving verbs need to be lexicalized to prevent misparsing due to an incorrect subcategorization.
(c) Domain specific expressions (e.g. *z. nm* in the MUC-II corpus) which frequently occur in phrases with omitted elements, need to be lexicalized.
(d) Otherwise, rely on syntactic rules defined in terms of part-of-speech.

In this section, we discuss typical misparses for the syntactic grammar on experiments in the MUC-II corpus. We then illustrate how these misparses are corrected by lexicalizing the grammar rules for verbs, prepositions, and some domain-specific phrases.

3.1 Typical Misparses Caused by Syntactic Grammar

The misparses we find in the MUC-II corpus, when tested on a syntactic grammar, are largely due to the three factors specified in (9).

(9) i. Misparsing due to prepositional phrase attachment (hereafter PP-attachment) ambiguity
ii. Misparsing due to incorrect verb subcategorizations
iii. Misparsing due to the omission of a preposition, e.g. *1410 z* instead of *at 1410 z*

Examples of misparses due to an incorrect verb subcategorization and a PP-attachment ambiguity are given in Figure 2 and Figure 3, respectively. An example of a misparse due to preposition omission is given in Figure 4.
In Figure 2, the verb *intercepted* incorrectly subcategorizes for a finite complement clause.
In Figure 3, the prepositional phrase *with 12 rounds* is wrongly attached to the noun phrase *the contact*, as opposed to the verb phrase *vp_active*, to which it properly belongs.
Figure 4 shows that the prepositional phrase *1410 z* with *at* omitted is misparsed as a part of the noun phrase expression *hostile raid composition*.

3.2 Correcting Misparses by Lexicalizing Verbs, Prepositions, and Domain Specific Phrases

Providing the accurate subcategorization frame for the verb *intercept* by lexicalizing the higher level category 'vp' ensures that it never takes a finite clause as its complement, leading to the correct parse, as in Figure 5.

As for PP-attachment ambiguity, lexicalization of verbs and prepositions helps in identifying the proper attachment site of the prepositional phrase, cf. (Brill and Resnik, 1993), as illustrated in Figure 6.

Misparses due to omission are easily corrected by deploying lexicalized rules for the vocabulary items which occur in phrases with omitted elements. For the misparse illustrated in Figure 3, utilizing the lexicalized rules in (10) prevents *1410 z* from being analyzed as part of the subsequent noun phrase, as in Figure 7.

(10) a. .time_expression
[*at*] numeric gmt

b. .gmt
z

4 Experimental Results

In this section we report two types of experimental results. One is the parsing results on two sets of unseen data TEST and TEST' (discussed in Section 2) using the syntactic grammar defined purely in terms of part-of-speech. The other is the parsing results on the same sets of data using the grammar which combines lexicalized semantic grammar rules and syntactic grammar rules. The results are compared with respect to the parsing coverage and the misparse rate. These experimental results are also compared with the parsing results with respect to the lexicalized semantic grammar discussed in Section 2.

4.1 Experimental Results on Data Set TEST

Total No. of sentences	111
No. of parsed sentences	84/111 (75.7%)
No. of misparsed sentences	24/84 (29%)

Table 3: TEST Data Evaluation Results on the Syntactic Grammar

Total No. of sentences	111
No. of parsed sentences	86/111 (77%)
No. of misparsed sentences	9/86 (10%)

Table 4: TEST Data Evaluation Results on the Mixed Grammar

In terms of parsing coverage, the two grammars perform equally well (around 76%). In terms of misparse rate, however, the grammar which utilizes only syntactic categories shows a much higher

[5]The term *misparse* in this paper should be interpreted with care. A number of the sentences we consider to be misparses are not "syntactic" misparses, but "semantically anomalous." Since we are interested in getting the accurate interpretation in the given context at the parsing stage, we consider parses which are semantically anomalous to be misparses.

