REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
07-01-2016	Final Report	18-Jun-2012 - 30-Jun-2013

4. TITLE AND SUBTITLE
Final Report: Mathematical Modelling for the Evaluation of Automated Speech Recognition Systems -- Research Area 3.3.1 (c)

5a. CONTRACT NUMBER
W911NF-12-1-0195

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Gerald Penn

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Toronto
McMurrich Building, 3rd floor
12 Queen's Park Crescent West
-

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62272-MA-II.2

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Automated speech recognizers (ASR) are now more often found as components inside other applications than as a standalone application for transcribing speech word-for-word into text. Statistical pattern recognition techniques allow us to acquire a better task-specific evaluation measure for embedded applications than word error rates (WER), which are used for transcription.

15. SUBJECT TERMS
speech recognition, evaluation, ablation, adaptive boosting, decision stumps, summarization, entity recognition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gerald Penn
UU	UU	UU	UU		19b. TELEPHONE NUMBER 416-978-7390

Report Title

Final Report: Mathematical Modelling for the Evaluation of Automated Speech Recognition Systems -- Research Area 3.3.1(c)

ABSTRACT

Automated speech recognizers (ASR) are now more often found as components inside other applications than as a standalone application for transcribing speech word-for-word into text. Statistical pattern recognition techniques allow us to acquire a better task-specific evaluation measure for embedded applications than word error rates (WER), which are used for transcription.

Our approach considered two applications of ASR: a decision support software system for meetings, in which a summary of a meeting is audited to record all of the decisions that were taken during the meeting, and a specific entity identification task, in which an intelligence analyst identifies triples of "who," "where" and "when" for each event described in transcribed broadcast news. Both of these resemble typical activities of intelligence analysts in OSINT processing and production applications.

We assessed two task evaluation measures. The first fixes the input, and learns to predict human subject performance as the transcript for the input varies in accuracy. This measure is well-suited to developers of ASR systems who wish to measure the effects of modifications they make to their software during development. The second measure does not hold the input fixed, and does not require new human-subject data to be collected for new input.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received	Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received	Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper	
01/02/2016	1.00	Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare Voss, Frauke Zeller. Automatic Human Utility Evaluation of ASR Systems: Does WER Really Predict Performance?, INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association . 25-AUG-13, . : ,

TOTAL: **1**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Aditya Bhargava	0.07	
FTE Equivalent:	**0.07**	
Total Number:	**1**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Cosmin Munteanu	0.25
FTE Equivalent:	**0.25**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Gerald Penn	0.20	
FTE Equivalent:	**0.20**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
Matthew Giamou	0.17	Engineering Science
FTE Equivalent:	**0.17**	
Total Number:	**1**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

(4) Statement of Problem studied

Speech recognition has classically been thought of as the task of transcribing speech word-for-word into running text. With the widespread availability of digital audio on computers, this task has almost no applications of its own apart from support for hearing-impaired users of audio media. It is more often the case that an automated speech recognizer (ASR), the software and digital signal processing hardware that performs this task, can be found as a component inside some other application, such as a speech-to-speech machine translation system, or a spoken information retrieval system.

Our means of evaluating ASR systems, however, have not changed with the times. The staple measure of success within the speech community is still "word error rate" (WER), a simple ratio that counts the number of edit operations required to transform an automatically generated transcript into the correct transcript per unit length (in words) of the correct transcript. The present research programme began with the hypotheses that (1) WER is a poor predictor of embedded ASR performance within more realistic applications, and that (2) statistical pattern recognition techniques would allow us to acquire a better task-specific evaluation measure.

Our approach to this subject was to consider two specific, realistic applications of ASR. The first is a decision support software system for meetings, in which a summary of a meeting is audited after the conclusion of the meeting by a non-participant, recording all of the decisions that were taken during the meeting. The second is a specific named-entity identification task in broadcast news speech, in which an intelligence analyst identifies triples of "who," "where" and "when" for each event described in transcribed broadcast news under timed conditions. Both of these tasks resemble typical activities of intelligence analysts in a number of OSINT processing and production applications.

