CRL/NMSU
Description of the CRL/NMSU Systems Used for MUC-6

Jim Cowie
Computing Research Laboratory
New Mexico State University
jcowie@nmsu.edu
(505) 646-5181

Introduction

This paper discusses the two CRL named entity recognition systems submitted for MUC-6. The systems are based on entirely different approaches. The first is a data intensive method, which uses human generated patterns. The second uses the training data to develop decision trees which detect the start and end points of names.

Background

CRL submitted two systems for the Named Entity task. One of these (Basic) is an improved version of the CRL name recognizer developed in phase one of Tipster[1]. The second (AutoLearn) is a system which learns automatically from training data. The Basic system had approximately six man months of work in its original development. Improvements for MUC-6 were carried out by one graduate student (about one man month). The AutoLearn system was developed by one graduate student specifically for MUC-6 (also one man month).

Availability

The Basic system can be accessed for testing through the mail server - ne_tag@crl.nmsu.edu. the subject field of the message should consist of the word "tag". A web interface to the tagger is also available from CRL's home page. Source code and data files can be ftp'ed from CRL (mail cable@crl.nmsu.edu for ftp address and access password). The system has been successfully tested on Linux running on a PC.

Basic NE System

The Basic system consists of a pipeline of Unix process. These can be identified as carrying out four different types of task.

1. Recognizing names based on character patterns (numbers, dates).
2. Recognizing names based on pre-stored names.
3. Applying mark-up to potential components of complete names.
4. Recognizing names based on patterns of components.

For the most part the system is context free. A few of the patterns used require some additional context before a name is recognized. For example an ambiguous human name in isolation may be recognized if it is followed closely by a title.

The system consists of suite of 'C' and lex programs.

Component units recognized by the system are cities, provinces, countries, company prefixes and suffixes, company beginning and ending words (Club, Association etc.), unambiguous and ambiguous human first and last names, titles and human position words.

Additional Fix-up Procedures

Final patterns are used to join together units of the same type which are immediately next to each other in the text.

After all the pattern based procedures have operated on the text a final pass is made to recognize abbreviated forms of names. This takes the lists of names found so far and truncates them. (right to left for person and left to right for companies). These new lists are then used as lists of known organizations and persons and any occurrences of these in the text are marked. In particular for organizations the headline is not processed apart from this last stage. This avoids recognition of organizations such as "Leaves Bank". The assumption that names mentioned in the heading will be repeated in the body of the text holds almost universally.

Data Sources

The data used in the Basic system is derived from public domain source, university phone lists, the Tipster Gazetteer and government data-bases of company names.

Performance

The performances of the for the test set and for the walk through article are given in Appendix A. Overall performance was Recall - 85% and Precision - 87% giving an F measure of 85.8.

Walk through article

Performance here was Recall - 63% and Precision - 83%.

The main source of error was missing patterns in the system. For example Robert L. James was partially recognized (as L. James), McCann-Erickson was missed as no hyphenated company pattern had been added. Once a frequently mentioned name is ignored in its full form the system unfortunately misses all abbreviated forms. This article also shows the importance of context in reliably recognizing some names (e.g. an analyst with PayneWebber).

AutoLearn NE System

The AutoLearn system was developed to explore the possibility of using simple learning algorithms to detect specific features in text. An implementation of Quinlan's ID3 Algorithm was used [2,3]. This algorithm constructs a decision tree which decides whether an element of a collection satisfies a property or not. Each element of a collection has a finite number of attributes each of which may take one of several values. Quinlan's original paper suggests the range of values of the attributes should be "small". In the case of the AutoLearn system the values are every word occurring in the training collection.

Collections for Name Recognition

In order to apply the ID3 algorithm the data needs to be structured into a collection, each member of which has specific values for a set of attributes and for each of which it is known whether the member has a specific property or not. For the name recognition problem the training data was converted in tuple of five words. Each tuple was marked as having the start (or end) of a specific type of proper name at the middle word of the tuple. This data can be easily generated from the training articles. Thus for the beginning of a person -

many differences between Robert L. -1

differences between Robert L. James 1

between Robert L. James , -1

etc.

Fourteen sets of training data were generated using the 318 development articles supplied for MUC-6. The quality of the tagging is not particularly uniform, but no attempt has been made to improve this.