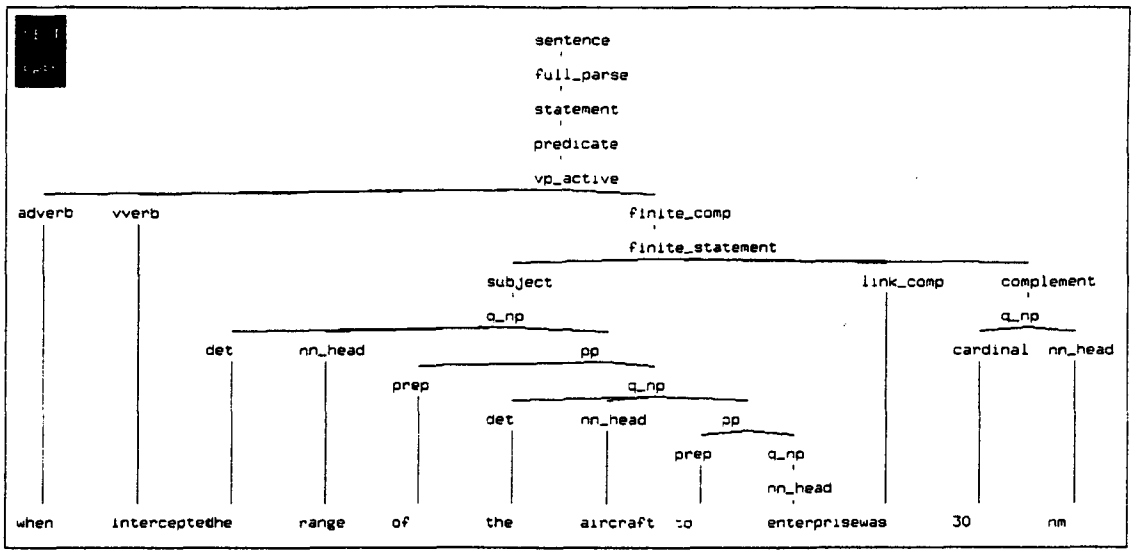

Figure 2: Misparse due to incorrect verb subcategorization

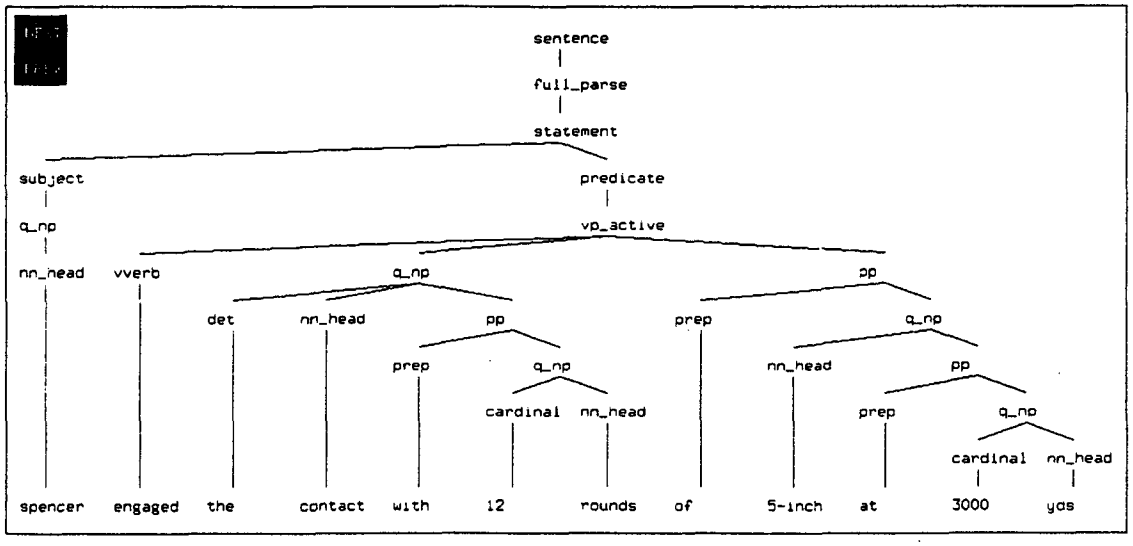

Figure 3: Misparse due to PP-attachment ambiguity

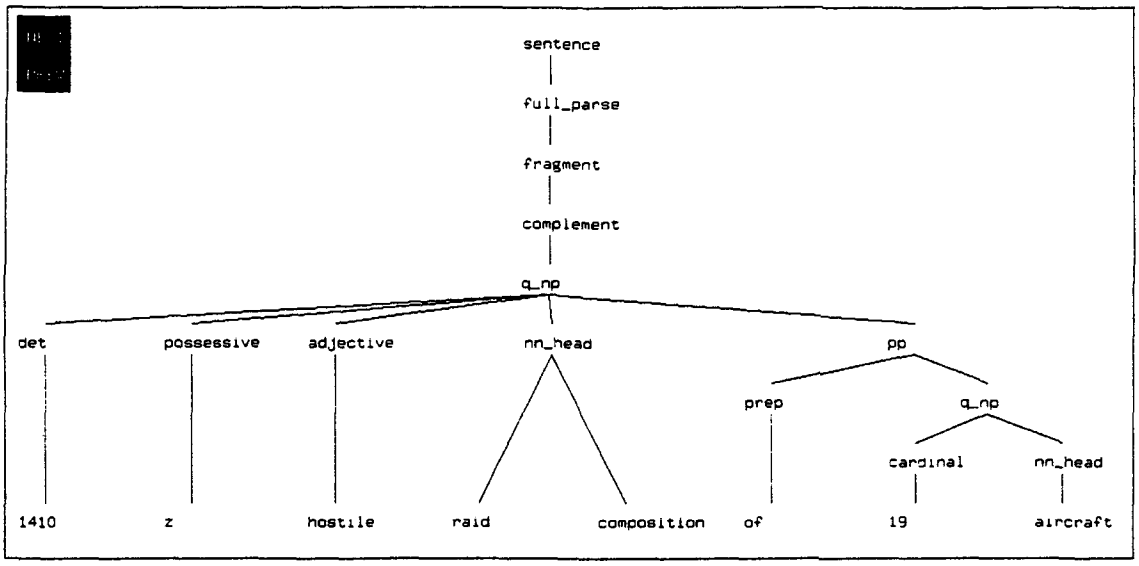

Figure 4: Misparse due to Omission of Preposition

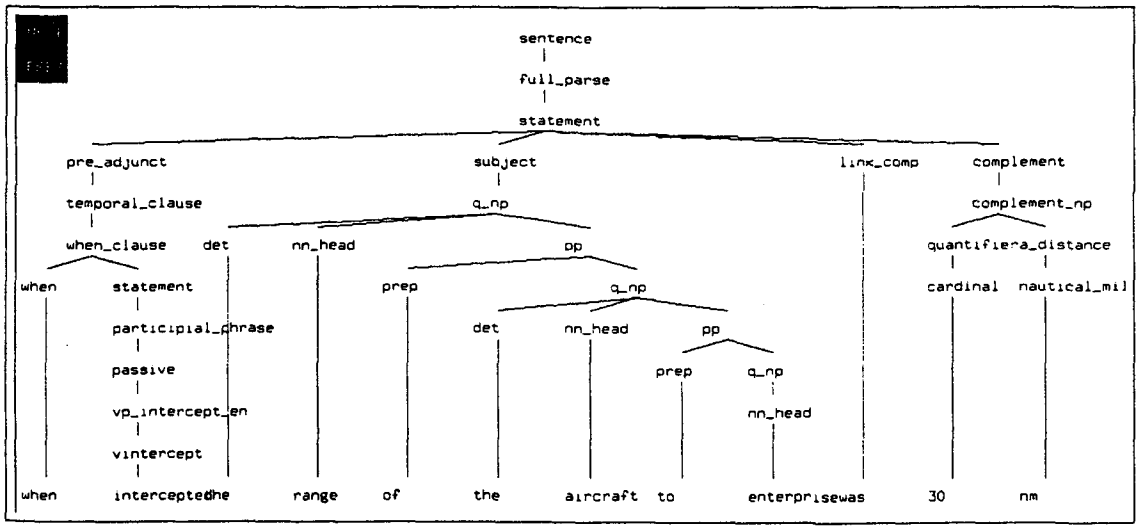

Figure 5: Parse Tree with Correct Verb Subcategorization

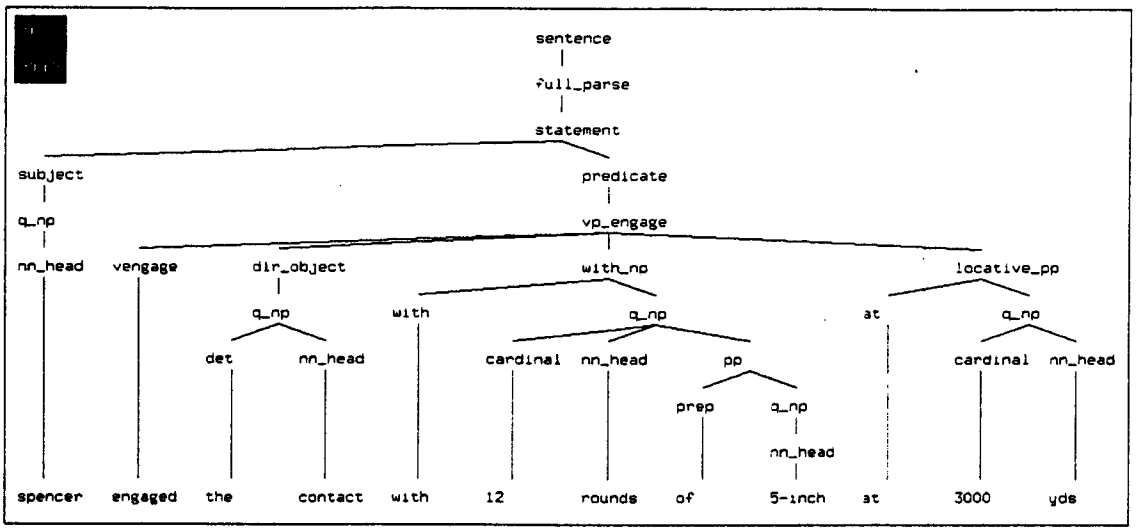

Figure 6: Parse Tree with Correct PP-attachment

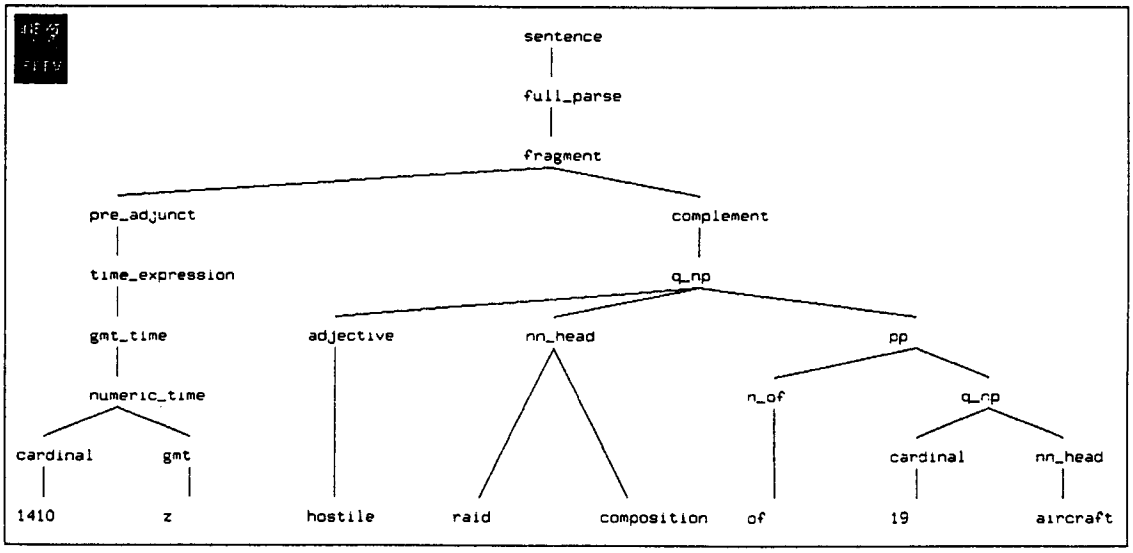

Figure 7: Corrected Parse Tree

rate of misparse (i.e. 29%) than the grammar which utilizes both syntactic and semantic categories (i.e. 10%). Comparing the evaluation results on the mixed grammar with those on the lexicalized semantic grammar discussed in Section 2, the parsing coverage of the mixed grammar is much higher (77%) than that of the semantic grammar (59.5%). In terms of misparse rate, both grammars perform equally well, i.e. around 9%.[6]

4.2 Experimental Results on Data Set TEST'

Total No. of sentences	281
No. of sentences which parse	215/281 (76.5%)
No. of misparsed sentences	60/215 (28%)

Table 5: TEST' Data Evaluation Results on Syntactic Grammar

Total No. of sentences	289
No. of parsed sentences	236/289 (82%)
No. of misparsed sentences	23/236 (10%)

Table 6: TEST' Data Evaluation Results on Mixed Grammar

Evaluation results of the two types of grammar on the TEST' data, given in Table 5 and Table 6, are similar to those of the two types of grammar on the TEST data discussed above.

To summarize, the grammar which combines syntactic rules and lexicalized semantic rules fares better than the syntactic grammar or the semantic grammar. Compared with a lexicalized semantic grammar, this grammar achieves a higher parsing coverage without increasing the amount of ambiguity/misparsing. When compared with a syntactic grammar, this grammar achieves a lower degree of ambiguity/misparsing without decreasing the parsing rate.