(5) Summary of the most important results

In both settings, there is indeed the temptation to measure the success of speech recognition according to the WER of the transcripts that are used either by automated downstream applications (e.g., for summarization) or by human intelligence analysts (e.g., who/where/when identification). Using transcripts with controlled WERs, we examined performance in an ecologically more valid assessment that used unequivocal, task-specific measures of performance, and attempted to correlate these with the would-be predictions of WER. Our most important results have been: (1) a statistically tuned evaluation protocol that provides a far more accurate assessment of performance while minimizing human annotation effort in the assessment process, and (2) a validated human-computer interface for the extraction of event triples that was developed for the purpose of assessing the second task.

In our human-subject studies, we have measured a statistically significant difference in task performance as a function of WER, with significance being measurable relative to manual transcription beginning at roughly 30%. Note that this differs from the 20% threshold established by earlier research, beyond which it is easier to begin manually transcribing again from scratch than to manually correct the errors. This is definitive proof that there are transcripts that, although unusable as transcripts for the classical task of word-for-word reading, nevertheless can play an important role as embedded artefacts in more modern speech

applications.

What will probably more surprising, as shown in Attachment 1, is that to the extent that WER correlates with human-subject performance at all (rho = 0.017), it correlates positively. We have surmised that this is because, once WERs cross the significance threshold, reading erroneous transcripts word-for-word becomes so onerous that human subjects to a more keyword-spotting or browsing behaviour, which improves their performance at the tasks we studied. We have also determined that time limits on the tasks were not a determining factor in the polarity of the correlation. We tried various alternative formulations of WER, including the calculation of WER on only topic-specific keywords, but these alternatives only improved predictability to the extent that they increased the magnitude of the positive correlation, i.e., they are more strongly anti-correlated with accuracy.

In our investigation of better task-specific evaluation measures, we have explored two statistical models of two specific classes of measures. The first measure is trained on pairs of automatically generated transcripts for the same spoken audio input and task-specific performance measures collected from human subjects who work with that transcript. It is then deployed on new, unpaired automatically generated transcripts for the same task and input, and predicts task-specific performance. This is the more conservative measure, in that it predicts the outcome of using a newer, hopefully better ASR system on the same input. It is also a more labour-intensive measure, as it must be retrained when different input is used. That retraining involves the collection of new human-subject performance data, which is more expensive and time-consuming than calculating WER. Nevertheless, this measure is well-suited to developers of ASR systems who wish to measure the effects of modifications they make to their software using a corpus of test data annotated with human-subject performance, sometimes called development test data.

The second measure is trained on triples of spoken audio input, automatically generated transcripts, and task-specific performance measures collected from human subjects. It is then deployed on new spoken input paired with an automatically generated transcript, and predicts task-specific performance. This measure is less conservative and less labour-intensive. In this measure, the spoken input varies, and so the evaluation protocol must learn to react to variations in the input, as human subjects do. It does not require new human-subject data to be collected for new input, but instead merely new automated transcripts, which are comparatively inexpensive.

Taking the decision support application as a running example, the third page of the attached Interspeech conference proceedings paper lists the features that were used to characterize the transcript and acoustic input, together with an ablated experiment in which features were selectively removed to determine the overall effect on evaluation accuracy in the sense of the more conservative, first measure (Figure 3, Interspeech paper). That accuracy (Figure 2, Interspeech paper) has an equal precision/recall rate of roughly 55%. WER alone can be seen in the same figure to have an equal precision/recall rate of roughly 30%.

All results were obtained using an adaptive boosting classifier that iteratively searches for the best combination of one-level decision trees formed from the features listed on the third page. We have also experimented with the other pattern recognition methods found on the mlcomp.org reference website, a comparison of which is

shown in Attachment 2.

