A Critical Survey of the Methodology for IE Evaluation

A. Lavelli[*], M. E. Califf[◦], F. Ciravegna[⋄], D. Freitag[‡], C. Giuliano[*], N. Kushmerick[†], L. Romano[*]

[*]ITC-irst, Trento, Italy {lavelli, giuliano, romano}@itc.it

[◦]Department of Applied Computer Science, Illinois State University, Illinois, USA mecalif@ilstu.edu

[⋄]Computer Science Department, University of Sheffield, UK f.ciravegna@dcs.shef.ac.uk

[‡]Fair Isaac Corporation, California, USA dayne@cs.cmu.edu

[†]Computer Science Department, University College Dublin, Ireland nick@ucd.ie

Abstract

We survey the evaluation methodology adopted in Information Extraction (IE), as defined in the MUC conferences and in later independent efforts applying machine learning to IE. We point out a number of problematic issues that may hamper the comparison between results obtained by different researchers. Some of them are common to other NLP tasks: e.g., the difficulty of exactly identifying the effects on performance of the data (sample selection and sample size), of the domain theory (features selected), and of algorithm parameter settings. Issues specific to IE evaluation include: how leniently to assess inexact identification of filler boundaries, the possibility of multiple fillers for a slot, and how the counting is performed. We argue that, when specifying an information extraction task, a number of characteristics should be clearly defined. However, in the papers only a few of them are usually explicitly specified. Our aim is to elaborate a clear and detailed experimental methodology and propose it to the IE community. The goal is to reach a widespread agreement on such proposal so that future IE evaluations will adopt the proposed methodology, making comparisons between algorithms fair and reliable. In order to achieve this goal, we will develop and make available to the community a set of tools and resources that incorporate a standardized IE methodology.

1. Introduction

Evaluation has a long history in Information Extraction (IE), mainly thanks to the MUC conferences, where most of the IE evaluation methodology (as well as most of the IE methodology as a whole) was developed (Hirschman, 1998). In particular the DARPA/MUC evaluations produced and made available some annotated corpora that have been used as standard testbeds. More recently, a variety of other corpora have been shared by the research community, such as Califf's job postings collection (Califf, 1998), and Freitag's seminar announcements, corporate acquisition, university Web page collections (Freitag, 1998).

However, the definition of an evaluation methodology and the availability of standard annotated corpora do not guarantee that the experiments performed with different approaches and algorithms proposed in the literature can be reliably compared. Some of the problems are common to other NLP tasks (e.g., see (Daelemans and Hoste, 2002)): the difficulty of exactly identifying the effects on performances of the data used (the sample selection and the sample size), of the information sources used (the features selected), and of the algorithm parameter settings.

One issue specific to IE evaluation is how leniently to assess inexact identification of filler boundaries. Another question concerns the possibility of multiple fillers for a slot and how the counting is performed. Finally, because of the complexity of the task, the limited availability of tools, and the difficulty of reimplementing published algorithms (usually quite complex and sometimes not fully described in papers), in IE there are very few comparative articles in the sense mentioned in (Daelemans and Hoste, 2002). Most of the papers simply present the results of the new proposed approach and compare them with the results reported in previous articles. There is rarely any detailed analysis to ensure that the same methodology is used across different experiments.

Given this predicament, it is obvious that a few crucial issues in IE evaluation need to be clarified. This paper aims at providing a solid foundation for carrying out meaningful comparative experiments. The goal of the paper is to provide a critical survey of the methodologies employed in the main IE evaluation tasks. In this paper we concentrate our attention on the preliminary steps of the IE evaluation. First, we describe the IE evaluation methodology as defined in the MUC conference series and in other reference works. Then, we point out both the problems common also to the evaluation of other NLP tasks and those specific to IE. Finally, we draw some directions for future work.

2. IE Evaluation Methodology

The MUC conferences can be considered the starting point of the IE evaluation methodology as currently defined. The MUC participants borrowed the Information Retrieval concepts of precision and recall for scoring filled templates. Given a system response and an answer key prepared by a human, the system's precision was defined as the number of slots it filled correctly, divided by the number of fills it attempted. Recall was defined as the number of slots it filled correctly, divided by the number of possible correct fills, taken from the human-prepared key. All slots were given the same weight. F-measure, a weighted combination of precision and recall, was also introduced to provide a single figure to compare different systems' performances.

