BBN:
Description of the PLUM System as Used for MUC-3

Ralph Weischedel, Damaris Ayuso, Sean Boisen, Robert Ingria, Jeff Palmucci

BBN Systems and Technologies
10 Moulton Street
Cambridge, MA 02138
weischedel@bbn.com

BACKGROUND

Traditional approaches to the problem of extracting data from texts have emphasized handcrafted linguistic knowledge. In contrast, BBN's **PLUM** system (**P**robabilistic **L**anguage **U**nderstanding **M**odel) was developed as part of a DARPA-funded research effort on integrating probabilistic language models with more traditional linguistic techniques. Our research and development goals are

- more rapid development of new applications,
- the ability to train (and re-train) systems based on user markings of correct and incorrect output,
- more accurate selection among interpretations when more than one is found, and
- more robust partial interpretation when no complete interpretation can be found.

We have previously performed experiments on components of the system with texts from the Wall Street Journal, however, the MUC-3 task is the first end-to-end application of PLUM. All components except parsing were developed in the last 5 months, and cannot therefore be considered fully mature. The parsing component, the MIT Fast Parser [4], originated outside BBN and has a more extensive history prior to MUC-3.

A central assumption of our approach is that in processing unrestricted text for data extraction, a non-trivial amount of the text will not be understood. As a result, all components of PLUM are designed to operate on partially understood input, taking advantage of information when available, and not failing when information is unavailable.

The following section describes the major PLUM components.

SYSTEM ARCHITECTURE

The PLUM architecture is presented in Figure 1.

Preprocessing

The input to the system is a file containing one or more messages. The sectioning module determines message boundaries, identifies the header, and determines paragraph and sentence boundaries. In addition, we have built a preprocessor which classifies text according to its relevance and topic. We expect this component to allow the system to ignore paragraphs that are irrelevant and to focus on those that contain relevant information, greatly increasing the efficiency of the overall system. Time constraints did not permit us to integrate this approach with the rest of our system, however; it was therefore not used for the MUC-3 task.

Morphological Analysis

The first phase of the text processing is assignment of part-of-speech information. In our current system, we use the MIT Fast Parser [4]. In the MITFP, a bi-gram probability model, frequency models for known words (derived from large corpora) and heuristics based on word endings for unknown words, assign part of speech to the highly ambiguous words of the corpus.[1] Since the MITFP predictions for unknown words were very inaccurate for input that is all upper case, we augmented this part-of-speech tagging with probabilistic models (automatically

[1] We are now in the process of integrating BBN's POST probabilistic part-of-speech tagger [8] for the tagger in MITFP.

trained) for recognizing words of Spanish origin and words of English origin. This allowed us to tag new words that were actually Latin American names highly reliably. The Spanish classifier uses a 5 character hidden Markov model, trained on about 30,000 words of Spanish text. The five-gram model of words of English was derived from text from the Wall Street Journal.