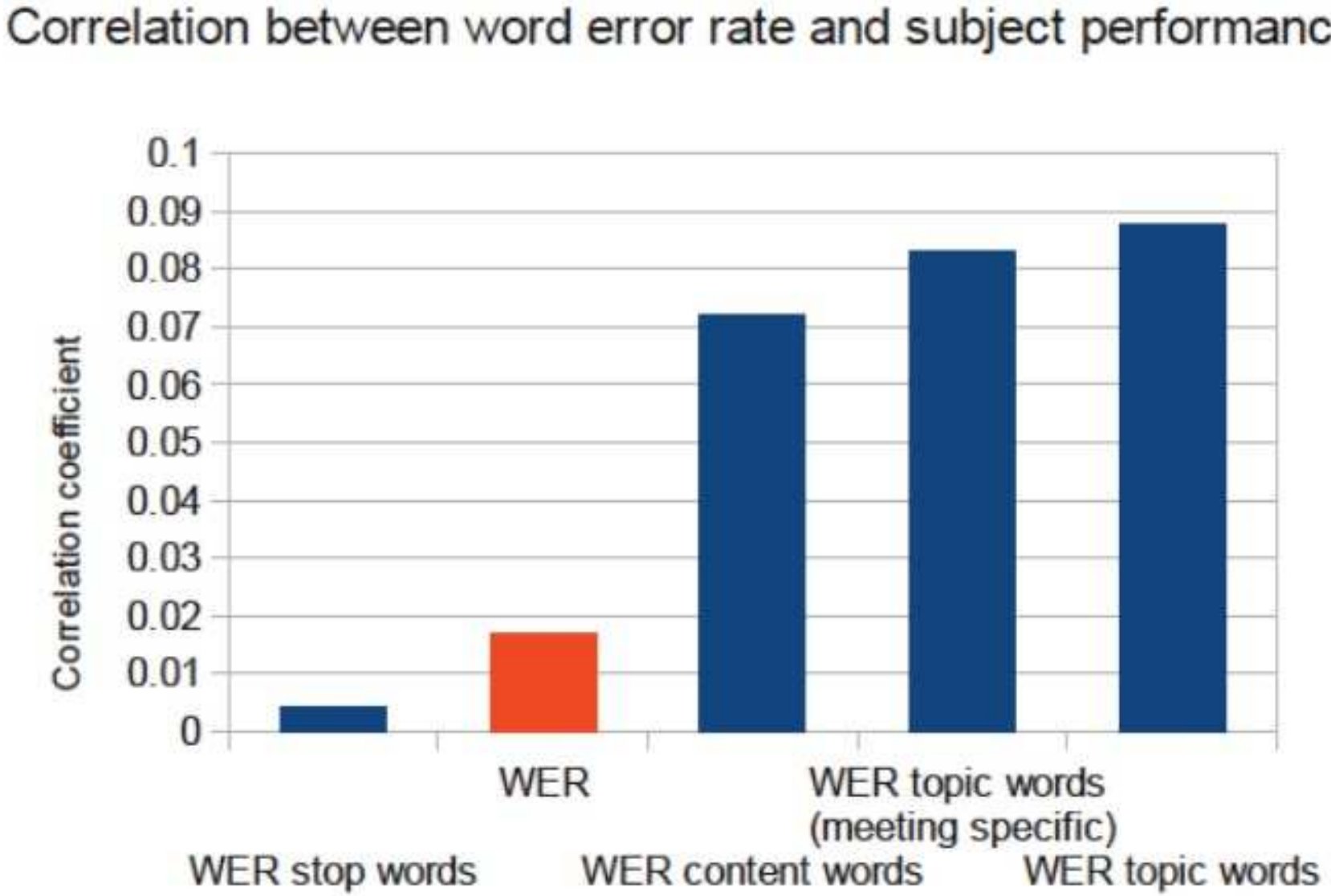
The second measure, using the same features, has an equal rate of roughly 40%, a number that has been observed to be stable across spoken audio inputs. WER is completely unsuited to this task, yielding less than 1%, as shown in the feature ablation graph for the second measure in Attachment 3. A comparison with the other pattern recognition methods found on mlcomp.org is shown in Attachment 4. In a separate experiment, we compared the second measure with and without the artificial addition of development test data analogous to those available for the first measure. With development test data, the F-measures are significantly better, approaching 45%. This reassures us that the addition of more data could be used to address the observed accuracy shortfall of the second measure relative to the first.

