ARMY RESEARCH LABORATORY

Extrinsic Evaluation of Automated Information Extraction Programs

by Frank Small and William Tanenbaum

ARL-TN-391 **May 2010**

prepared by

U.S. Army Research Laboratory
Aberdeen Proving Ground, MD

under contract

W81XWH-09-D-0081

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-TN-391 **May 2010**

Extrinsic Evaluation of Automated Information Extraction Programs

Frank Small and William Tanenbaum
Computational and Information Sciences Directorate, ARL

prepared by

U.S. Army Research Laboratory
Aberdeen Proving Ground, MD

under contract

W81XWH-09-D-0081

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2010	Final	June 2009–August 2009

4. TITLE AND SUBTITLE
Extrinsic Evaluation of Automated Information Extraction Programs

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Frank Small and William Tanenbaum

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-C
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-391

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Information extraction (IE) plays a vital role in Natural Language Processing and also serves as the foundation for computational visualization of information. Information can be converted into a user-defined, ontology-friendly format much faster and more efficiently by automating IE. Programs like General Architecture for Text Engineering (GATE) and Automap allow entities such as name, date, location, and organization to be extracted from a corpus written in a natural language. Verbs and other parts of speech can be extracted using these programs as well. The extracted information can then be formatted into a computer-readable language for visualization and populating a database for use by the fusion community to provide actionable intelligence for the Warfighter. This technical note documents the results of the comparison of the IE tools offered by GATE versus those in Automap.

15. SUBJECT TERMS
information extraction, Automap, GATE, precision, recall, natural language processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ann Bornstein
Unclassified	Unclassified	Unclassified	UU	16	19b. TELEPHONE NUMBER (*Include area code*) 410-278-8947

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Tables

Acknowledgments

We would like to give our most sincere thanks to our mentors—Mark Thomas and Ann Bornstein—for their help and guidance. We would also like to thank Fred Brundick for his expertise with computer code and text-splitting capabilities and Gary Moss for his assistance throughout the project. Our thanks also to Michelle McVey and Brianna Larrimore for reviewing and editing our papers.

INTENTIONALLY LEFT BLANK.

1. Introduction

The capability to extract particular pieces of information from a data set while maintaining both high precision and recall is difficult and time-consuming. The need for faster information extraction (IE) without significant loss of accuracy has lead to the creation of automated IE programs. Empirical evaluation plays a key role in estimating the performance of promising IE tools.

General Architecture for Text Engineering (GATE) and Automap are two such promising IE tools; the former was developed by the Natural Language Processing (NLP) Group of the University of Sheffield and the latter by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. Both programs have similar named entity, date, and location extraction capabilities.

Comparison of the two programs was based on functionality, usability, customization with empirical performance evaluation keyed to the one-dimensional metrics precision, recall, and the F-measure. Each program was evaluated against three independent corpora: the Database Creation for Information Processing Methods, Metrics, and Models (DCIPM3) message set, the Soft Target Exploitation and Fusion Human Intelligence (STEF HUMINT) message set, and a Google message set.

Extracted information can be visualized or formatted and stored as Resource Descriptive Framework (RDF) triples for later use in the construction of an ontology. The result would be an information system that provides fast and actionable intelligence to the Warfighter.

2. Automated IE Tools

2.1 GATE

GATE v.5.0 is an IE open-source program. The program features a user-friendly graphical user interface (GUI) that was used for this project.

2.1.1 Graphical User Interface (GUI)

Many of GATE's features could be utilized via the GUI. A few syntactical and visual errors were encountered when using the GUI, but overall it was a helpful aid to the extraction process.

Pipelines,* Processing Resources,† and Language Resources‡ could be loaded into GATE with minimal effort using the GUI. Various resources could be implemented to assist in the extraction effort given the capability to load plug-ins into GATE by accessing the plug-in manager from the GUI.

The GUI allowed a corpus to be saved as an XML document, with the option of including annotations. This was a useful capability.

2.1.2 Annotations

GATE displays extracted information by annotation. Each annotation is classified as a type, such as date, location, organization, person, etc. Each annotation includes features such as type, gender, kind, rules, etc. A user can also manually add features to any annotation. These features will appear alongside the annotations when saved as an XML document.

A pre-defined pipeline named ANNIE, which can be loaded into GATE, uses a variety of processing resources located in GATE's plug-in manager to automatically annotate a corpus. ANNIE analyzes a corpus by tokenizing the text,§ running a gazetteer,** and transducing.†† Once ANNIE has analyzed the text, the annotation types which have been created are compiled into an annotation set. While viewing a document, a user can select the annotation types to be viewed. The selected annotations will be highlighted in the text.

