Utility Evaluation of Cross-document Information Extraction

Heng Ji[a], Zheng Chen[a], Jonathan Feldman[a], Antonio Gonzalez[a], Ralph Grishman[b], Vivek Upadhyay[a]
[a]Computer Science Department, Queens College and the Graduate Center, City University of New York
New York, NY 11367, USA
[b]Computer Science Department, New York University, New York, NY 10003, USA
hengji@cs.qc.cuny.edu, zchen1@gc.cuny.edu, agonzalez117@qc.cuny.edu, grishman@cs.nyu.edu,
vivekqc@gmail.com

Abstract

We describe a utility evaluation to determine whether cross-document information extraction (IE) techniques measurably improve user performance in news summary writing. Two groups of subjects were asked to perform the same time-restricted summary writing tasks, reading news under different conditions: with no IE results at all, with traditional single-document IE results, and with cross-document IE results. Our results show that, in comparison to using source documents only, the quality of summary reports assembled using IE results, especially from cross-document IE, was significantly better and user satisfaction was higher. We also compare the impact of different user groups on the results.

1 Introduction

Information Extraction (IE) is a task of identifying 'facts' (entities, relations and events) within unstructured documents, and converting them into structured representations (e.g., databases). IE techniques have been effectively applied to different domains (e.g. daily news, Wikipedia, biomedical reports, financial analysis and legal documentations) and different languages. Recently we described a new cross-document IE task (Ji et al., 2009) to extract events across-documents and track them on a time line. Compared to traditional single-document IE, this new task can extract more salient, accurate and concise event information.

However, a significant question remains: will the events extracted by IE, especially this new cross-document IE task, actually help end-users to make better use of the large volumes of news? In order to investigate whether we have reached this goal, we performed an extrinsic utility (i.e., usefulness) and usability evaluation on IE results. Two groups of subjects were asked to perform the same time-restricted summary writing tasks, reading news under different conditions: with no IE results at all, with traditional single-document IE results, and with cross-document IE results. Our results show that, in comparison to using source documents only, the quality of summary reports assembled using IE techniques, especially from cross-document IE, was significantly better. Also, as extraction quality increases from no IE at all to single-document IE and then to cross-document IE, user satisfaction increases. We also compare the impact of different user groups on the results. To the best of our knowledge, this is the first systematic evaluation of cross-document IE from a usability perspective.

2 Overview of IE Systems

We applied the English single-document IE system (Ji and Grishman, 2008) and cross-document IE system presented in (Ji et al., 2009). Both systems were developed for the ACE program[1].

The single-document IE system can extract events from individual documents. The core stages include entity extraction, time expression extraction and normalization, relation extraction and event extraction. Events include the 33 distinct types defined in ACE05. The extraction results are presented in tabular form.

The cross-document IE system can identify important person entities which are frequently in-

[1] http://www.itl.nist.gov/iad/mig/tests/ace/2005/

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 285–288,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

volved in events as ‘centroid entities’; and then for each centroid entity, link and order the events centered around it on a time line and associate them to a geographical map. The event chains are presented in a user-friendly graphical interface (Ji and Chen, 2009). Both systems link the events back to their context documents.

3 Evaluation Methods

3.1 Study Execution

Our measurement challenge is to assess how IE techniques affect users’ abilities to perform real-world tasks. We followed the summary writing task described in the Integrated Feasibility Experiment of the DARPA TIDES program (Colbath and Kubala, 2003) and the daily task conducted by intelligence analysts (Bodnar, 2003). Each task in our evaluation is based on writing a summary of ACE-type events involving a specific centroid entity, using one of three levels of support:

- Level (I): Read the news articles, with assistance of keyword based sentence search;
- Level (II): (I) + with assistance from single-document IE results;
- Level (III): (I) + with assistance from cross-document IE results.

The summary writing task for each entity using any level should be finished in 10 minutes. The users can choose to trust the IE results to create new sentences or select relevant sentences from the source documents. The IE systems were applied to a corpus of 106 articles from ACE 2005 training data.

3.2 Summary Scoring

We measure user responses in three aspects:

- *Observer-based Quantity* -- How many sentences are extracted in each summary? How many of them are uniquely correct?
- *Observer-based Quality*-- How fluent and coherent are the sentences in each summary?
- *User-based Usability* -- How does the user feel about the system?