Generating the decision trees

As each word of the training data is read it is hashed and stored in a hash table. Thereafter words are referred to by their hash values. For each of the values of the five attributes (words 1 through 5) a count is maintained of the number of times this value contributed to an element holding a proper named occurrence at the middle attribute. The attribute to be tested first is chosen by computing for each value the relative frequency of positive and negative outcomes for this value. This is used to approximate the information content of that attribute

$$-p^{+}\log_2 p^{+} \; - p^{-}\log_2 p^{-} \qquad \text{(EQ 1)}$$

The sum of the approximate information contents for each column is calculated and the column with the highest value is chosen as the primary decision. Here all the values which always contributed to a positive outcome are used as the primary decision. Values which are always negative are ignored (this is primarily to reduce the size of the data being handled). New sub-collections are formed with elements containing one value which contributed both to positive or negative outcomes are collected and the tree building process is repeated for each of these new collections.

The decision trees thus formed can be output in a readable if somewhat lengthy form. In most cases the first choice is the third word in a group taking one of a large number of values. Thereafter a group off fairly impenetrable tests occur. For example for location beginnings -

> If word 3 is one of the following - Milwaukee Ridgefild Pa ST.. (around 300 more words) then location_beginning
>
> else if word 3 is Illinois and word 1 is Indiana then location_beginning
>
> else if word 3 is Northeast and word 1 is 'in' then location_beginnning

The printed decision table takes about 5 pages.

Running the AutoLearn System

A pass through the texts is made for each decision tree (beginning and end) of each named entity.

First the hash table of words is read and the corresponding decision tree. The text is then processed in groups of five words. Whenever a positive decision is made a new tag is added to the output stream.

Ideally at this stage the tagging would be done. However, given that we are processing new texts, there are many occasions where an end or a beginning is identified, but the corresponding beginning or end is not. For example a surname may have been seen previously, but not the attached forename. At this point a heuristic is applied which for every un-matched bracket in the text works forward or backward until some appropriate point is reached. The actual skipping heuristics need to be different for organizations, persons, locations, dates and numbers.

Data Sources

The only data source used for the AutoLearn system was the 318 MUC-6 training texts.

Performance

A high precision was expected from this system. Most of the errors that occur are due to failures of the bracket insertion heuristic. The overall scores were Recall - 47% and Precision- 81% giving an F measure of 59.3.

No specific code was inserted to handle numbers or dates. The method was more successful with organizations and locations then with persons. More training data is perhaps required to make the system aware of the spread of examples for human names.

Walk through article

The performance here was Recall - 36% and Precision - 88%.

The major problem here is that the system has not learned a rule which uses "Mr." to identify the word previous to a name.

Relationship of Performance to Amount of Training Data.

The evaluation texts were processed with decision trees generated using subsets of the MUC-6 development data. This was done in increasing units of 50 texts. The results are shown in Figure 1 below. Both recall and precision increase with increasing training data. Precision appears to tail off at around 82%. Recall, however, increases (with one exception) steadily over the whole range.

