DESCRIPTION OF LOCKHEED MARTIN'S NLTOOLSET

AS APPLIED TO MUC-7 (AATM7)

Deborah Brady

Lois Childs

David Cassel

Bob Magee

Norris Heintzelman

Dr. Carl Weir

Lockheed Martin

Management & Data Systems (M&DS)

Building 10, Room 1527

P.O. Box 8048

Philadelphia, PA 19101

BACKGROUND

The NLToolset has been used to build a variety of information extraction applications, ranging from military message traffic to newswire accounts of corporate activity. AATM7 is an acronym for As Applied To MUC-7. AATM7 was not tailored specifically for MUC-7, but rather represents the NLToolset in a state of flux, as TIPSTER experimentation and the delivery of a real-world application were taking place, simultaneously. This contrast in domains proved beneficial for our real-world applications, perhaps to the detriment of the MUC-7 system, which had to compete for developers.

NLToolset applications are delivered under the Windows NT, as well as the UNIX Solaris operating system.

TEMPLATE ELEMENT TASK

AATM7 was applied to the MUC-7 Template Element task in order to test some theories of coreference that were being investigated under the TIPSTER III research activity. The Template Element task requires an automatic system to build templates for every person, organization, and artifact entity, as well as every location.

Entities

The Entities are defined as follows:

An organization object consists of:

organization's name and aliases found in the text,

a type slot of ORGANIZATION,

one descriptor phrase, and

the category of the organization: ORG_CO, ORG_GOVT, or ORG_OTHER.

A person object consists of:

person's name and aliases found in the text,

a type slot of PERSON,

one descriptor phrase, and

the category of the person: PER_CIV or PER_MIL

An artifact object consists of:

artifact's name and aliases found in the text,

a type slot of ARTIFACT,

one descriptor phrase, and

the category of the artifact: ART_AIR, ART_LAND, or ART_WATER.

To perform this task perfectly, an automatic system must link all references to the same entity within a text, and collect those references, whether they be names or descriptive noun phrases. The entire list of unique names for an entity is placed in the "NAME" slot. Of the descriptors, the system must pick one of those found, and put it in the "DESCRIPTOR" slot, as long as it is not "insubstantial" according to the fill rules, e.g. "the company" or "Dr." Pronouns are also excluded from the entity object. Additionally, the system must decide to what category the entity belongs, either through its knowledge base or the surrounding context, e.g. "Gen. Smith" vs. "Ms. Smith" as PER_MIL vs. PER_CIV.

The limitation to one descriptor can have the effect of hiding how well the coreference resolution has performed, since a system may have found all descriptive phrases, plus one incorrect descriptor, and chosen the incorrect descriptor, thus getting a score of incorrect for the entire slot. Lockheed Martin is planning to test a multiple-descriptor version of MUC-7, in the near future.

Of the three entity types, those of "PERSON" and "ORGANIZATION" are the most similar, since language is used in similar ways to describe them. They both can be named, where the "name" is an identity which, within the context of a story, is usually unique. The artifact, which in MUC terms can be a land, air, sea, or space vehicle, is sometimes named, but often the tag which is considered the name is merely a type. For example, a story that tells about three different F-14 crashes may, according to MUC rules, produce three different entities named "F-14", whose only difference would be found in information not captured by the TE object.

Locations

Locations are defined as follows:

A location object consists of:

locale found in the text,

the country where the locale exists, and

The location object's locale slot is filled with the most specific reference to a location. For example, if the location were "Philadelphia, PA," the locale slot would be filled with "Philadelphia." The country would be "United States" and the locale type would be "CITY." The deficiency of this design is obvious; it fails to differentiate between the actual location and any other city named "Philadelphia" in the nation. An alternative design, which has been used for other NLToolset applications, contains a locale slot which holds the entire phrase describing the locale. Some examples are:

"at the checkpoint on Route 30"

"southwest of Miami"

"Wilmington, Delaware"

Additionally, the location object contains slots for whatever other information can be gleaned from the text or from on-line resources, such as a gazetteer. This includes slots for city, country, province, latitute/longitude, region, or water.