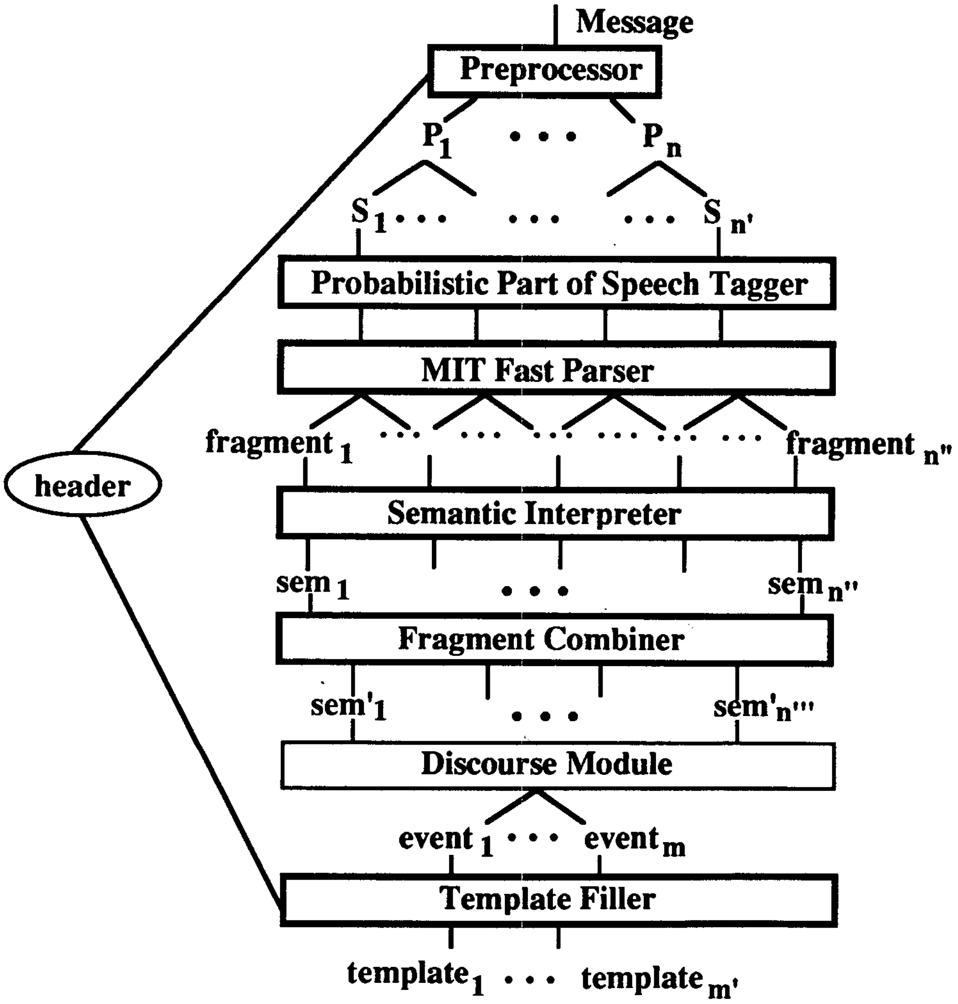

Figure 1. PLUM System Architecture

Parsing

Each sentence identified by the sectioning module is passed to the parsing component. The MITFP is a deterministic stochastic parser which does not attempt to generate a single syntactic interpretation of the whole sentence, rather, it generates one or more parse fragments spanning the input sentence, deferring difficult decisions on attachment ambiguities. Consequently, every sentence is assigned some (set of) syntactic interpretations, producing an average of seven fragments for sentences of the complexity seen in the MUC-3 corpus.

Here are the parse fragments generated by MITFP for the second sentence of message 99 in the TST1 corpus, "THE BOMBS CAUSED DAMAGE BUT NO INJURIES" (the full text of the message is in Appendix H):

```
("THE BOMBS CAUSED DAMAGE"
  (S (NP (DET "THE") (N "BOMBS"))
     (VP (AUX) (VP (V "CAUSED")
                   (NP (N "DAMAGE")))))
("BUT"
  (CONJ "BUT"))
("NO INJURIES"
  (NP (DET "NO")(N "INJURIES")))
("."
  (PUNCT "."))
```

Semantic Interpreter

The semantic interpreter operates on each fragment produced by MITFP in a bottom-up, compositional fashion. Throughout the system, defaults are provided so that missing semantic information or rules do not produce errors, but simply mark semantic elements or relationships as unknown. This is consistent with our belief that partial understanding has to be a key element of text processing systems, and missing data has to be regarded as a normal event.

The semantic component encompasses both lexical semantics and semantic rules. The semantic lexicon is separate from the parser's lexicon and has much less coverage. At present it contains the following numbers of entries:

Adjectives:	**98**
Verbs	**205 (roots)**
Common nouns	**1433 (roots)**
Location names	**1310**
Proper names	**200**

Lexical semantic entries typically include a domain model concept, as well as predicates pertaining to it. For example, here is the lexical semantics for the verb BOMB:

```
(defverb BOMB-V-1 "BOMB" BOMBING
   (:case
    (subject PEOPLE TI-PERP-OF)
    (object ANYTYPE OBJECT-OF)))
```

This entry indicates that the domain model concept is BOMBING, that a subject argument whose type is PEOPLE should be given the role TI-PERP-OF, and that an object argument of any type should be given the role OBJECT-OF. BOMB-V-1 is the unique identifier of this word sense.