A user can also create and edit annotations manually. This is a necessary step to calculate precision and recall since manually annotated corpora, which serve as the ground truth, must be compared to the automatically annotated corpora. Creating and editing annotations is a long, grueling process in GATE, mainly caused by a few specific yet annoying errors.

2.2 Automap

Automap v.2.7.4 takes a different approach to automated IE than GATE. A corpus loaded into Automap undergoes numerous preprocessing steps. Key preprocessing tools may include stemming functions, deletion, thesauri, and removing symbols, numbers, and punctuation. Once preprocessing is complete, the user tags the remaining concepts in a meta-matrix thesaurus then selects and applies a sub-matrix. The extracted information is then ready for output.

2.2.1 Text Preprocessing

Application of a delete list and generalization thesaurus was sufficient to streamline the data for this study.

---

*An application built from different processing resources to process language resources.

†Sub-processes that can be used to build a pipeline.

‡Corpora, single documents, etc.

§Splits texts into tokens such as word, punctuation, space, etc.

**Sets of lists containing names of entities such as cities, organizations, days of the week, etc.

††Implementing grammars (patterns) to define entities not defined by gazetteer.

2.2.1.1 Deletion. Deletion was a tricky process—clarity and concision had to be weighed against losing potentially valuable information. The delete list included with Automap was a very basic list of 35 words, the majority of which were pronouns and prepositions. A far larger, customized delete list was written to pare the number of concepts. Since the project involved training Automap on only one of the three corpora, the STEF HUMINT message set, the delete list was aimed at removing all but military-relevant words.

2.2.1.2 Thesauri. Two thesauri found in Automap were utilized. The first, the Generalization Thesaurus, allowed two or more concepts to be indentified as identical and provided a common concept label. For instance, this function identified JCOC as the Joint Civilian Orientation Conference and inserted a new symbol grouping, Joint_Civilian_Orientation_Conference, into the processed document. In this way, the concept is processed as a single concept and not 3 pairs of concepts. The Generalization Thesaurus had several other uses, it: (1) turned plural concepts into their singular form, (2) converted dialectal spelling differences, and (3) changed similar concepts into the same concept (e.g., "street," "roadway," "thoroughfare," and "throughway" were all deemed synonymous with "road").

The second and, perhaps, more important thesaurus, was the Meta-Matrix Thesaurus. Words are tagged in this Thesaurus according to the program's embedded ontology; that is, this thesaurus associates text-level concepts with meta-matrix concepts. AutoMap's ontology offers several classifications for any given concept including knowledge, agent, resource, event, organization, location, when, and attribute.

Output from Automap can be analyzed at any of three levels: (1) the concept network level, (2) the entire meta-matrix level, or (3) the sub-matrix level. The sub-matrix controls how the different classes of Automap's ontology interact with one another. The full sub-matrix was selected for this study to ensure every class interaction was displayed.

3. Performance Evaluation

To objectively compare the two IE programs, the single-dimension quantitative metrics precision, recall and the traditional F-measure were selected. Precision is the proportion of documents retrieved that are relevant to a user's information needs; precision takes all retrieved documents into account. It is defined as

$$\text{Precision (P)} = tp/(tp + fp), \quad (1)$$

where tp* and fp† are the numbers of true positive and false positive, respectively.

*Document is retrieved by system and is relevant.

†Document is retrieved by system and is not relevant.

Recall is the proportion of successfully retrieved documents that are relevant to the user's query; recall corresponds to the true positive rate. It is defined by

$$\text{Recall (R)} = tp / (tp + fn), \tag{2}$$

where tp is defined as in equation 1 and fn* is the number of false negative.

The F-measure can be interpreted as the (equally) weighted harmonic mean of precision and recall; the F-measure is defined as

$$F = 2 \bullet [(P \bullet R) / (P + R)]. \tag{3}$$

There are several methods to measure precision and recall depending on how strictly or leniently partially correct true positives are taken into account. A partially correct true positive occurs when a piece of information is not correctly extracted, such as not including a full name or adding words which aren't part a name. The "strict"† method considers partially correct true positives as false positives. The "lenient"‡ method considers partially true correct positives as true positives. The third method uses the mean of the strict and lenient methods to compute the overall precision and recall measures. The third method was selected to evaluate GATE and Automap.

4. Results

Table 1 gives a summary of the statistical results.

Table 1. Precision, recall, and F-measure statistics for the three corpora.