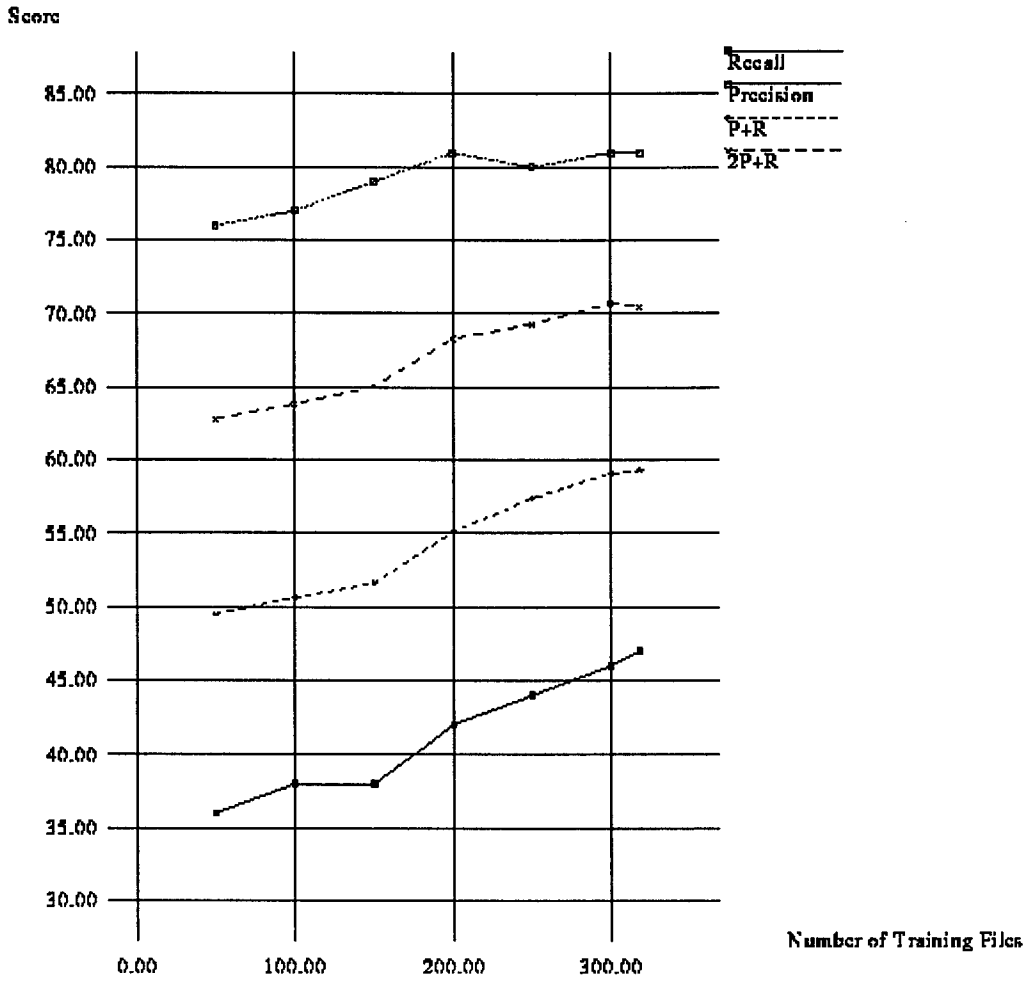

FIGURE 1. Relationship of Performance to Amount of Training

Future Developments

We intend to rebuild the Basic system. One of the principle drawbacks of the system is its sequential application of component tags. In many cases a second tag is not applied because the word or

phrase is ambiguous. The correct solution here is to apply all tags in a manner that allows the correct tags to be selected by the pattern processing mechanisms. In addition we plan to improve our collection of patterns. The current version of the system is being made generally available. This, we hope, will provide us with some feedback on patterns and errors in the data files.

Some further experiments are also planned with the AutoLearn system. The main drawback with the system is that it doesn't make maximal use of the training data in so far that with small training samples one word may be sufficient to make a decision. This situation can probably be improved by replacing specific words with a *NULL word.* This will force he system to develop rules based more on context. In particular when the system encounters unknown words these will be considered equivalent to the *NULL word.*

We also intend to apply the learning method described here to other NLP tasks such as part of speech tagging and disambiguation.

References

[1] Cowie, J.,Guthrie, L., Pustejovsky, J., Waterman, S., and Wakao, T., The CRL/Bradneis System us Used for MUC-5 In *Proceedings of the Fifth Message Understanding Conference (MUC-5)* Baltimore, Ma., Morgan Kaufmann, 1993.

[2] Quinlan, J.R. Discovering Rules by Induction from Large Collections of Examples.In *Expert Systems in the Micro-Electronic Age*, ed Michie D., Edinburgh University Press, 1979.

[3] Quinlan, J.R. Machine Learning: Easily Understood Decision Rules.In *Computer Systems that Learn*, eds. Weiss S.M. and Kulikowski C.A., Morgan Kaufmann, 1991.

Appendix A - Basic System Scores

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
<enamex>	925	889	841	0	0	48	84	0	91	95	9	5	14	0
type	925	889	784	0	57	48	84	0	85	88	9	5	19	7
text	925	889	755	0	86	48	84	0	82	85	9	5	22	10
subtotals	1850	1778	1539	0	143	96	168	0	83	86	9	5	21	8
<timex>	111	108	102	0	0	6	9	0	92	94	8	6	13	0
type	111	108	102	0	0	6	9	0	92	94	8	6	13	0
text	111	108	92	0	10	6	9	0	83	85	8	6	21	10
subtotals	222	216	194	0	10	12	18	0	87	90	8	6	17	5
<numex>	93	102	91	0	0	11	2	0	98	89	2	11	12	0
type	93	102	91	0	0	11	2	0	98	89	2	11	12	0
text	93	102	88	0	3	11	2	0	95	86	2	11	15	3
subtotals	186	204	179	0	3	22	4	0	96	88	2	11	14	2
ALL OBJECTS	2258	2198	1912	0	156	130	190	0	85	87	8	6	20	8
MATCHED ONLY	2068	2068	1912	0	156	0	0	0	92	92	0	0	8	8