The semantic rules are based on general syntactic patterns, using wildcards and similar mechanisms to provide an extra measure of robustness. The basic elements of our semantic representation are "semantic forms", each of which introduces a variable (e.g. ?13) with a type taken from the domain model, and a collection of predicates pertaining to that variable.

There are three basic types of semantic forms: entities of the domain, events, and states of affairs. Each of these three can be further categorized as known, unknown, and referential. Entities correspond to the people, places, things, and time intervals of the domain. These are related in important ways, such as through events (who did what to whom) and states of affairs (properties of the entities). Entity descriptions typically arise from noun phrases; events and states of affairs may be described in clauses.

Not everything that is represented in the semantics has actually been understood. For example, the predicate PP-MODIFIER indicates that two entities (expressed as noun phrases) are connected via a certain preposition. In this way, we have a "placeholder" for the information that a certain structural relation holds between these two items, even though we do not know what the actual semantic relation is. Sometimes understanding the relation more fully is of no consequence, since the information does not contribute to the template-filling task. The information is maintained, however, so that later expectation-driven processing can use it if necessary.

Here is a semantic rule which handles, for example, "group of businessmen", "murder of a man", and "terrorists of the FMLN":

For an NP dominating an NP1, and a PP whose PREP is "OF" and which dominates NP2:

If NP1 is in {"GROUP, "BAND"} ; return semantics of NP2

If NP1 is an EVENT of type TERRORIST ; make NP2 the OBJECT-OF NP1 and return result

If type of NP1 is PEOPLE and type of NP2 is ORGANIZATION, merge semantics, showing that NP1 BELONGS-TO NP2

otherwise use a more general NP => NP PP rule

An important consequence of the fragmentation produced by MITFP is that top-level constituents are typically more shallow and less varied than full sentence parses. As a result, more semantics coverage was obtained early on in the development process with few semantic rules than would have been expected if the system had had to cover widely varied syntactic structures before producing any semantic structures. In this way, semantic coverage was added gradually, while the rest of the system was progressing in parallel.

Another novel aspect of our use of the MITFP was in combining its output fragments. After having assigned semantic representations to the fragments, it is often possible to make some of the attachment decisions deferred by the MITFP. For example, it is possible to combine two NPs of compatible semantic types that are conjoined, or attach prepositional phrases preferentially, using information automatically derived from a corpus [7]. While we lacked sufficient time to pursue this as fully as we would have liked, we did use this for certain proper name constructions, and anticipate using further fragment combining strategies as our semantic coverage increases. Figure 2 shows a graphical version of the semantics generated for the first fragment of sentence 1 in message 99:

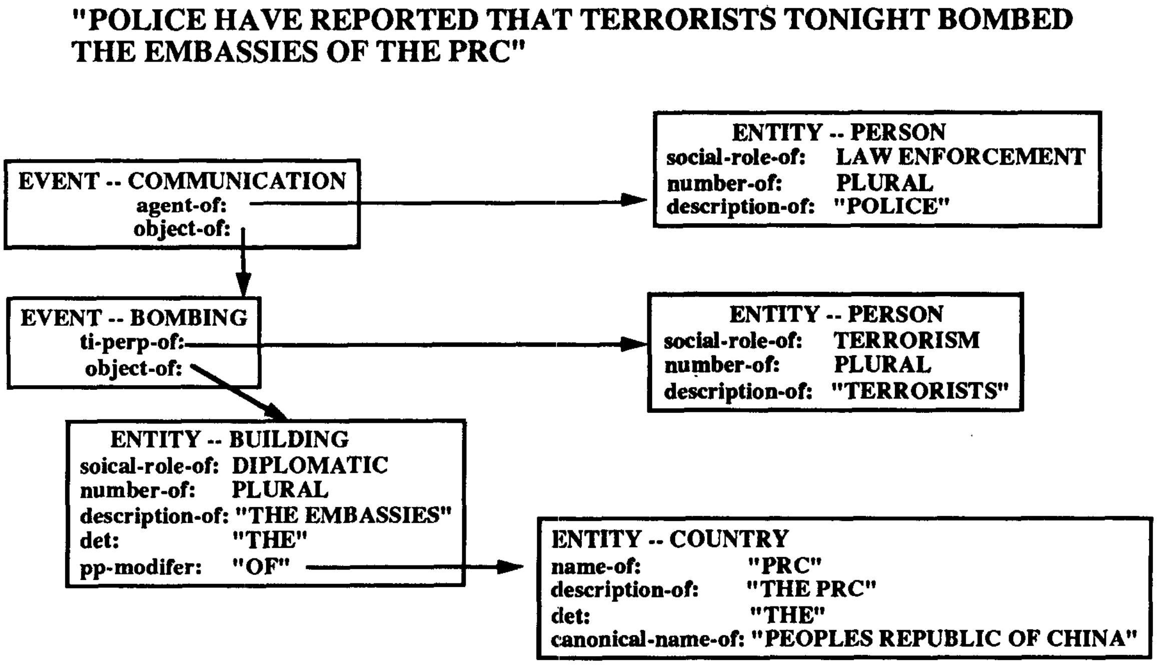
"POLICE HAVE REPORTED THAT TERRORISTS TONIGHT BOMBED THE EMBASSIES OF THE PRC"
EVENT -- COMMUNICATION
agent-of:
object-of:
ENTITY -- PERSON
social-role-of: LAW ENFORCEMENT
number-of: PLURAL
description-of: "POLICE"
EVENT -- BOMBING
ti-perp-of:
object-of:
ENTITY -- PERSON
social-role-of: TERRORISM
number-of: PLURAL
description-of: "TERRORISTS"
ENTITY -- BUILDING
soical-role-of: DIPLOMATIC
number-of: PLURAL
description-of: "THE EMBASSIES"
det: "THE"
pp-modifer: "OF"
ENTITY -- COUNTRY
name-of: "PRC"
description-of: "THE PRC"
det: "THE"
canonical-name-of: "PEOPLES REPUBLIC OF CHINA"

Figure 2: Example Semantic Representation

In this example note that the prepositional phrase in "embassies of the PRC" was not connected properly semantically, as evidenced by the use of the general "pp-modifier" relation. This is because we had no case frame rule for <diplomatic building> of <country>.

Discourse Processing

The discourse component of PLUM performs the operations necessary to derive, from the semantic representation of the fragments in the input message, a high level "discourse event structure", or a representation of the events of interest that occurred in the message. Each event in the discourse event structure is similar in principle to the notion of a "frame", with its corresponding "slots" or fields. There is a correspondence between a discourse event and the semantics that the semantic interpreter assigns to an event in the text. However, the semantic representation assigned by the interpreter can only include relations contained locally in a fragment (after fragment combination); the discourse module must infer other long-distance or indirect relations not explicitly found by the interpreter. The template generator then uses the structures created by the discourse component to generate the final templates. Currently only terrorist incidents (and "possible terrorist incidents") generate discourse events, since these are the core events for MUC-3 template generation. The discourse component is further discussed in the paper "Computational Aspects of Discourse in the Context of MUC-3" in these proceedings.

Two primary structures are created by the discourse processor which are used by the template generator: the discourse predicate-database and the discourse event structure. The database contains all the predicates mentioned in the semantic representation of the message (e.g., that some entity is the object of an event). It supports unification of semantic variables, so that all the information can be easily retrieved when references in the text are resolved. Any other inferences done by the discourse component also get added to the database. While only one database is produced at present, ideally there should be several, to handle multiple inference paths.