5 System Engineering

An input to the parser driven by a grammar which utilizes both syntactic and lexicalized semantic rules consists of words (to be covered by lexicalized semantic rules) and parts-of-speech (to be covered by syntactic rules). To accommodate the part-of-speech input to the parser, the input sentence has to be part-of-speech tagged before parsing. To produce an adequate translation output from the input containing parts-of-speech, there has to be a mechanism by which parts-of-speech are used for parsing purposes, and the corresponding lexical items are used for the semantic frame representation.

5.1 Integration of Rule-Based Part-of-Speech Tagger

To accommodate the part-of-speech input to the parser, we have integrated the rule-based part-of-speech tagger, (Brill, 1992), (Brill, 1995), as a preprocessor to the language understanding system TINA, as in Figure 8. An advantage of integrating a part-of-speech tagger over a lexicon containing part-of-speech information is that only the former can tag words which are new to the system, and provides a way of handling unknown words.

While most stochastic taggers require a large amount of training data to achieve high rates of tagging accuracy, the rule-based tagger achieves performance comparable to or higher than that of stochastic taggers, even with a training corpus of a modest size. Given that the size of our training corpus is fairly small (total 7716 words), a transformation-based tagger is well suited to our needs.

[6]The parsing coverage of the semantic grammar, i.e. 34.8%, is after discounting the parsing failure due to words unknown to the grammar. The reason why we do not give the statistics of the parsing failure due to unknown words for the syntactic and the mixed grammar is because the part-of-speech tagging process, which will be discussed in detail in Section 5, has the effect of handling unknown words, and therefore the problem does not arise.

Figure 8: Integration of the Rule-Based Part-of-Speech Tagger as a Preprocessor to the Language Understanding System

The transformation-based part-of-speech tagger operates in two stages. Each word in the tagged training corpus has an entry in the lexicon consisting of a partially ordered list of tags, indicating the most likely tag for that word, and all other tags seen with that word (in no particular order). Every word is first assigned its most likely tag in isolation. Unknown words are first assumed to be nouns, and then cues based upon prefixes, suffixes, infixes, and adjacent word co-occurrences are used to upgrade the most likely tag. Secondly, after the most likely tag for each word is assigned, contextual transformations are used to improve the accuracy.

We have evaluated the tagger performance on the TEST Data both before and after training on the MUC-II corpus. The results are given in Table 7. Tagging statistics 'before training' are based on the lexicon and rules acquired from the BROWN CORPUS and the WALL STREET JOURNAL CORPUS. Tagging statistics 'after training' are divided into two categories, both of which are based on the rules acquired from training data sets of the MUC-II corpus. The only difference between the two is that in one case (After Training I) we use a lexicon acquired from the MUC-II corpus, and in the other case (After Training II) we use a lexicon acquired from a combination of the BROWN CORPUS, the WALL STREET JOURNAL CORPUS, and the MUC-II database.

Training Status	Tagging Accuracy
Before Training	1125/1287 (87.4%)
After Training I	1249/1287 (97%)
After Training II	1263/1287 (98%)

Table 7: Tagger Evaluation on Data Set TEST

Table 7 shows that the tagger achieves a tagging accuracy of up to 98% after training and using the combined lexicon, with an accuracy for unknown words ranging from 82 to 87%. These high rates of tagging accuracy are largely due to two factors: (1) Combination of domain specific contextual rules obtained by training the MUC-II corpus with general contextual rules obtained by training the WSJ corpus; And (2) Combination of the MUC-II lexicon with the lexicon for the WSJ corpus.

5.2 Adaptation of the Understanding System

The understanding system depicted in Figure 1 derives the semantic frame representation directly from the parse tree. The terminal symbols (i.e. words in general) in the parse tree are represented as vocabulary items in the semantic frame. Once we allow the parser to take part-of-speech as the input, the parts-of-speech (rather than actual words) will appear as the terminal symbols in the parse tree, and hence as the vocabulary items in the semantic frame representation. We adapted the system so that the part-of-speech tags are used for parsing, but are replaced with the original words in the final semantic frame. Generation can then proceed as usual. Figures 9 and (11) illustrate the parse tree and semantic frame produced by the adapted system for the input sentence *0819 z unknown contacts replied incorrectly.*

