ITP INTERPRETEXT SYSTEM: MUC-3 TEST RESULTS AND ANALYSIS

Kathleen Dahlgren
Carol Lord
Hajime Wada
Joyce McDowell
Edward P. Stabler, Jr.

Intelligent Text Processing, Inc.

1310 Montana Avenue, Suite 201
Santa Monica, CA 90403
213-576-4910
Internet: 72550,1670@compuserve.com

Intelligent Text Processing is a small start-up company participating in the MUC-3 exercise for the first time this year. Our system, Interpretext, is based on a prototype text understanding system. With three full-time and three part-time people, dividing time between MUC-3 and other contract projects, ITP made maximum use of modest resources.

SLOT	POS	ACT	COR	PAR	INC	ICR	IPA	SPU	MIS	NON	REC	PRE	OVG	FAL
Matched Only	794	479	243	91	76	30	76	78	393	400	35	58	16	
Matched/Missing	1372	479	234	91	76	30	76	78	971	793	20	58	16	
All Templates	1372	604	234	91	76	30	76	203	971	1031	20	46	34	
Set Fills Only	575	191	99	31	25	8	31	36	420	492	20	60	19	0

Figure 1. Intelligent Text Processing Final Scores Test 2

ITP's results are shown in Figure 1. The ITP system was second highest in precision (46%) when all templates were considered, and at the same time achieved a credible recall percentage (20%). Our overgeneration rate was second best (34%). ITP was a very close second in both precision and overgeneration, as the top percentages were 48 and 33 to ITP's 46 and 34. The major limiting factor in ITP's MUC-3 performance was parser failure. We are building a parser with wide coverage and a comprehensive approach to disambiguation. Because our parser is not yet complete, in order to participate in the MUC-3 exercise we used a parser on loan.

It proved to lack the robustness necessary to parse the MUC-3 messages, failing on 50% of the sentences. For those sentences which it did parse, the Interpretext system returned precise semantic interpretations. ITP's word-based approach required minimal reorientation in shifting to the new domain of terrorism texts; the main new material was the straightforward addition of a relatively small number of new words to the syntactic and naive semantic lexicons, not whole new semantic modules. The semantic structures and analyses already implemented proved to be appropriate for texts in the new domain.

The source of the precision in our performance was the Cognitive Model built by the Natural Language Understanding Module. The Cognitive Model contains specific reference markers identifying events and individuals in the text. The same events and individuals are given the same

reference markers by the Anaphora Resolution Module. The Cognitive Model distinguishes between events, individuals and sets. It directly displays the argument structure of events. Thus, to find a terrorist incident, the template-filling code looked for an event which implied harm, damage or some other consequence of terrorism in the Naive Semantics for the verb naming the event. The agent of the event had to be described as having a role in clandestine activity, the government or the military. The ITP naive semantic lexicon distinguishes between nouns which names objects and nouns which name events, so that the template-filling code had only to look for events, even those introduced by phrases such as *the destruction of homes in ...*

Furthermore, the Cognitive Model connects head nouns with prepositional phrase modifiers and adjectival or nominal modifiers via the same reference marker. Thus the template-filling code could look for a variety of modifiers of an individual as a source of information about the individual. For example, the phrase *member of the guerrilla troop* connects *member* with *troop* and *guerrilla*, so that the template-filling code could recognize a semantically empty term like *member* as referring to an agent. This type of connection works everywhere, not just with the particular string pattern *member of the guerrilla troop*. Furthermore, it is much more precise than a pattern-matching method which would find *guerrilla* as perpetrator everywhere it occurs, even when a phrase like "member of the guerrilla troop" is the object of a verb which implies harm, and is therefore not indicative of guerrilla terrorism.

Another source of precision is that the formal semantic module interprets the cardinality of sets. "None", "plural" or "three" come out in the formal representation as the number of objects in a set. Finding target number and amount of injury and damage is trivial given a precise treatment of cardinality in the formal semantics.

Finally, the Cognitive Model indicated discourse segments. These are portions of the text which function as a unit around one topic. The recognition of segments simplified the anaphora resolution and the process of identifying the same individuals and events with each other. It prevented the overgeneration of templates. Some competitor systems generated a new template for each sentence containing a terrorism word and then they had to try to merge them. Without segment information, merging was very difficult.

A Cognitive Model with this level of precision can be built only when a deep natural language analysis of the text is performed. Syntactic, formal semantic, discourse semantic and pragmatic (or naive semantic) complexities of text are addressed by the ITP Natural Language Understanding Module. Some researchers have rejected a principled linguistic approach as hopeless at this stage in the history of computational linguistic research. They assume that the only feasible methods are statistical. Such systems match to certain string patterns and rely upon the statistical probability that they co-occur with a particular semantic interpretation. The problem is that many times the pattern occurs in phrase which is irrelevant, or has the opposite meaning to the predicted one. The pattern can occur in the scope of a negative or modal, as in *the bomb did not explode,* and produce a false alarm for a pattern-matching method. Such methods will tend to over-generate templates, because patterns indicate a terrorist incident where there is none. For the same false alarm texts, more precise linguistic analysis can correctly rule out a terrorist incident.

Furthermore, the patterns for matching must be coded anew for each domain. In contrast, ITP Naive Semantic and syntactic lexicons need only be built once, and they work across all domains. For MUC-3 we added to an existing naive semantic lexicon prepared originally for texts in other domains.