Apart from the definition of precise evaluation measures, the MUC conferences made other important contributions to the IE field: the availability of large amount of annotated data (which have made possible the development of Machine Learning based approaches), along with the evaluation software (i.e., the MUC scorer (Douthat, 1998)), the emphasis on domain-independence and portability, and the identification of a number of different tasks which can be evaluated separately(Hirschman, 1998).

It should be noticed that MUC evaluation concentrated mainly on IE from relatively unrestricted text, i.e. newswire articles. In independent efforts, other researchers developed and made available annotated corpora developed from somewhat more constrained texts. Califf compiled and annotated a set of 300 job postings from the Internet (Califf, 1998), and Freitag compiled corpora of seminar announcements and university web pages, as well as a corporate acquisitions corpus from newswire texts (Freitag, 1998). Several of these corpora are available from the RISE repository (RISE, 1998) where a number of tagged corpora have been made available by researchers in Machine Learning for IE.

Freitag (1998) uses the term Information Extraction in a more restricted sense than MUC. In the Seminar Announcement collection, the templates are simple and include slots for the seminar speaker, location, start time, and end time. This is in strong contrast with what happened in MUC where templates might be nested (i.e., the slot of a template may take another template as its value), or there might be several templates from which to choose, depending on the type of document encountered. In addition, MUC domains include irrelevant documents which a correctly behaving extraction system must discard. A template slot may be filled with a lower-level template, a set of strings from the text, a single string, or an arbitrary categorical value that depends on the text in some way (a so-called "set fill").

Califf (1988) takes an approach that is somewhat inbetween Freitag's approach and more complex MUC extraction tasks. All of the documents are relevant to the task, and the assumption is that there is precisely one template per document, but that many of the slots in the template can have multiple fillers.

Although the tasks to be accomplished are different, the methodology adopted by (Freitag, 1998) and (Califf, 1998) is similar to the one used in the MUC competition: precision, recall, and F-measure are employed as measures of the performances of the systems.

3. Problematic Issues in IE Evaluation

In Section 2. we have summarized the current status of the methodology adopted in IE. However, the definition of an evaluation methodology and the availability of standard annotated corpora do not guarantee that the experiments performed with different approaches and algorithms proposed in the literature can be reliably compared. Some of the problems are common to other NLP tasks (e.g., see (Daelemans and Hoste, 2002)): the difficulty of exactly identifying the effects on performances of the data used (the sample selection and the sample size), of the information sources used (the features selected), and of the algorithm parameter settings.

One of the most relevant issues is that of the exact split between training set and test set, considering both the numerical proportions between the two sets (e.g., a 50/50 *vs.* a 80/20 split) and the procedure adopted to partition the documents (e.g., n repeated random splits *vs.* n-fold cross-validation).

Furthermore, the question of how to formalize the learning-curve sampling method and its associated cost-benefit trade-off may cloud comparison further. For example, the following two approaches have been used: (1) For each point on the learning curve, train on some fraction of the available data and test on the remaining fraction; or (2) Hold out some fixed test set to be used for all points on the learning curve. The second approach is generally preferable: with the first procedure, points on the "high" end of the learning curve will have a larger variance than points on the "low" end.

Another important issue concerns the features used by the algorithm and their contribution to the performances of the algorithm. In IE, for instance, it would be relevant to extensively investigate the effectiveness of the use of simple orthographic features with respect to the use of more complex linguistic features such as PoS tags or semantic labels extracted from gazetteers (Ciravegna, 2001b).

Apart from those problematic issues mentioned above, there are some others that are specific to IE evaluation. A first issue concerns how to deal with issues related to tokenization, which is often considered something obvious and non problematic but it is not so and can affect the performance of the IE algorithms.

A second issue is related to how to evaluate an extracted fragment - e.g., if an extra comma is extracted should it count as correct, partial or wrong? This issue is related to the question of how relevant is the exact identification of the boundaries of the extracted items. (Freitag, 1998) proposes three different criteria for matching reference instances and extracted instances:

Exact The predicted instance matches exactly an actual instance.

Contains The predicted instance strictly contains an actual instance, and at most k neighboring tokens.

Overlap The predicted instance overlaps an actual instance.

Each of these criteria can be useful, depending on the situation, and it can be interesting to observe how performance varies with changing criteria. (De Sitter and Daelemans, 2003) mention such criteria and present the results of their algorithm for all of them.

A third issue concerns which software has been used for the evaluation. The only publicly available tool for such aim is the MUC scorer. Usually IE researchers have implemented their own scorer, relying on a number of implicit assumptions that have a strong influence on performance's evaluation.