TIPSTER Research

AATM7 was developed with a focus on the investigation of a number of techniques involved in coreference resolution. Coreference Resolution can be thought of as the identification and linking of all references to a particular entity. References may be in the form of names, pronouns, or noun phrases.

Syntax is frequently used by an author to associate a descriptive phrase with an entity. This can be seen in the following examples:

APPOSITIVE: "Lockheed Martin, an aerospace firm,"

PRENOMIAL: "the aerospace firm, Lockheed Martin"

NAME-MODIFIED HEAD NOUN: "the Lockheed Martin aerospace firm"

PREDICATIVE NOMINATIVE: "Lockheed Martin is an aerospace firm"

When an entity is referred to only by a descriptive phrase, finding its true identity is very challenging. The following sentence

"*The president has announced that he will resign.*"

has varying degrees of import, depending on its preceding sentence…

"*Coca Cola Company today revealed the future plans of its president, James Murphy.*"

"*Impeachment hearings were scheduled to begin today against President Clinton.*"

An automatic system can use the information closely related by syntax to the entity, in this case the title "President" or the prenominal "its president", to identify the entity referred to by "the president." This is the heart of our current research. Our aim is to find all descriptive information closely related by syntax and to build a story-specific ontology for each entity so that far-flung references that depend on this semantic information can be identified. As part of this research, the Template Element development keys were analyzed to determine how often the descriptors of an organization and person are directly associated by syntax. A surprisingly large number of descriptive phrases within the keys can be directly associated to an entity by way of syntax. Of a total of approximately 900 descriptors, 125 were organization descriptors, and 775, person descriptors --- a disproportionate number, since there are actually more organization entities (985) than person entities (802) in the keys.

The following table shows the breakdown by category and entity type. "Association by Context" refers to descriptors that have been found in titles, prenominal phrases, appositives, and predicate nominatives. "Association by Reference" refers to

a remote reference which refers to a named entity. "Un-named" refers to entities described by noun phrases alone, e.g. "a local bank."

Category	Person	Organization
Association by Context	548 (71%)	33 (26%)
Association by Reference	103 (13%)	53 (42%)
Un-named	119 (15%)	38 (30%)

Table 1: Training Set Analysis

This data supports the hypothesis that much reliable descriptive information can be obtained through syntactic association. This descriptive information can be associated with the entity object and then be used to help resolve associations by reference, in a manner similar to that used for organizations in the Lockheed Martin MUC-6 system, LOUELLA. This is the idea of a semantic filter, which was used to compare descriptive phrases with the semantic content of organization names, as in the following example.

"Buster Brown Shoes" => (buster brown shoes shoe footwear)

"the footwear maker" => (footwear maker make manufacturer)

Since person names rarely include semantic content, we must rely on other descriptive information to build the semantics, either through world knowledge stored in the system's knowledge base or through associations found in the text itself.

As part of Lockheed Martin's TIPSTER research, the freeware Brill part-of-speech tagger was connected to the NLToolset to see if it could help streamline the process of building patterns to find descriptors. Since standard NLToolset processing provides all possible parts of speech for each token, a part-of-speech tagger was introduced to see if it could simplify the process of pattern writing. It was found that a package for finding and correctly linking the majority of person descriptors could be written in about a week by incorporating the information that Brill provides with that provided by the NLToolset, i.e. symbol name, semantic category, and possible parts of speech as found in the NLToolset's lexicon. The contrast between the descriptor scores for persons and organizations in the test set is striking.

DESCRIPTORS	RECALL	PRECISION
PERSON	61	55
ORGANIZATION	28	20

Table 2: Descriptor Scores

Finding artifacts and linking up all references to the same entity has proved especially challenging because of the unusual way that artifacts are described in text, and the way that the descriptions are categorized for MUC-7. For instance, "Boeing 747" and "F-14" are considered names, whereas "TWA Flight 800" is considered a descriptor. Under the TIPSTER research, a new algorithm was developed to find vehicles and resolve coreferences. The algorithm differs from that for organizations and people in that a match is assumed to belong to the most recently seen entity, unless there is some information to contradict this assumption. The possible types of contradictory information are: model information, manufacturer, military branch, airline, and flight number. Further, if the comparison reveals that one entity has military information and the other has airline information, there is a contradiction. Further, the variable-binding feature of the NLToolset's pattern matching allows the developer to extract type information while finding the entities in the text. This type information helps the system to distinguish between entities during coreference resolution.