A screenshot of the human-computer interface for extracting event triples is shown in Attachment 5. It consists of three panes, containing, from left to right, a copy of the transcribed speech, a list of keywords/keyphrases extracted from the transcript, either automatically by entity identification software or using a mouse, and a list of who-where-when triples, again assembled either from the keyphrases by clicking and dragging, or automatically using relation identification software.

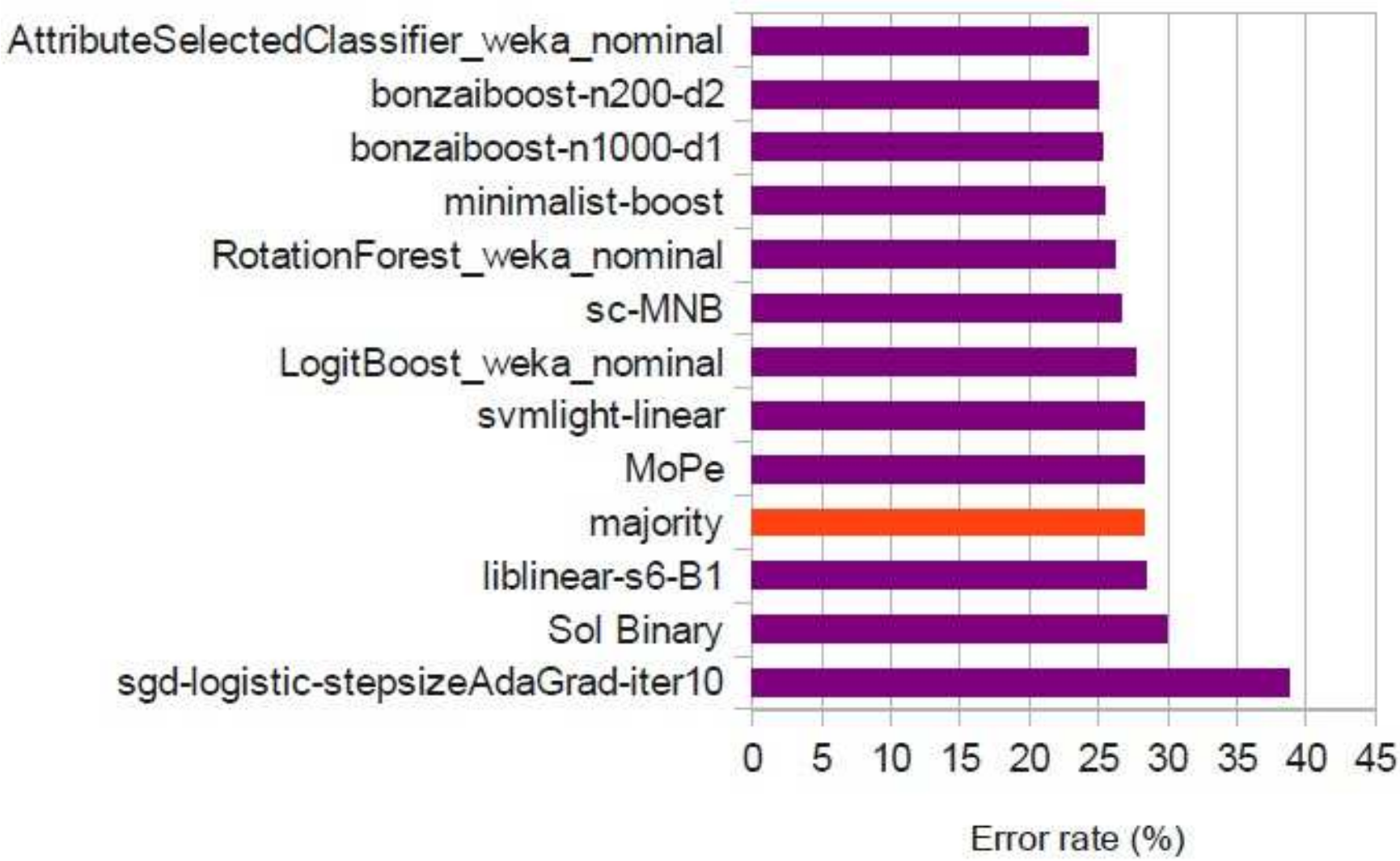
This interface has been validated first through a pre-piloting experiment with intelligence analysts, conducted at ARL by Drs. Clare Voss and Stephen Tratz in the summer of 2012, then again, after modifications arising from pre-piloting feedback, in a final test of the interface through Amazon Mechanical Turk with untrained crowdsource workers.