When multiple fillers are possible for a single slot, there is an additional ambiguity – usually glossed over in papers – that can influence performance. For example, (Califf and

Mooney, 2003) remark that there are differences in counting between RAPIER (Califf, 1998), SRV (Freitag, 1998), and WHISK (Soderland, 1999). In his test on Job Postings (Soderland, 1999) does not eliminate duplicate values. When applied to Seminar Announcements SRV and RAPIER behave differently: SRV assumes only one possible answer per slot, while RAPIER makes no such assumption since it allows for the possibility of needing to extract multiple independent strings.

De Sitter and Daelemans (2003) also discuss this question and claim that in such cases there are two different ways of evaluating performance in extracting slot fillers: to find *all occurrences* (AO) of an entity (e.g. every mention of the job title in the posting) or only one occurrence for each template slot (one best per document, OBD). The choice of one alternative over the other may have an impact on the performance of the algorithm. (De Sitter and Daelemans, 2003) provide results for the two alternative ways of evaluating performances. This issue is often left underspecified in papers and, given the lack of a common software for evaluation, this further amplifies the uncertainty about the reported results.

Note that there are actually three ways to count:

- one answer per slot (where "2pm" and "2:00" are considered one correct answer)
- one answer per occurrence in the document (each individual appearance of a string to be extracted in the document where two separate occurrences of "2pm" would be counted separately)
- one answer per different string (where two separate occurrences of "2pm" are considered one answer, but "2:00" is yet another answer)

Freitag takes the first approach, Soderland takes the second, and Califf takes the third.

To summarize, an information extraction task should specify all of the following:

1. A set of fields to extract.
2. The legal numbers of fillers for each field, such as "exactly one value", "zero or one values", "zero or more values", or "one or more values". For example, in Seminar Announcements, the fields `stime`, `etime` and `location` are "0-1", `speaker` is "1+"; for Job Postings, `title` is "0-1 or 0+", required programming languages is "0+", etc. Thus, in the following seminar announcement:

   > Speakers will be Joel S. Birnbaum and Mary E.S. Loomis.

   if the task specifies that there should be one or more speaker, then to be 100% correct the algorithm must extract both names, while if the task specifies that zero or more speakers are allowed, then extracting either name would result in 100% correct performance.
3. The possibility of multiple varying occurrences of any particular filler. For example, a seminar announcement with 2 speakers might refer to them each twice, but slightly differently:

   > Speakers will be Joel S. Birnbaum and Mary E.S. Loomis. Dr. Birnbaum is Vice President of Research and Development and Dr. Loomis is Director of Software Technology.

   In this case, if we adopt the "one answer per slot" approach any of the following extractions should count as 100% correct: 'Joel S. Birnbaum, Mary E.S. Loomis'; 'Joel S. Birnbaum, Dr. Loomis'; 'Dr. Birnbaum, Mary E.S. Loomis'; 'Dr. Birnbaum, Dr. Loomis'; 'Joel S. Birnbaum, Dr. Birnbaum, Dr. Loomis'; 'Joel S. Birnbaum, Dr. Birnbaum, Mary E.S. Loomis'; 'Joel S. Birnbaum, Dr. Loomis, Mary E.S. Loomis'; 'Dr. Birnbaum, Dr. Loomis, Mary E.S. Loomis'; 'Joel S. Birnbaum, Dr. Birnbaum, Dr. Loomis, Mary E.S. Loomis'. On the other hand, both of the following get only partial credit: 'Joel S. Birnbaum, Dr. Birnbaum'; 'Mary E.S. Loomis, Dr. Loomis'.
4. How stringently are matches evaluated (exact, overlap or contains)?

While issue #1 above is always specified, issues #2, #3 and #4 are usually specified only implicitly.

4. Towards Reliable Evaluations

In the previous section, we have outlined a number of issues that can hamper the efforts for comparatively evaluating different IE approaches. To fix this situation, some steps are necessary. We propose a precise and reproducible evaluation methodology for IE tasks. This includes the definition of the exact experimental setup (both the numerical proportions between the training and test sets and the procedure adopted to select the documents). This will guarantee a reliable comparison of the performance of different algorithms.

Other initiatives that would help the evaluation within the IE community include the correction of errors and inconsistencies in annotated corpora. During the years a lot of researchers have used the IE testbeds for performing experiments. During such experiments minor errors and inconsistencies in annotations have been discovered, and sometimes corrected versions of the corpora have been produced. We have been collecting such versions and will produce and distribute new, "improved" versions of the annotated corpora.