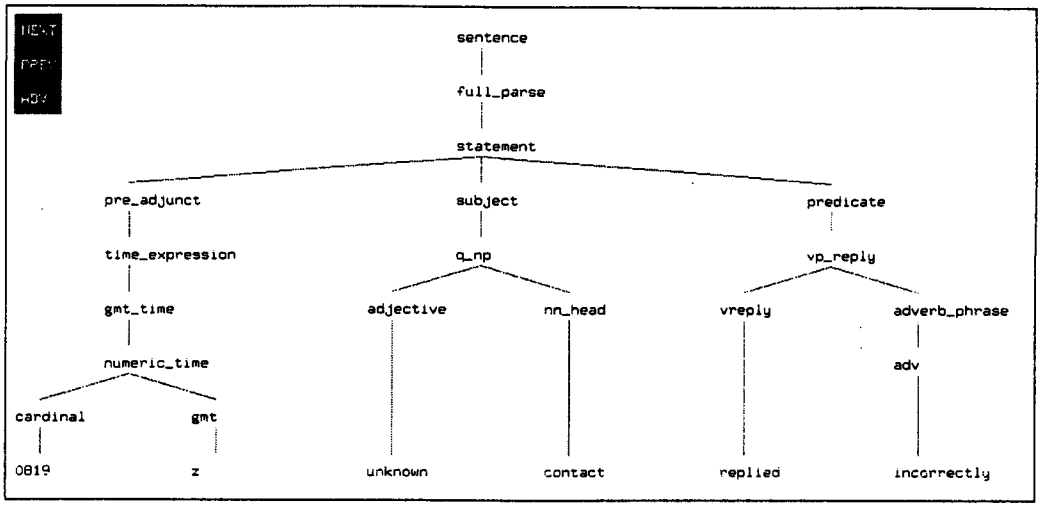

Figure 9: Parse Tree Based on the Mix of Word and Part-of-Speech Sequence

(11)

```
{c statement
   :time_expression {p numeric_time
                       :topic {q gmt
                                 :name "z" }
                       :pred {p cardinal
                                :topic "0819" } }
   :topic {q nn_head
             :name "contact"
             :pred {p unknown
                      :global 1 } }
   :subject 1
   :pred {p reply_v
            :mode "past"
            :adverb {p incorrectly }  } }
```

6 Summary

In this paper we have proposed a technique which maximizes the parsing coverage and minimizes the misparse rate for machine translation of telegraphic messages. The key to the technique is to adequately mix semantic and syntactic rules in the grammar. We have given experimental results of the proposed grammar, and compared them with the experimental results of a syntactic grammar and a semantic grammar with respect to parsing coverage and misparse rate, which are summarized in Table 8 and Table 9. We have also discussed the system adaptation to accommodate the proposed technique.

Grammar Type	Parsing Rate	Misparse Rate
Semantic Grammar	34.8%	8.7%
Syntactic Grammar	75.7%	29%
Mixed Grammar	77%	10%

Table 8: TEST Data Evaluation Results on the Three Types of Grammar

Grammar Type	Parsing Rate	Misparse Rate
Semantic Grammar	43.1%	14.6%
Syntactic Grammar	76.5%	28%
Mixed Grammar	82%	10%

Table 9: TEST' Data Evaluation Results on the Three Types of Grammar

References

Eric Brill. 1992. A Simple Rule-Based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing, ACL, Trento, Italy.*

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21-4, pages 543-565.

Eric Brill and Philip Resnik. 1993 A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation. Technical report, Department of Computer and Information Science, University of Pennsylvania.

James Glass, Joseph Polifroni and Stephanie Seneff. 1994. Multilingual Language Generation Across Multiple Domains. Presented at the 1994 International Conference on Spoken. Language Processing, Yokohama, Japan.

Ralph Grishman. 1989. Analyzing Telegraphic Messages. *Proceedings of Speech and Natural Language Workshop,* DARPA.

Stephanie Seneff. 1992. TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics*, 18:1, pages 61-88.

Beth M. Sundheim. Navy Tactical Incident Reporting in a Highly Constrained Sublanguage: Examples and Analysis. Technical Document 1477, Naval Ocean Systems Center, San Diego.

Clifford Weinstein, Dinesh Tummala, Young-Suk Lee, Stephanie Seneff. 1996. Automatic Engish-to-Korean Text Translation of Telegraphic Messages in a Limited Domain. To be presented at the International Conference on Computational Linguistics '96.