RESULTS ANALYSIS

Overall, AATM7's scores for MUC-7 are good. There are a few errors, as well as some quirks of the MUC-7 domain, that will be discussed which significantly effected the scores for entity names and locations. The artifact scores are significantly below the NLToolset's usual performance; this is due to the newness of this entity, particularly of the space vehicle artifacts. This capability is still a work in progress, as the need arises for our real-world applications

.

*   *   *   **SUMMARY SCORES**   *   *   *

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
SUBTASK SCORES														
entity														
artifact	197	241	165	0	17	15	59	12	**84**	**68**	8	24	9	36
organization	866	910	800	0	33	33	77	11	**92**	**88**	4	8	4	15
person	469	534	457	0	7	5	70	0	**97**	**86**	1	13	2	15
location														
airport	21	18	1	0	17	3	0	2	**5**	**6**	14	0	94	95
city	226	226	197	0	9	20	20	2	**87**	**87**	9	9	4	20
country	260	239	221	0	10	29	8	7	**85**	**92**	11	3	4	18
province	52	53	42	0	6	4	5	19	**81**	**79**	8	9	13	26
region	89	40	29	0	11	49	0	4	**33**	**73**	55	0	28	67
unk	33	14	4	0	10	19	0	3	**12**	**29**	58	0	71	88
water	12	10	10	0	0	2	0	1	**83**	**100**	17	0	0	17
OBJ SCORES														
location	693	626	554	0	32	107	40	18	**80**	**88**	15	6	5	24
entity	1532	1685	1432	0	47	53	206	23	**93**	**85**	3	12	3	18
SLOT SCORES														
location														
locale	697	626	511	0	75	111	40	24	**73**	**82**	16	6	13	31
locale_type	693	600	504	0	63	126	33	38	**73**	**84**	18	6	11	31

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
country	691	624	493	0	90	108	41	26	**71**	**79**	16	7	15	33
entity														
ent_name	1761	1731	1305	0	159	297	267	28	**74**	**75**	17	15	11	36
ent_type	1532	1685	1422	0	57	53	206	23	**93**	**84**	3	12	4	18
ent_descrip	680	819	338	0	175	167	306	585	**50**	**41**	25	37	34	66
ent_categor	1532	1685	1340	0	139	53	206	53	**87**	**80**	3	12	9	23
ALL SLOTS	7586	7771	5913	0	758	915	1100	1061	**78**	**76**	12	14	11	32

	P&R	2P&R	P&2R
F-MEASURES	**77.01**	76.45	77.57

Table 3: Overall MUC-7 Scores

Since the TE task spans four separate subtasks with very different characteristics, an analysis was done on each. The formal run keys were split into four sets: organization, person, artifact, and location keys. The formal run was then also split into organization, person, artifact, and location responses. Each set was then respectively scored with SAIC's version 3.3 of the MUC scoring program. The results are described below. This scoring method removes the mapping ambiguity between entities of different types and allows an accurate analysis of the performance of each individual entity type.

People

AATM7 found 97% of the people objects, with 86% of the names correctly. The slot scores are high, even the descriptor slot, which has traditionally been at less than 50%. To improve on this performance, one problem that could very easily be resolved is an incorrect interpretation of expressions like "(NI FRX)" in the formal text. "NI" is a common first name in some languages and therefore, AATM7 interpreted all thirteen of these as person names. This error accounted for 13 of the overgenerated or incorrect person names, or the equivalent of 2 points of precision.