3.3 User Group Selection

We selected user groups based on the principles that we should run as many tests as we can afford (Nielsen, 1994), and at least 5 to insure that we detect any major usability problems (Faulkner, 2003). Two different groups of users were asked to conduct the evaluation:

(1) Hallway Evaluation

We chose the first group of users with a “Hallway Testing” user-study method described in (Nielsen, 1994). We randomly asked 11 PhD students in the field of natural language processing to conduct the evaluation. In order to evaluate these three levels independently, each student was asked to write at most one summary, using one of the three levels, for any single centroid entity. To avoid the impact of diverse text comprehension abilities, each student was involved in all of these three levels for different centroid entities.

(2) Remote Evaluation

An effective utility evaluation will require users with a diversity of prior knowledge and computer experience. Therefore we asked the second group of 11 users in a remote usability testing mode (Hammontree et al., 1994). We sent out the request to university-wide undergraduate student mailing lists and found 11 users to work on the evaluation. The evaluation procedure follows the Hallway Testing method, except that the tests are carried out in the user’s own environment (rather than labs) helping further simulate real-life scenario testing. Also the users didn’t meet with the observers and thus they were not aware of any expectations for results.

4 Evaluation Results

In this section we will focus on reporting the results from Hallway Evaluation, while providing comparisons with Remote Evaluation.

4.1 Observer-based Quantity

The summaries were judged by two annotators and the judgements reconciled. A summary sentence is judged as uniquely correct if it: (1) includes relevant events involving the centroid entity; and (2) the same information was not included in previous sentences in the current summary. This metric can be considered as an approximate com bination of the “content responsiveness”, “non-redundancy”and “focus” criteria in the NIST TAC summarization track[2]. Table 1 presents the

[2] http://www.nist.gov/tac/2009/Summarization/update.summ.09.guidelines.html

Centroid	(I)	(II)	(III)	Centroid	(I)	(II)	(III)	Centroid	(I)	(II)	(III)
Bush	3/1/0	5/1/2	6/0/0	Al-douri	4/3/3	4/2/0	6/0/1	Ba'asyir	3/1/0	3/0/0	5/0/0
Ibrahim	4/0/1	5/0/0	8/0/0	Giuliani	2/0/0	3/2/0	5/0/0	Erdogan	1/0/1	4/0/0	4/0/0
Toefting	0/0/0	7/1/0	4/0/0	Blair	2/0/1	3/0/0	5/0/0	Diller	3/0/0	4/1/0	3/0/0
Putin	2/1/0	4/3/2	7/1/1	Pasko	3/0/0	3/0/0	2/0/0	**Overall**	**27/6/6**	**45/10/5**	**55/1/2**

Table 1. # (uniquely correct sentences)/ #(redundant correct sentences)/ #(spurious sentences) in a summary in Hallway Evaluation

quantified Hallway Testing results for each centroid separately and the overall score. It shows that overall Level (II) contained 18 more correct sentences than the baseline (I), while (III) achieved 11 further correct sentences. (I) obtained significantly fewer sentences without assistance from IE tools. We conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on a query entity basis for accuracy - number of (uniquely correct sentences)/number of (total extracted sentences in a summary). The results show that (III) is significantly better than (I) at a 99.2% confidence level, and better than (II) at a 96.9% confidence level. (II) is not significantly better than (I).

We can also see that for some centroid entities such as “Putin”, “Al-douri” and “Giuliani”, (II) generated more sentences but also introduced more redundant information. The user feedback has indicated that they did not have enough time to remove redundancy. In contrast, (III) yielded much less redundant information. In fact, the average time the users spent using (III) was only about 7.2 minutes. Therefore we can conclude that cross-document IE can produce more informative summaries in a more efficient way.

Error analysis showed that the major error types propagated from IE to summaries are as follows.

1. Event time errors. For example, the summary sentence “Toefting was convicted in September 2001 of assaulting a pair of restaurant workers in the capital” was judged as incorrect because the time argument should be “October 2002”.

2. Pronoun resolution errors. When a pronoun is mistakenly linked to an entity, incorrect event arguments will be included in the summaries.

3. Event type errors. When an event is misclassified, the users tend to use incorrect templates and thus generate wrong summaries.

4. Negative events. Sometimes the event attribute classifier makes mistakes and the users include negative events in the summaries.