	GATE Version 5.0			**Automap Version 2.7**		
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
DCIPM3 Msg Set						
Date/when	1.000	0.967	0.983	1.000	0.484	0.652
Location	0.750	0.050	0.094	0.333	0.033	0.061
Organization	0.750	0.214	0.333	0.250	0.071	0.111
Person/agent	0.550	0.668	0.604	0.475	0.340	0.400
Google Msg Set						
Date/when	0.989	0.989	0.989	0.662	0.179	0.282
Location	0.966	0.664	0.787	0.954	0.838	0.893
Organization	0.968	0.724	0.828	0.822	0.751	0.785
Person/agent	0.648	0.805	0.718	0.687	0.658	0.672
STEF HUMINT Msg Set						
Date/when	0.690	0.794	0.738	0.675	0.214	0.325
Location	0.947	0.471	0.629	0.680	0.439	0.533
Organization	0.789	0.366	0.500	0.784	0.500	0.612
Person/agent	0.333	0.497	0.358	0.818	0.248	0.381

*Document is not retrieved by system but is relevant.

†http://gate.ac.uk/sale/tao/split.html.

‡http://gate.ac.uk/sale/tao/split.html.

The numbers of correct, partially correct, false negatives, and false positives used for precision, recall, and F-Measure calculations are provided in table 2.

Automap was trained on the STEF HUMINT message set for this study. Since both the STEF HUMINT and Google message sets featured a military specific domain, Automap performed reasonably well for most entity types. Automap performed poorly with respect to the DCIPM3 message set, which features a high school environment domain. GATE, however, did not appear to be encumbered by the domain type and had higher or equivalent precision, recall, and F-Measure statistics when compared to Automap.

Table 2. Correct (C), partially correct (PC), false negatives (FN), and false positives (FP).

	GATE Version 5.0				**Automap Version 2.7**			
	C	**PC**	**FN**	**FP**	**C**	**PC**	**FN**	**FP**
DCIPM3 Msg Set								
Date/when	58	0	2	0	30	0	32	0
Location	1	1	28	0	1	0	29	2
Organization	1	1	5	0	0	1	6	1
Person/agent	74	111	9	50	0	132	62	7
Google Msg Set								
Date/when	135	1	1	1	12	25	100	0
Location	254	6	127	6	312	25	50	3
Organization	237	8	88	4	201	98	34	5
Person/agent	68	17	10	33	36	53	6	2
STEF HUMINT Msg Set								
Date/when	90	20	16	35	14	26	86	0
Location	114	3	128	5	58	99	88	1
Organization	14	2	25	3	15	11	15	0
Person/agent	32	58	55	82	28	16	101	0

5. Conclusions

GATE has more potential than Automap as an automated IE program. GATE offers a fully automated process, while Automap needs a fair amount of user definition. The performance metrics for GATE, for the most part, are equivalent or better than Automap's. Automap performed significantly worse at extraction when dealing with concepts it was not trained on, which is a large downfall when dealing with domain specific IE.

Bibliography

CASOS. http://www.casos.cs.cmu.edu/ (accessed 11 August 2009).

Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V.; Ursu, C.; Dimitrov, M.; Dowman, M.; Aswani, N.; Roberts, I.; Li, Y.; Shafirin, A.; Funk, A. Developing Language Processing Components with GATE Version 5 (a User Guide). The University of Sheffield. http://www.gate.ac.uk/sale/tao/split.html (accessed June–August 2009).

Mooney, R.; Cooper, J.; Ghosh, J.; Lee, D. Performance Evaluation of Information Retrieval Systems. Powerpoint Presentation for Information Retrieval and Web Search Class (accessed 14 July 2009).

NO. OF
COPIES ORGANIZATION

1 (PDF only) DEFENSE TECHNICAL INFORMATION CTR
DTIC OCA
8725 JOHN J KINGMAN RD
STE 0944
FORT BELVOIR VA 22060-6218

1 DIRECTOR
US ARMY RESEARCH LAB
IMNE ALC HRR
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIM L
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIM P
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL D
2800 POWDER MILL RD
ADELPHI MD 20783-1197

ABERDEEN PROVING GROUND

1 DIR USARL
RDRL CIM G (BLDG 4600)

NO. OF
COPIES ORGANIZATION

1 DIRECTOR
US ARMY RSRCH LAB
RDRL CII
B BROOME
2800 POWDER MILL RD
ADELPHI MD 20783-1197

2 DIRECTOR
US ARMY RSRCH LAB
RDRL CII T
L HERNANDEZ
M VANNI
2800 POWDER MILL RD
ADELPHI MD 20783-1197

ABERDEEN PROVING GROUND

17 DIR USARL
RDRL CII C
A BORNSTEIN (15 CPS)
M THOMAS
D WELSH