Another area for improvement is in the descriptor slot. Twenty-six of AATM7's person descriptors were marked incorrect because they contained only the head of the noun phrase and not the entire phrase, e.g. "commander" instead of "Columbia's commander" and "manager" instead of "project manager." The descriptor rule package will be improved to better encompass the entire phrase. If these descriptors had been extracted correctly for the MUC-7 test, the descriptor recall and precision would have improved to 70 and 63, while the overall person scores would have improved to 89 recall, 79 precision, and 83.7 F-measure.

* * * **SUMMARY SCORES** * * *

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
SUBTASK SCORES														
person	469	560	457	0	0	12	103	0	**97**	**82**	3	18	0	20
OBJ SCORES														

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
entity	469	560	457	0	0	12	103	0	97	82	3	18	0	20

SLOT SCORES

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
entity														
ent_name	568	564	491	0	14	63	59	1	86	87	11	10	3	22
ent_type	469	560	457	0	0	12	103	0	97	82	3	18	0	20
ent_descrip	302	335	184	0	61	57	90	147	61	55	19	27	25	53
ent_categor	469	560	444	0	13	12	103	28	95	79	3	18	3	22
obj_status	0	0	0	0	0	0	0	4	0	0	0	0	0	0
comment	0	0	0	0	0	0	0	13	0	0	0	0	0	0
ALL SLOTS	1808	2019	1576	0	88	144	355	193	87	78	8	18	5	27

	P&R	2P&R	P&2R
F-MEASURES	**82.36**	79.72	85.18

Table 4: Person Object Scores

Organizations

Organizations are complex entities to determine in text because organization names have a more complex structure than person names. A variation algorithm for one name may not work for another. For example, "Hughes" is a valid variation for "Hughes Aerospace, Inc." but "Space" is not a valid variation for "Space Technology Industries". An automatic system must, therefore, look at the surrounding context of variations and filter out those that are spurious.

AATM7 found 780 of the 877 organizations in the formal test corpus. Of the 780 it found, points were lost here and there for mistakes in two areas. First, current performance on organization descriptors is woefully inadequate and in sharp contrast to that on person descriptors. An effort is currently underway to improve this with the help of a part-of-speech tagger. Additionally, it was discovered that the mechanism for creating and linking variations of organization names was broken during the training period. The result of this was that 64 name variations were missed. When this problem was fixed, recall and precision for ent_name improved to 76 and 77, with the overall organization recall and precision improving to 80 and 77.

*   *   *   **SUMMARY SCORES**   *   *   *

POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
organization	865	889	800	0	0	65	89	12	**92**	**90**	8	10	0	16

OBJ SCORES

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
entity	865	889	800	0	0	65	89	12	**92**	**90**	8	10	0	16

SLOT SCORES

entity

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
ent_name	1062	984	742	0	111	209	131	25	**70**	**75**	20	13	13	38
ent_type	865	889	800	0	0	65	89	12	**92**	**90**	8	10	0	16
ent_descrip	196	265	54	0	44	98	167	126	**28**	**20**	50	63	45	85
ent_categor	865	889	733	0	67	65	89	14	**85**	**82**	8	10	8	23
obj_status	0	0	0	0	0	0	0	69	0	0	0	0	0	0
comment	0	0	0	0	0	0	0	50	0	0	0	0	0	0
ALL SLOTS	2988	3027	2329	0	222	437	476	296	**78**	**77**	15	16	9	33

	P&R	2P&R	P&2R
F-MEASURES	**77.44**	77.14	77.74

Table 5: Organization Object Scores

Artifacts

AATM7's artifact performance really suffers in the area of entity names. It missed almost half of the artifact entities purely from lack of patterns with which to recognize them. This is a sign of the immaturity of the artifact packages and can be overcome by more development. Another problem, which caused the low precision, was the result of incorrectly identifying the owner of the artifact as its name. This accounted for 38 of the spurious entity names and 2% of the precision. Since this is a new package, the coreference resolution is also not up to the NLToolset's usual performance. This is an on-going research effort.

