NEW YORK UNIVERSITY PROTEUS SYSTEM: MUC-4 TEST RESULTS AND ANALYSIS

Ralph Grishman, John Sterling, and Catherine Macleod

The PROTEUS Project
Computer Science Department
New York University
715 Broadway, 7th Floor
New York, NY 10003

{grishman,sterling,macleod}@cs.nyu.edu

RESULTS

The "ALL TEMPLATES" results of our "official" runs were as follows:

	RECALL	PRECISION
TST3	41	47
TST4	46	46

Evaluating the degree of improvement over the MUC-3 runs is complicated by the changes between MUC-3 and MUC-4: there were changes in the template structure, the MURDER templates were eliminated, content mapping constraints were incorporated into the scoring program, and the rules for manual remapping were much more constrained. We resumed system development specifically for MUC (with regular runs and rescorings) in mid-March, approximately two weeks before the "Dry Run" was due, and the modifications prior to the Dry Run primarily reflected the changes needed for the new template structure (no significant changes were made to concepts, verb models, inference rules, etc.). The changes between our final MUC-3 scores and our Dry Run scores thus roughly reflect the changes due to the change in the task -- for both TST1 and TST2, a loss of about 7 points of recall. During the following 8 weeks, we made a number of system modifications which recovered much of this loss of recall and substantially improved system precision.

	TST1 RECALL	TST1 PRECISION	TST2 RECALL	TST2 PRECISION
May 91	56	41	44	36
March 92	49	38	37	35
May 92	57	54	40	45

During the period from mid-March, when we adapted the system for the MUC-4 templates and began scoring runs, until the evaluation at the end of May, approximately 5 to 6 person-months were involved in development specifically addressed to MUC-4 performance. This does not count the time we spent since MUC-3 on research using the MUC-3 data, on such topics as semantic pattern acquisition, Wordnet, and grammar evaluation; most of this work was not directly used in the MUC-4 system.

IMPROVEMENTS

We made a number of small improvements in upgrading our MUC-3 system for the MUC-4 evaluation:

(1) We integrated the BBN stochastic part-of-speech tagger into our system. We had done this for MUC-3, but in a rather crude way, keeping only the most probable part-of-speech assigned by the tagger. This made the system run faster, but with some loss of recall. For MUC-4, we made full use of the probabilities assigned by the tagger, combining them with the other contributions to our scoring function (e.g., semantic scores, syntactic penalties) and selecting the highest-scoring analysis. This yielded a small improvement in system recall (1% on the TST1 corpus).

(2) We incorporated a more elaborate time analysis component to handle constructs such as "Three weeks later ..." and "Two weeks after <event 1>, <event 2> ...", in addition to the absolute times (explicit dates) and times relative to the dateline ("two weeks ago") which were handled in our MUC-3 system. The system now produces a time graph relating events, and computes absolute times as the information becomes available. This produced a small benefit in recall and precision.

(3) In our MUC-3 system, if no parse could be obtained of the entire sentence, we identified the longest string starting at the first word which could be analyzed as a sentence. We now have the option of taking the remaining words, identifying the longest clauses and noun phrases, and processing these (in addition to the longest initial substring). We refer to this as "syntactic debris". Because most sentences obtain a full-sentence parse, this option has only a small effect. On TST3, selecting "syntactic debris" increased recall by 1% and reduced precision by 1%.

(4) We implemented a simple mechanism for dynamically shifting the parsing strategy. For each sentence, up to a certain point, all hypotheses are followed, in a best-first order determined by our scoring function. Once a specified number of hypotheses have been generated (15000 in the official runs), we shift to a mode where only the highest-ranking hypothesis for each non-terminal and each span of sentence words is retained. This mode may yield a sub-optimal analysis (because many constraints are non-local), but will converge to some analyisis much more quickly (effectively shifting from an exponential to a polynomial-time algorithm).

(5) We made several improvements to reference resolution. In particular, we refined the semantic slot/filler representation we use for people in order to improve anaphor-antecedent matching.

(6) We have been steadily expanding our grammatical coverage.

Except as needed for our other system changes, we made relatively few additions to the sets of concepts and lexical models developed for MUC-3.[1] We did not extend the effort at extensive corpus analysis pursued prior to MUC-3; rather we experimented with various strategies which would lead to greater automation of this process in the future (see the sections below on "Wordnet" and "Acquiring Selectional Constraints").

DISCOURSE

At MUC-3, discourse analysis was frequently cited as a serious shortcoming of many of the systems. In our system, discourse analysis (beyond reference resolution) is reflected mainly in decisions about merging events to form templates. Roughly speaking, our MUC-3 system tried to merge events (barring conflicting time, location, etc.)

- when they affected the same target
- when they appeared in the same sentence
- when an attack (including bombing, arson, etc.) was followed by effect (death, damage, injury, etc.)

For MUC-4 we tried 3 variations on our discourse analysis procedure:

(1) blocking attack/effect merging across paragraph boundaries

(2) in addition, making use of anaphoric references to events in the merging procedure (so that "Five civilians were killed in the attack." would cause the templates for the attack and the killings to be merged even if the antecedent of "attack" were in a prior paragraph).

(3) identifying and attempting to merge general and specific descriptions of events (this happens quite often in newspaper-style articles, where the introductory paragraph is a summary of several distinct events which are reported separately later in the article). This linking of general and specific events was then used by reference resolution to order the search for antecedents. (This can be viewed as an attempt at a Grosz/Sidner focus stack.)