In summary, ITP was precise in the MUC-3 fills for the sentences which our loaner parser was able to process. When our own parser is available, ITP's technology will vastly improve in recall.

Naive Semantics

The basic approach to template-filling involved looking at feature types in the naive semantic knowledge for verbs and nouns. The feature types inspected had already been present in the theory and in the system prior to MUC-3. The verb feature "consequence of event" was important for recognizing terrorist incidents, because if the typical consequence of an event was damage or harm, it triggered a template fill. The theory of Naive Semantics as described in Dahlgren[1] identifies that feature type as important in lexical semantics and reasoning about discourse. Similarly, the "rolein" feature was used to distinguish between clandestine agents, government agents and military agents. Again, that feature type was antecedently present in our theory.

Test Settings

The effect of the MUC-3 reader was to exclude any sentences which did not contain a terror word, saving processing time. This setting tended to reduce precision, because a sentence like *She succeeded* contains no terrorism word, but could be very significant in the recognition of a terrorist incident. Recall was implicitly set very low by the fact that the parser was able to parse only 50% of the input.

Level of Effort

The greatest effort by ITP was the six years of research that went into the Natural Language Understanding Module. As for MUC-3-specific tasks, Table I indicates the level of effort on each one. ITP made a detailed linguistic analysis of the terrorism domain, and the way that terrorist incidents were described in the first messages sent out by NOSC, and in the DEV messages. The analysis guided the expansion of the lexicons and the writing of the template-filling code. During Test 1 we identified both parser failure and parse time to be problems in our performance. Therefore, for Test 2 we built a reader which could handle dates, abbreviations, and so on, and would return a sentence only if it contained a terrorism word. In addition, we pruned the output to shorten sentences for the parser. These tactics will not be necessary once our own wide-coverage parser is completed. The template-filling code took about as much of our time as the reader and pruner. Each element of the code reasons from the Cognitive Model using generalized lexical reasoning or DRS reasoning. The temporal-locative reasoning is general and will be used in other applications.

Tasks	Estimated Person-weeks
Linguistic analysis of terrorism domain	4
Syntactic Lexicon expansion	2
Naive Semantic Lexicon expansion	3
Reader, pruner	4
Temporal, locative reasoning	2
Template-filling code	4

Table 1. MUC-3 specific Tasks and their Estimated Person-Weeks

Limiting Factor

The main limiting factors were the parser and resources. With more persons and time, we could have written code for all of the fills and debugged the template-filling code thoroughly. Given the modest resources we had, we were forced to run the test before we had thoroughly debugged the code. In particular, our code for recognizing and building up proper names was in

place, but failed during the test in most cases. That explained our performance on Perpetrator Organization. Given that we missed the latter, we of course could not get Category of Incident correct for any of the State-sponsored Violence cases either.

Training

Training took place on the first 100 DEV messages, and on Test 1 messages with the new key. We did not have sufficient resources to fully debug and repeatedly test prior to MUC-3 week. The system improved dramatically between Test 1 and Test 2 (from recall of 3 to recall of 20). Improvement was mainly due to expansion of the template-filling code and the introduction of pruning to get more parses.

Success and Failure

For those sentences which we were able to parse, the reasoning performed well for incident recognition, segmentation (separating different incidents in the same message), perpetrator and target recognition. The only exceptions were perpetrators or targets with long proper names. We have an approach to these, but didn't get it working in time. The fills which failed were perpetrator organization (because of names), and target nationality. The latter code is working fine (it looks to see whether any descriptor of an individual is a foreign nation name or adjective). The failures were due to missing the whole template because of parsing, or missing the target in a recognized template. In addition, our target number code was not fully operational at the time of the test. We would most like to rewrite the template-filling code in even more general reasoning algorithms which could be used in applications beyond the terrorism domain. Our system's capabilities make possible a question-answering system which could reply to English queries like *Who did it?* and *How many people were killed?*.

Reusability

Everything but the template-filling code is reusable in a different application. All of the words we added to the lexicons have all of their senses common in American English. They can be used in any domain. As for the template-filling code, we plan to extract generalizable reasoning algorithms for use in other domains. Again, the code is reusable because it is a principled, general linguistic approach rather than a pattern-matching approach.

What we learned

We learned that anything a person wants to say or write can be said in an extremely large number of different ways. Therefore, a robust deep natural language understanding system must have a wide-coverage parser and formal semantics which directly display the similarity of content across many possible forms of expression. A sound theoretical approach such as DRT is particularly appropriate for a data extraction task. Secondly, we learned that natural language systems require ample testing against real-world texts. And, third, a system in which word meanings are central, developed to interpret text in the domains of geography and finance, can function in the domain of terrorism with the addition of a relatively small number of lexical items.

References

[1] Dahlgren, K. (1988). *Naive Semantics for Natural Language Understanding*. Kluwer Academic Publishers, Norwell, Mass.

[2] Dahlgren,K. (1989). "Coherence Relation Assignment," in *Proceedings of the Cognitive Science Society,* pp.588-596.

[3] Kamp, H. (1981). "A Theory of Truth and Semantic Representation," in Gronendijk, J.; T. Janssen; and M. Stokhof, editors, *Formal Methods in the Study of Language*, Mathematisch Centrum, Amsterdam.