A final issue concerning annotations is the fact that different algorithms may need different kinds of annotations: either tagged texts (e.g., BWI (Freitag and Kushmerick, 2000), $(LP)^2$(Ciravegna, 2001a)) or templates associated with texts (e.g., RAPIER). Note that two of the most frequently used IE testbeds (i.e., Seminar Announcements and Job Postings) adopt two different kinds of annotations. While transforming tagged texts into templates can be considered straightforward, the reverse is far from obvious and the differences in the annotations which the algorithms rely on can produce relevant differences in performances. This raises the issue of having two different but consistent annotations of the same corpus. We are collecting these different corpora and making them available to the community.

Finally, to simplify running experiments, it would be helpful to adopt a uniform format for all corpora, e.g. based on XML. Adopting XML would also help solving the consistency problem (mentioned above) between different versions of the same corpus. We are exploring the possibility of adopting the approach standard in the corpora community: creating one file containing the original text and one for each type of annotations.

5. Conclusions and Future Work

The work reported in this paper aims at elaborating a clear and detailed experimental methodology and proposing it to the IE community. The aim is to reach a widespread agreement so that future IE evaluations will adopt the proposed methodology, making comparisons between algorithms fair and reliable. In order to achieve this goal, we will develop and make available to the community a set of tools and resources that incorporate a standardized IE methodology. This will include the creation of web pages in the web site of the Dot.Kom project (`www.dot-kom.org`) where these guidelines and resources will be made available. They include:

Exact definition of the corpus partition One of the crucial issues is that of the exact split between training set and test set, considering both the numerical proportions between the two sets (e.g., a 50/50 *vs.* 80/20 split) and the procedure adopted to select the documents (e.g., n repeated random splits *vs.* n-fold cross-validation). As is well known, different partitions can affect the system results, therefore we will establish the partitions to be used for the experiments.

Fragment evaluation Errors in extraction can be evaluated differently according to their nature. For example, if an extra comma is extracted should it count as correct, partial or wrong? This issue is related to the question of how relevant the exact identification of the boundaries of the extracted items is.

Improved versions of corpora We are collecting the different versions of the standard corpora produced by researchers so to compare the corrections introduced and produce new versions which take such corrections into account. The final aim is to distribute new, "improved" versions of the annotated corpora.

Scorer Use of the MUC scorer for evaluating the results. We will define the exact matching strategies by providing the configuration file for each of the tasks selected and guidelines for further corpora.

Learning curve When working on learning algorithms, the simple global results obtained on the whole corpus are not very informative. The study of the learning curve is very important. Therefore all the evaluations will involve computing a full learning curve. We will define the strategy to be used for determining the learning curve for each corpus.

Some work in such direction has already been done in the framework of the EU Dot.Kom project, and further efforts will be spent in the future months.

Acknowledgments

Fabio Ciravegna, Claudio Giuliano, Alberto Lavelli and Lorenza Romano are supported by the IST-Dot.Kom project (`www.dot-kom.org`), sponsored by the European Commission as part of the Framework V (grant IST-2001-34038). Nicholas Kushmerick is supported by grant 101/F.01/C015 from Science Foundation Ireland and grant N00014-03-1-0274 from the US Office of Naval Research.

6. References

Califf, M. and R. Mooney, 2003. Bottom-up relational learning of pattern matching rules for information extraction. *Journal of Machine Learning Research*, 4:177–210.

Califf, Mary Elaine, 1998. *Relational Learning Techniques for Natural Language Information Extraction*. Ph.D. thesis, University of Texas at Austin.

Ciravegna, Fabio, 2001a. Adaptive information extraction from text by rule induction and generalisation. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*. Seattle, WA.

Ciravegna, Fabio, 2001b. $(LP)^2$, an adaptive algorithm for information extraction from web-related texts. In *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. Seattle, WA.

Daelemans, Walter and Véronique Hoste, 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Spain.

De Sitter, An and Walter Daelemans, 2003. Information extraction via double classification. In *Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining (ATEM 2003)*. Cavtat-Dubronik, Croatia.

Douthat, A., 1998. The message understanding conference scoring software user's manual. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_%manual.html`.

Freitag, Dayne, 1998. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. thesis, Carnegie Mellon University.

Freitag, Dayne and Nicholas Kushmerick, 2000. Boosted wrapper induction. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000)*. Austin, Texas.

Hirschman, Lynette, 1998. The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech and Language*, 12:281–305.

RISE, 1998. A repository of online information sources used in information extraction tasks. *[http://www.isi.edu/info-agents/RISE/index.html] Information Sciences Institute / USC*.

Soderland, S., 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272.