Technology Transfer

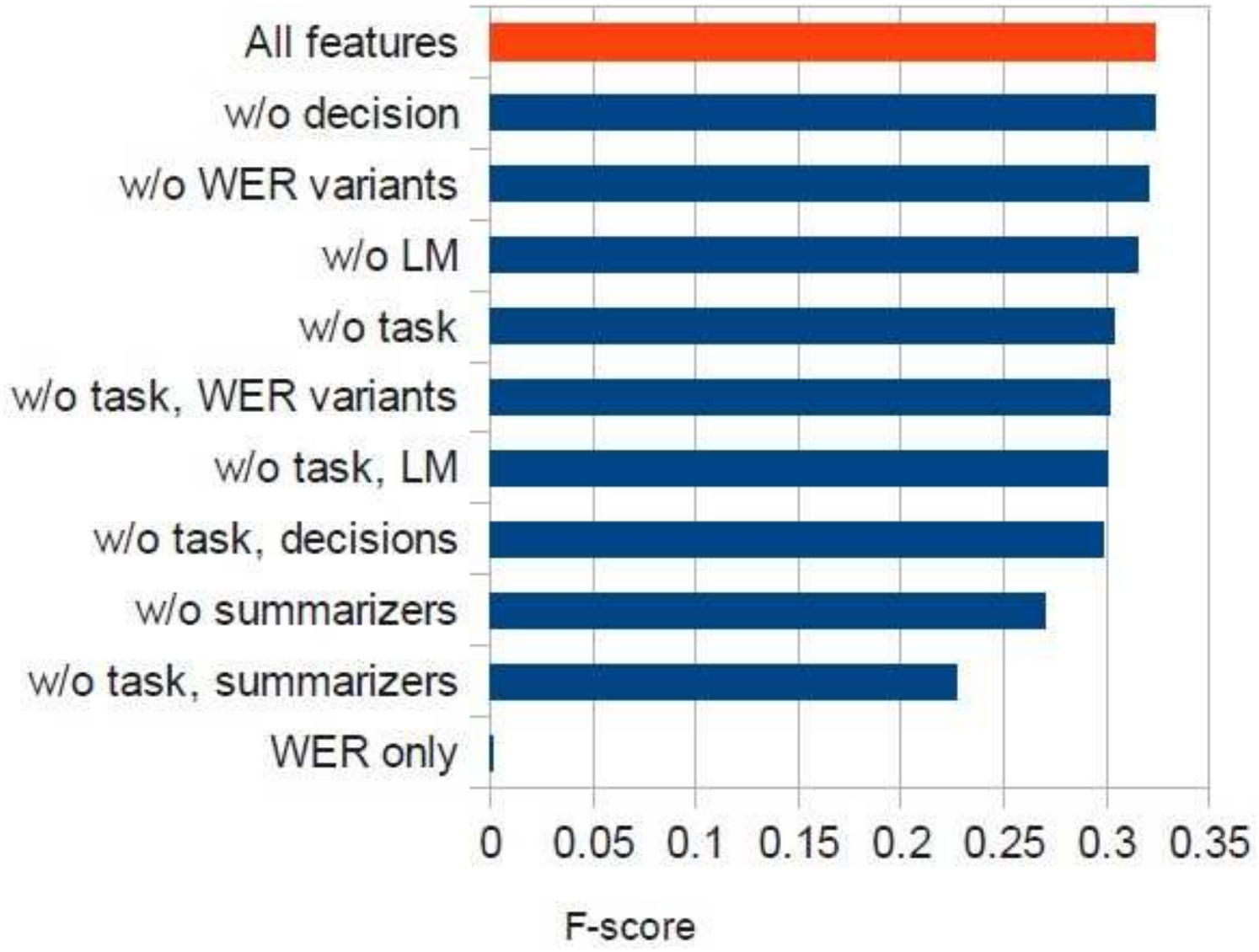
Extensive interaction with Dr. Clare Voss and colleagues, ARL, Beltsville, MD, on an interface for a named entity extraction task that identifies "who-where-when" triples from events described in broadcast news.