*   *   *   **SUMMARY SCORES**   *   *   *

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR

artifact	197	236	165	0	0	32	71	12	**84**	**70**	16	30	0	38

OBJ SCORES

entity	197	236	165	0	0	32	71	12	**84**	**70**	16	30	0	38

SLOT SCORES

entity

ent_name	130	183	60	0	15	55	108	3	**46**	**33**	42	59	20	75
ent_type	197	236	165	0	0	32	71	12	**84**	**70**	16	30	0	38
ent_descrip	181	219	98	0	48	35	73	313	**54**	**45**	19	33	33	61
ent_categor	197	236	165	0	0	32	71	12	**84**	**70**	16	30	0	38
obj_status	0	0	0	0	0	0	0	27	0	0	0	0	0	0
comment	0	0	0	0	0	0	0	46	0	0	0	0	0	0
ALL SLOTS	705	874	488	0	63	154	323	413	**69**	**56**	22	37	11	53

	P&R	2P&R	P&2R
F-MEASURES	**61.81**	58.08	66.05

Table 6: Artifact Object Scores

Locations

The NLToolset performs well at finding and disambiguating locations. Determining the country for a given location can be complicated since many named locations exist in multiple countries. A small number of minor changes have been identified to significantly boost the score to its normal level. One of the obvious problems AATM7 had was with the airports. Eleven occurrences of Kennedy Space Center were identified as locale type "CITY" instead of the correct type of "AIRPORT". This was caused by a simple inconsistency in our location processing. Fixing this one problem, improved the airport-specific recall and precision to 57 and 67 respectively, and improved the precision overall by 1 percentage point.

The location recall for MUC-7 is slightly depressed because of some challenges which this particular domain presented. AATM7 was not configured to process planet names or other extra-terrestrial bodies as locations. This accounted for sixty-three missing items, at three slots per item; thirty-one of the missing were occurrences of "earth" alone. This is reflected in the subtask scores for region and unk. By just adding these locations to the NLToolset's knowledge base, recall and precision was improved to 82 and 83 for the location object.

Another quirk of the MUC-7 domain was that adjectival forms of nation names were to be extracted as location objects, if they were the only references to the nation in the text. In other words, if the text contains the phrase "the Italian satellite" but no other mention of Italy, a location object with the locale "Italian" would be extracted. This was not addressed in AATM7 and resulted in a loss of thirty-two location objects, at three slots per object. This feature could be added just for the MUC-7 test. It is unlikely that a real-world application would want this information extracted. If it is added, recall and precision for the location object rise to 86 and 84 with an overall F-measure of 85.

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
SUBTASK SCORES														
location														
airport	21	18	1	0	17	3	0	2	**5**	**6**	14	0	94	95
city	226	226	197	0	9	20	20	2	**87**	**87**	9	9	4	20
country	260	239	221	0	10	29	8	7	**85**	**92**	11	3	4	18
province	52	53	42	0	6	4	5	19	**81**	**79**	8	9	13	26
region	89	40	29	0	11	49	0	4	**33**	**73**	55	0	28	67
unk	33	14	4	0	10	19	0	3	**12**	**29**	58	0	71	88
water	12	10	10	0	0	2	0	1	**83**	**100**	17	0	0	17
OBJ SCORES														
location	693	626	554	0	32	107	40	18	**80**	**88**	15	6	5	24
SLOT SCORES														
location														
locale	697	626	511	0	75	111	40	24	**73**	**82**	16	6	13	31
locale_type	693	600	504	0	63	126	33	38	**73**	**84**	18	6	11	31
country	691	624	493	0	90	108	41	26	**71**	**79**	16	7	15	33
obj_status	0	0	0	0	0	0	0	23	0	0	0	0	0	0
comment	0	0	0	0	0	0	0	52	0	0	0	0	0	0
ALL SLOTS	2081	1851	1508	0	228	345	115	163	**72**	**81**	17	6	13	31

	P&R	2P&R	P&2R
F-MEASURES	**76.70**	79.49	74.10

WALKTHROUGH MESSAGE

Our overall score for the walkthrough message is slightly below our overall performance.