To create the discourse event structure, the discourse component processes each semantic form produced by the interpreter, adding its information to the database and performing reference resolution (currently only pronouns and proper name references) when needed. When a semantic form for an event of interest is encountered, a discourse event is generated, and any slots already found by the interpreter are filled in the event. This event is then merged with a previous event if they are compatible. This heuristic assumes that the events were derived from repeated references to a single real event in the text.

Once all the semantic forms have been processed, heuristic rules are applied to fill in any unfilled slots by looking at text surrounding the forms which triggered a given event. Each filler found is assigned a score based on where it was found in relation to an event trigger, indicating a higher confidence for fillers found closer to a trigger. This will not always be a valid assumption, but has proved to be a good approximation.

Following is the discourse event structure created by using information in the first three sentences (spanning 2 paragraphs) of message 99:

> "POLICE HAVE REPORTED THAT TERRORISTS TONIGHT BOMBED THE EMBASSIES OF THE PRC AND THE SOVIET UNION. THE BOMBS CAUSED DAMAGE BUT NO INJURIES."
>
> "A CAR-BOMB EXPLODED IN FRONT OF THE PRC EMBASSY, WHICH IS IN THE LIMA RESIDENTIAL DISTRICT OF SAN ISIDRO. MEANWHILE, TWO BOMBS WERE THROWN AT A USSR EMBASSY VEHICLE THAT WAS PARKED IN FRONT OF THE EMBASSY LOCATED IN ORRANTIA DISTRICT, NEAR SAN ISIDRO."

```
Event: BOMBING
Trigger: "BOMBED" (?29)
   Slots:
      TI-PERP-OF: "TERRORISTS" (?9, score=0)
      EVENT-TIME-OF:
      EVENT-LOCATION-OF:
        "EL SALVADOR" (?100, score=6)
        " SAN ISIDRO" (?104, score=6)
        " RESIDENTIAL DISTRICT" (?105, score=6)
        "ORRANTIA DISTRICT" (?169, score=6)
     TI-INSTRUMENT-OF : "THE BOMBS" (?41, score=4)
     TI-RESULT-OF:
        "DAMAGE" (?46, score=4)
        "NO INJURIES" (?54, score=4)
```

OBJECT-OF: "THE EMBASSIES" (?22, score=0)

In the example above, a score of 0 indicates the filler was found directly by the semantics; 4 indicates it was found in the same paragraph; and 6 that it was found in an adjacent paragraph. Note that El Salvador, though not in the text, was introduced by the definition of San Isidro in the lexicon, which had only been seen previously as a town of El Salvador.

Template Generation

The template generator takes the event structure produced by discourse processing and fills out the application-specific templates. Clearly much of this process is governed by the specific requirements of the application, considerations which have little to do with linguistic processing. For example, in our domain model, all terrorist incidents have a result, but the MUC-3 task description states that, if the incident type is MURDER, the RESULT slot is to be left unspecified. The template generator must incorporate these kinds of arbitrary constraints, as well as deal with the basic details of formatting.

The template generator uses a combination of data-driven and expectation-driven strategies. First the information in the event structure is used to produce initial values. At this point, values which should be filled in but are not available in the event structure are supplied from defaults, either from the header (e.g., date and location information) or from reasonable guesses (e.g. that the object of a murder is usually a suitable filler for the human target slot when the semantic type of the object is unknown).

We expect to eventually use a classifier at this stage of processing. This is especially appropriate for template slots with a set list of possible fillers, e.g. perpetrator confidence, category of incident, etc.