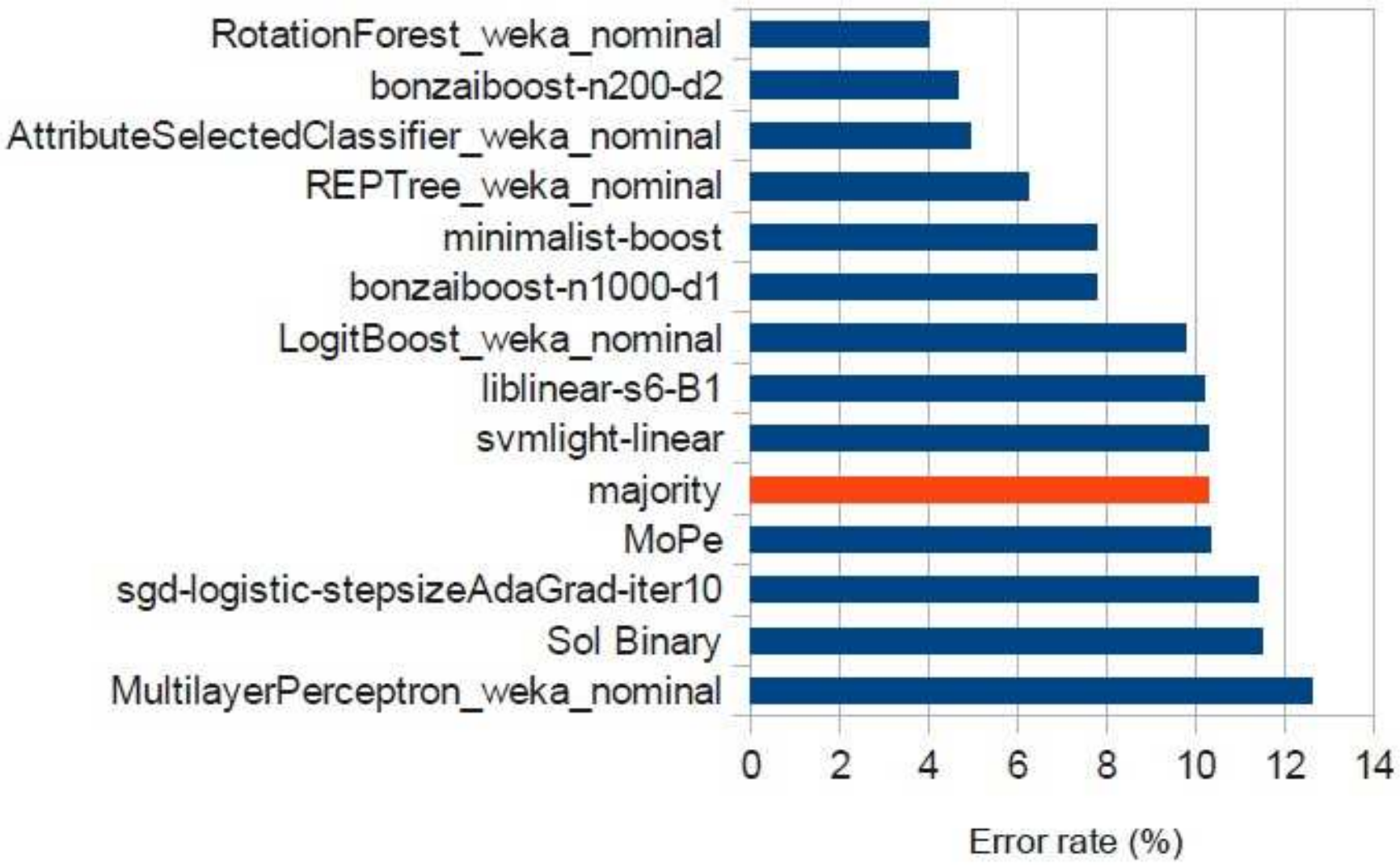
Attachment 1



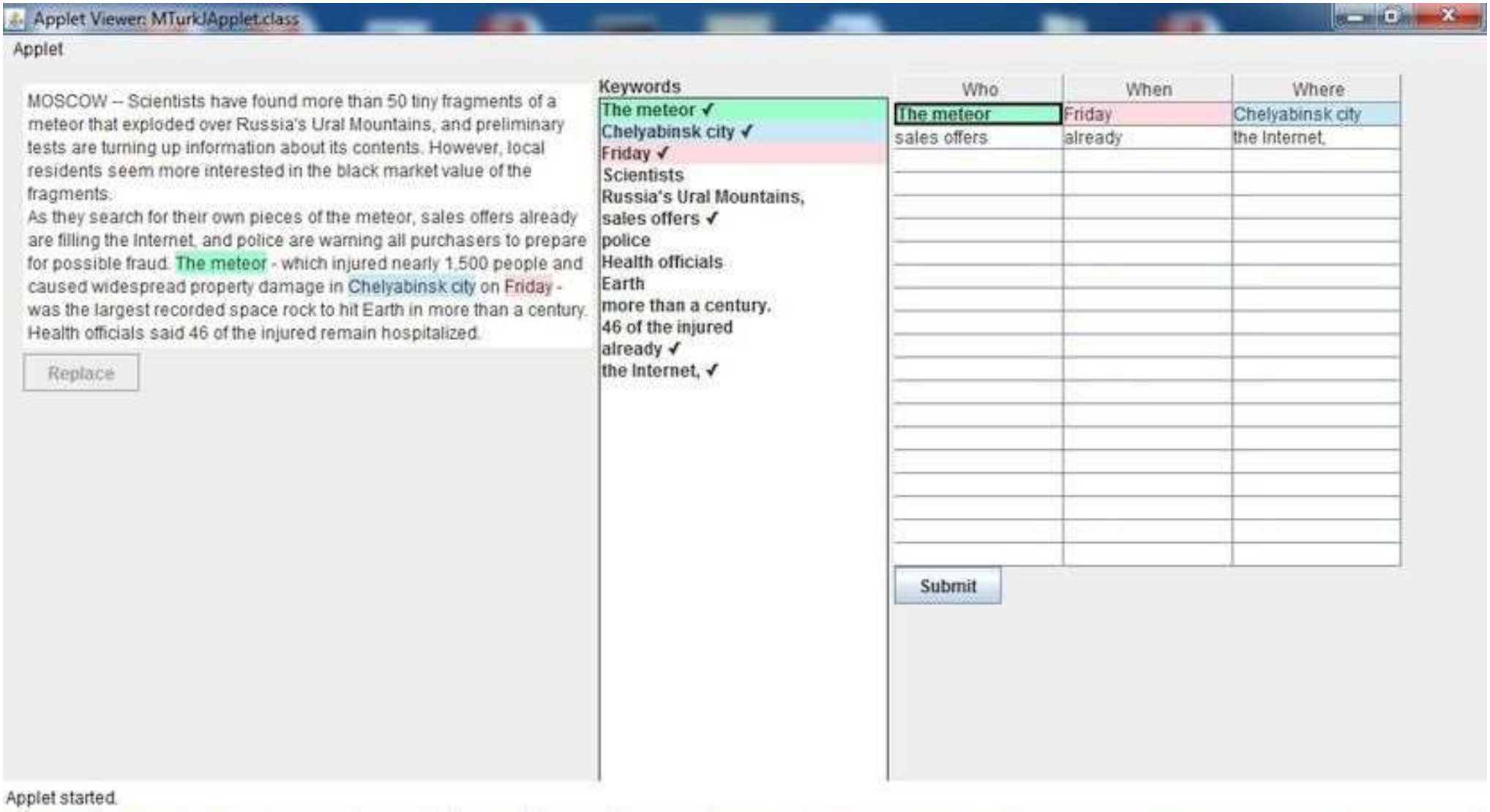
Attachment 2



Attachment 3



Attachment 4



Attachment 5