Variation 1 did slightly better than the MUC-3 base system (on TST3, it got 1% better recall at no loss in precision). Variations 2 and 3, although more "linguistically principled", did slightly worse (variation 2 lost 2% recall, 1% precision on TST3). We therefore used variation 1 for our official run.

[1] The set of lexico-semantic models grew by about 25% over MUC-3; the set of concepts (except for geographical names) by about 15%. A partial failure analysis for TST3 suggested that many of the template errors could be attributed to gaps or errors in the models or concepts, and hence that further improvements in these two components were crucial to improved performance.

An examination of some of the errors indicated that, while variations 2 and 3 did OK in and of themselves, they were sensitive to errors in prior stages of processing (in particular, shortcomings in semantic interpretation led to occasional incorrect anaphora resolution, which in turn led to excess event merging). In contrast, paragraph boundaries, while not as reliable a discourse indicator, are more reliably observed. Thus, the best component in isolation may not be the best choice for a system, because it may be too sensitive to errors made by prior components.

RELATED RESEARCH

Much of our time since MUC-3 was involved in research using the MUC-3/MUC-4 corpus and task. We describe here very briefly some of our work related to semantic acquisition, evaluation, and multi-lingual systems.

WORDNET

One of our central interests lies in improving the methods used for acquiring semantic knowledge for new domains. As we noted earlier, we did not invest much additional effort (beyond that for MUC-3) in manual data analysis in order to augment the conceptual hierarchy and lexico-semantic models. We instead conducted several experiments aimed at more automatic semantic acquisition.

One of these experiments involved using Wordnet, a large hierarchy of word senses (produced by George Miller at Princeton), as a source of information to supplement our semantic classification hierarchy. We added to our hierarchy everything in Wordnet under the concepts *person* and *building*.

We identified a number of additional events in this way. Some were correct. Some were incorrect, involving unintended senses of words. For example, the sentence

El Salvador broke diplomatic relations.

would be interpreted as an attack because "relations" (such as "close relations", i.e., relatives) are people in Wordnet. Even more obscure is that

He fought his way back.

becomes an attack because "back" (as in "running back", a football player) is a person. Some of the additional events were correct as events, but should not have appeared in templates, either because they were military ("the enemy") or because they were anaphoric references to prior phrases ("the perpetrator") and so should have been replaced by appropriate antecedents.

These results suggest that Wordnet may be a good source of concepts, but that it will not be of net benefit unless manually reviewed with respect to a particular application.

ACQUIRING SELECTIONAL CONSTRAINTS

An alternative source of semantic information is the texts themselves. NYU has conducted a number of studies aimed at gleaning selectional constraints and semantic classes from the co-occurrence patterns in the sample texts in a domain.

In the past year, we focussed on the task of acquiring the selectional constraints needed for the MUC texts. We have tried to automate this task by parsing 1000 MUC messages (without semantic constraints) and collecting frequency information on subject-verb-object and head-modifier patterns. Where possible, we used the classification hierarchy (which we had built by hand) to generalize words in these patterns to word classes. We then used these patterns as selectional constraints in parsing new text; we found that they did slightly better than the constraints we had created by hand last year [1]. The gain was small -- not likely to affect template score -- but should be an advantage in moving to a new domain, particularly if even larger corpora are available.

We have not yet completed the complementary task of building the word classes from this distributional information.

GRAMMAR EVALUATION

To understand why some systems did better than others, we need some glass-box evaluation of individual components. As we know, it is very hard to define any glass-box evaluation which can be applied across systems.

We have experimented with one aspect of this, grammar (parse) evaluation, which can at least be applied across those systems which generate a full sentence parse.

We use as our standard for comparison the Univ. of Pennsylvania Tree Bank, which includes parse trees for a portion of the MUC terrorist corpus. We take our parse trees, restructure them (automatically) to conform better to the Penn parses, strip labels from brackets, and then compare the bracket structure to that of the Tree Bank. The result is a recall/precision score which should be meaningful across systems.

We have experimented with a number of parsing strategies, and found that parse recall is well correlated with template recall [2].

In principle, we would like to try to extend these comparisons to "deeper" relations, such as functional subject/object relations. These will be harder to define, but may be applicable over a broader range of systems.

MULTI-LINGUAL MUC

We were fortunate to have two researchers from Spain, Antonio Moreno Sandoval and Cristina Olmeda Moreno, who over the past nine months have built a Spanish version of our MUC system (a Spanish grammar, dictionary, and lexico-semantic models) [3]. As this system has developed, we have gradually revised and extended our system so that we can have a language-independent core with language-specific modules.

REFERENCES

[1] Ralph Grishman and John Sterling. Acquisition of Selectional Patterns. To appear in *Proc. 14th Int'l Conf. on Computational Linguistics (COLING 92)*, Nantes, France, July 1992.

[2] Ralph Grishman, Catherine Macleod, and John Sterling. Evaluating Parsing Strategies Using Standardized Parse Files. *Proc. Third Conference on Applied Natural Language Processing*. Trento, Italy, April, 1992.

[3] Cristina Olmeda Moreno and Antonio Moreno Sandoval. El tratamiento semántico en un sistema automático de extracción de información To appear in *Proceedings of Semantica I*, Zaragoza, Spain, May, 1992.