P&R	2P&R	P&2R
F-MEASURES	85.82	86.52	85.13

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
Enamex:														
organization	442	392	330	0	40	22	72	0	75	84	16	6	29	11
person	373	378	353	0	9	16	11	0	95	93	3	4	9	2
location	110	119	101	0	8	10	1	0	92	85	1	8	16	7
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Timex:														
date	111	108	102	0	0	6	9	0	92	94	8	6	13	0
time	0	0	0	0	0	0	0	0	*	*	*	*	*	*
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Numex:														
money	76	77	74	0	0	3	2	0	97	96	3	4	6	0
percent	17	25	17	0	0	8	0	0	100	68	0	32	32	0
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*

* * * DOCUMENT SECTION SCORES * * *

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
HL	136	130	115	0	11	4	10	0	84	88	7	3	18	9
DD	60	60	60	0	0	0	0	0	100	100	0	0	0	0
DATELINE	52	50	49	0	1	0	2	0	94	98	4	0	6	2
TXT	2010	1958	1688	0	144	126	178	0	84	86	9	6	21	8

Appendix A - Autolearn System Scores

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
<enamex>	915	514	469	0	0	45	446	0	51	91	49	9	51	0
type	915	514	445	0	24	45	446	0	49	86	49	9	54	5
text	915	514	371	0	98	45	446	0	40	72	49	9	61	21
subtotals	1830	1028	816	0	122	90	892	0	44	79	49	9	58	13
<timex>	111	65	59	0	0	6	52	0	53	91	47	9	50	0
type	111	65	59	0	0	6	52	0	53	91	47	9	50	0
text	111	65	48	0	11	6	52	0	43	74	47	9	59	19
subtotals	222	130	107	0	11	12	104	0	48	82	47	9	54	9
<numex>	93	72	65	0	0	7	28	0	70	90	30	10	35	0
type	93	72	65	0	0	7	28	0	70	90	30	10	35	0
text	93	72	63	0	2	7	28	0	68	88	30	10	37	3
subtotals	186	144	128	0	2	14	56	0	69	89	30	10	36	2
ALL OBJECTS	2238	1302	1051	0	135	116	1052	0	47	81	47	9	55	11
MATCHED ONLY	1186	1186	1051	0	135	0	0	0	89	89	0	0	11	11

	P&R	2P&R	P&2R
F-MEASURES	59.38	70.57	51.25

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
Enamex:														
organization	433	287	251	0	8	28	174	0	58	87	40	10	46	3
person	372	155	131	0	14	10	227	0	35	84	61	6	66	10
location	110	72	63	0	2	7	45	0	57	88	41	10	46	3
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Timex:														
date	111	65	59	0	0	6	52	0	53	91	47	9	50	0
time	0	0	0	0	0	0	0	0	*	*	*	*	*	*
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Numex:														
money	76	55	52	0	0	3	24	0	68	94	32	5	34	0
percent	17	17	13	0	0	4	4	0	76	76	24	24	38	0
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*

* * * DOCUMENT SECTION SCORES * * *

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
HL	128	92	65	0	0 7	20	56	0	51	71	44	22	56	10
DD	60	0	0	0	0	0	60	0	0	*	100	*	100	*
DATELINE	52	24	24	0	0	0	28	0	46	100	54	0	54	0
TXT	1998	1186	962	0	128	96	908	0	48	81	45	8	54	12