EXAMPLE

Here is the first template generated by PLUM for message 99 in the TST1 corpus:

0.	**MESSAGE ID**	**TST1-MUC3-0099**
1.	**TEMPLATE ID**	**1**
2.	**DATE OF INCIDENT**	**- 25 OCT 89**
3.	**TYPE OF INCIDENT**	**BOMBING**
4.	**CATEGORY OF INCIDENT**	**TERRORIST ACT**
5.	**PERPETRATOR: ID OF INDIV(S)**	**"TERRORISTS"**
6.	**PERPETRATOR: ID OR ORG(S)**	**-**
7.	**PERPETRATOR CONFIDENCE**	**-**
8.	**PHYSICAL TARGET: ID(S)**	**"THE EMBASSIES"**
9.	**PHYSICAL TARGET: TOTAL NUM**	**PLURAL**
10.	**PHYSICAL TARGET: TYPE(S)**	**DIPLOMAT OFFICE OR RESIDENCE: "THE EMBASSIES"**
11.	**HUMAN TARGET: ID(S)**	**-**
12.	**HUMAN TARGET: TOTAL NUM**	**-**
13.	**HUMAN TARGET: TYPE(S)**	**-**
14.	**TARGET: FOREIGN NATIONS**	**-**
15.	**INSTRUMENT: TYPE(S)**	*****
16.	**LOCATION OF INCIDENT**	**EL SALVADOR: SAN ISIDRO (TOWN)**
17.	**EFFECT ON PHYSICAL TARGET**	**SOME DAMAGE: "THE EMBASSIES"**
18.	**EFFECT ON HUMAN TARGET**	**NO INJURY: "-"**

Several things were processed correctly here:

- we correctly identified the nature of the attack, the identity of the attacking individuals, and the identity and type of the target, and
- we correctly determined the nature of the damage, including the negation in "NO INJURIES".

However, several points were missed:

- we failed to understand "TONIGHT", and so filled in the default of some time before the header date;
- the identity of the terrorist organization was missed because our strategy for looking for perpetrators was too inflexible and did not keep looking once "TERRORISTS" was found;

- our system does not yet attempt to fill the foreign target slot, so naturally we missed that filler; and
- our semantics for locations are too limited, listing only the town of San Isidro (which is in El Salvador) and not the neighborhood of San Isidro (which is in Lima, Peru). There is a reference to Lima; the syntactic structure assigned, however, does not permit the proper semantics to identify it as a location.

ACKNOWLEDGEMENTS

The work reported here was supported in part by the Defense Advanced Research Projects Agency and was monitored by Rome Laboratory under Contract Nos. F30602-87-D-0093 and F30602-91-C-0051. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

REFERENCES

[1] Ayuso, D.M., Bobrow R., MacLaughlin, D., Meteer, M., Ramshaw, L., Schwartz, R. and Weischedel, R. Toward Understanding Text with a Very Large Vocabulary. In *Proceedings of the Speech and Natural Language Workshop*, Morgan-Kaufmann Publishers, Inc. June, 1990.

[2] Church, K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proceedings of the Second Conference on Applied Natural Language Processing,* pages 136-143. ACL, 1988.

[3] Crowther, W. A Common Facts Data Base. In *Proceedings of the Speech and Natural Language Workshop,* pages 89-93. Morgan Kaufmann Publishers Inc., San Mateo, CA, February 1989.

[4] de Marcken, C.G. Parsing the LOB Corpus. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 243-251. 1990.

[5] Marcus, M., Santorini , B., and Magerman, D. 1990, "First Steps Towards an Annotated Database of American English" *Readings for Tagging Linguistic Information in a Text Corpus* Langendoen and Marcus, tutorial for the 28th Annual Meeting of the Association for Computational Linguistics.

[6] Santorini, B. *Annotation Manual for the Penn Treebank Project.* CIS Department. University of Pennsylvania. May 1990.

[7] Weischedel, R., Ayuso, D. M., Bobrow, R., Boisen, S., Ingria, R., and Palmucci, J. Partial Parsing, A Report on Work in Progress, *Proceedings of the Fourth DARPA Speech and Natural Language Workshop,* February 1991.

[8] Weischedel, R., Meteer, M., and Schwartz, R., Empirical Studies in Part of Speech Labelling, *Proceedings of the Fourth DARPA Speech and Natural Language Workshop,* February, 1991.