*   *   *   **SUMMARY SCORES**   *   *   *

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
SUBTASK SCORES														
entity														
artifact	3	9	3	0	0	0	6	0	**100**	**33**	0	67	0	67
organization	23	23	20	0	3	0	0	0	**87**	**87**	0	0	13	13
person	10	12	10	0	0	0	2	0	**100**	**83**	0	17	0	17
location														
airport	0	0	0	0	0	0	0	0	**0**	**0**	0	0	0	0
city	9	8	8	0	0	1	0	0	**89**	**100**	11	0	0	11
country	6	6	6	0	0	0	0	0	**100**	**100**	0	0	0	0
province	1	1	1	0	0	0	0	2	**100**	**100**	0	0	0	0
region	3	2	2	0	0	1	0	0	**67**	**100**	33	0	0	33
unk	0	0	0	0	0	0	0	0	**0**	**0**	0	0	0	0
water	0	0	0	0	0	0	0	0	**0**	**0**	0	0	0	0
OBJ SCORES														
location	19	17	17	0	0	2	0	0	**89**	**100**	11	0	0	11
entity	36	44	33	0	3	0	8	0	**92**	**75**	0	18	8	25
SLOT SCORES														
location														

locale	19	17	16	0	1	2	0	1	84	94	11	0	6	16
locale_type	19	17	17	0	0	2	0	2	89	100	11	0	0	11
country	19	16	15	0	1	3	0	0	79	94	16	0	6	21
entity														
ent_name	41	46	28	0	7	6	11	0	**68**	**61**	15	24	20	46
ent_type	36	44	33	0	3	0	8	0	**92**	**75**	0	18	8	25
ent_descrip	19	25	12	0	4	3	9	17	**63**	**48**	16	36	25	57
ent_categor	36	44	29	0	7	0	8	0	**81**	**66**	0	18	19	34
ALL SLOTS	189	209	150	0	23	16	36	33	**79**	**72**	8	17	13	33

P&R   2P&R P&2R

F-MEASURES   **75.38** 73.17 77.72

Table 8: Walkthrough Scores

Persons

AATM7 found all of the persons in the walkthrough document. Of the five person descriptors, it missed only two; it made a separate entity for one of the descriptors and found only part of the other. The other spurious person entity is really an organization ("ING Barings") that was mistaken for a person, due to the fact that Ing is in the firstnames list. AATM7 did confuse another organization ("Bloomberg Business") as a person because of the context ("the parent of"), but this was marked incorrect, instead of spurious, because it was mapped to the organization object in the keys.

Organizations

Of the twenty-three organization entities, AATM7 found twenty-one. It missed "International Technology Underwriters" and "Axa SA." Two other organizations were typed incorrectly as people, as has been mentioned. Five of the nine organization descriptors were found correctly. The remaining error in the organization area is the result of the breaking of the variation linking mechanism that has been mentioned.

Artifacts

AATM7 correctly identified all three of the artifacts in the walkthrough article; however, because it overgenerated, precision for this object is a low 33%. This was due to the previously discussed mistake in which an organization that owned the satellite was incorrectly identified as the name. In fact, the organizations "Intelsat" and "United States" account for five of the six spurious artifacts. Two of the three descriptors were identified correctly.

Locations

AATM7 correctly identified sixteen of the nineteen locations, but missed "Arlington," "China," and the "Central" part of "Central America." This was due to overzealous context-based filtering.

CONCLUSIONS

A cursory analysis of AATM's MUC-7 scores revealed seven specific improvements to improve MUC-7 performance. Of these seven, five will be made in order to improve NLToolset performance. The sixth, adding extra-terrestrial bodies to the knowledge base, will be done to expand the NLToolset's reach. The seventh, making nation adjectives into locations, will not be done until a real-world application requires it.

If one were to make all of the changes specified, AATM7's overall scores would be improved to:

RECALL	PRECISION	F-MEASURE
83	78	80.42

The NLToolset continues to improve, as it is applied to new problems, whether real-world application or standardized test. Its accuracy remains high and its speed is constantly improving, currently standing, in its compiled state, at under twenty seconds for an average document.