Appendix A - Basic System Walk-through Scores

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
<enamex>	66	52	47	0	0	5	19	0	71	90	29	10	34	0
type	66	52	42	0	5	5	19	0	64	81	29	10	41	11
text	66	52	42	0	5	5	19	0	64	81	29	10	41	11
subtotals	132	104	84	0	10	10	38	0	64	81	29	10	41	11
<timex>	6	5	5	0	0	0	1	0	83	100	17	0	17	0
type	6	5	5	0	0	0	1	0	83	100	17	0	17	0
text	6	5	5	0	0	0	1	0	83	100	17	0	17	0
subtotals	12	10	10	0	0	0	2	0	83	100	17	0	17	0
<numex>	6	7	6	0	0	1	0	0	100	86	0	14	14	0
type	6	7	6	0	0	1	0	0	100	86	0	14	14	0
text	6	7	6	0	0	1	0	0	100	86	0	14	14	0
subtotals	12	14	12	0	0	2	0	0	100	86	0	14	14	0
ALL OBJECTS	156	128	106	0	10	12	40	0	68	83	26	9	37	9
MATCHED ONLY	116	116	106	0	10	0	0	0	91	91	0	0	9	9

P&R	2P&R	P&2R
F-MEASURES	74.65	79.34	70.48

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
Enamex:														
organization	29	12	5	0	5	2	19	0	17	42	66	17	84	50
person	35	38	35	0	0	3	0	0	100	92	0	8	8	0
location	2	2	2	0	0	0	0	0	100	100	0	0	0	0
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Timex:														
date	6	5	5	0	0	0	1	0	83	100	17	0	17	0
time	0	0	0	0	0	0	0	0	*	*	*	*	*	*
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Numex:														
money	5	6	5	0	0	1	0	0	100	83	0	17	17	0
percent	1	1	1	0	0	0	0	0	100	100	0	0	0	0
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*

* * * DOCUMENT SECTION SCORES * * *

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
HL	8	6	6	0	0	0	2	0	75	100	25	0	25	0
DD	2	2	2	0	0	0	0	0	100	100	0	0	0	0
DATELINE	0	0	0	0	0	0	0	0	*	*	*	*	*	*
TXT	146	120	98	0	10	12	38	0	67	82	26	10	38	9

Appendix A - Autolearn System Walk-through Scores

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
<enamex>	69	22	20	0	0	2	49	0	29	91	71	9	72	0
type	69	22	20	0	0	2	49	0	29	91	71	9	72	0
text	69	22	18	0	2	2	49	0	26	82	71	9	75	10
subtotals	138	44	38	0	2	4	98	0	28	86	71	9	73	5
<timex>	6	5	5	0	0	0	1	0	83	100	17	0	17	0
type	6	5	5	0	0	0	1	0	83	100	17	0	17	0
text	6	5	4	0	1	0	1	0	67	80	17	0	33	20
subtotals	12	10	9	0	1	0	2	0	75	90	17	0	25	10
<numex>	6	6	6	0	0	0	0	0	100	100	0	0	0	0
type	6	6	6	0	0	0	0	0	100	100	0	0	0	0
text	6	6	5	0	1	0	0	0	83	83	0	0	17	17
subtotals	12	12	11	0	1	0	0	0	92	92	0	0	8	8
ALL OBJECTS	162	66	58	0	4	4	100	0	36	88	62	6	65	6
MATCHED ONLY	62	62	58	0	4	0	0	0	94	94	0	0	6	6

	P&R	2P&R	P&2R
F-MEASURES	50.88	68.08	40.62

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
Enamex:														
organization	32	12	12	0	0	0	20	0	38	100	62	0	62	0
person	35	7	6	0	0	1	29	0	17	86	83	14	83	0
location	2	3	2	0	0	1	0	0	100	67	0	33	33	0
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Timex:														
date	6	5	5	0	0	0	1	0	83	100	17	0	17	0
time	0	0	0	0	0	0	0	0	*	*	*	*	*	*
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*
Numex:														
money	5	5	5	0	0	0	0	0	100	100	0	0	0	0
percent	1	1	1	0	0	0	0	0	100	100	0	0	0	0
other	0	0	0	0	0	0	0	0	*	*	*	*	*	*

* * * DOCUMENT SECTION SCORES * * *

SLOT	POS	ACT	COR	PAR	INC	SPU	MIS	NON	REC	PRE	UND	OVG	ERR	SUB
HL	8	2	2	0	0	0	6	0	25	100	75	0	75	0
DD	2	0	0	0	0	0	2	0	0	*	100	*	100	*
DATELINE	0	0	0	0	0	0	0	0	*	*	*	*	*	*
TXT	152	64	56	0	4	4	92	0	37	88	60	6	64	7