4.2 Impact of User Groups

In the Remote Testing, the accuracy results from the three levels are as follows: 21/37, 28/37 and 31/36. Thus both user groups benefited from using IE techniques, but the enhancements vary a lot. In the Hallway Testing, the users were better trained and more familiar with IE tools (including the graphical interface of cross-document IE); and thus they can benefit more from the IE techniques. In contrast, in the Remote Evaluation, the users had quite diverse knowledge backgrounds. For example, one remote user was only able to find 1-2 sentences using any of the three levels; while another, more skilled remote user found more than 5 sentences with any level. However the Remote Evaluation is important to gather the feedback of the more subjective usability evaluation in section 4.4. Because the users in Hallway Testing may be aware of the observations that the observer is hoping to achieve, they may provide potentially biased feedback.

4.3 Observer-based Quality

The evaluation also showed that (III) produced summaries with better quality. We asked the observers to give a score between [1, 10] to each summary according to the following TAC summarization quality criteria: Readability/Fluency, Referential Clarity and Structure/Coherence. Table 2 shows the evaluation results for the three different methods.

Criteria	(I)	(II)	(III)
Readability/Fluency	9.4	8.5	8.2
Referential Clarity	6.1	8.3	8.7
Structure/Coherence	7.1	7.6	8.5

Table 2. Observer-based Average Quality

In their detailed feedback, the users indicated that (III) has the following advantages: (1) Better pronoun resolution; (2) More complete and accurate temporal order because (III) Can recover unknown time arguments using cross-document inference. (3) Can generate abstractive summaries. For the biographical events (e.g. employment), some users were able to use specific templates such as "PER was hired by ORG at TIME" to write summaries. For example, a sentence "Bush and Blair met at Camp David and the UK three times in March 2003" was derived from three different "Contact-Meeting" events in the event chains. (4) Can connect related events into more concise summaries. For example, several events were connected to generate the following sentences "Pasko was appealed for treason crime on April 16, 2003 *and then* released on June 15, 2003". The readability scores in Table 2 also indicate that a more effective template generation method should be developed to produce more fluent summaries based on IE results.

4.4 User-based Usability

The user feedback from both evaluations also showed that (II) and (III) results were trusted almost equally, and (III) was claimed to provide the most useful functions. The positive comments about (III) include "Temporal Linking allows logical reasoning and generalization", "Centroid search helps to focus immediately", "Spatial Linking allows to browse all the places which a person has visited", "Name disambiguation helps to filter irrelevant information", "Can find key information from event chains", "Timeline helps correlate events"; and the negative comments include "Sometimes IE errors mislead locating the sentences", "No support of name pair search for meeting events", "No color emphasis of events on the original documents" and "No suggestions of templates to compose summary sentences".

5 Conclusion and Future Work

Through a utility evaluation on summary writing we have proved that IE techniques, especially cross-document IE, can aid news browsing, search and analysis. In particular, temporal event tracking across documents helps users perform better at fact-gathering than they do without IE. Users also produced more informative summaries with cross-document IE than with traditional single-document IE. We also compared and analyzed the differences between two user groups. Such measures of the benefits to the eventual end users also provided feedback on what works well and identified additional research problems, such as to expand the centroid to a pair of entities and to provide confidence metrics in the interface. In the future we aim to set up an online news article analysis system and perform larger and regular utility evaluations.

Acknowledgement

This work was supported by the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, Google, Inc., CUNY Research Enhancement Program, Faculty Publication Program and GRTI Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

John W. Bodnar. 2003. Warning Analysis for the Information Age: Rethinking the Intelligence Process. *Center for Strategic Intelligence Research, Joint Military Intelligence College*, Washington, D.C.

Sean Colbath and Francis Kubala. 2003. TAP-XL: An Automated Analyst's Assistant. *Proc. HLT-NAACL 2003 (demonstrations).*

Laura Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods Instruments and Computers* 35(3), 379-383.

Monty Hammontree, Paul Weiler and Nandini Nayak. 1994. Remote Usability Testing. *Interactions*. Volume 1, Issue 3. Pages: 21-25.

Heng Ji and Zheng Chen. 2009. Cross-document Temporal and Spatial Person Tracking System Demonstration. *Proc. HLT-NAACL 2009.*

Heng Ji, Ralph Grishman, Zheng Chen and Prashant Gupta. 2009. Cross-document Event Extraction, Ranking and Tracking. *Proc. Recent Advances in Natural Language Processing 2009.*

Jakob Nielsen. 1994. Usability Engineering. Morgan Kaufmann Publishers.