IE evaluation: Criticisms and recommendations

A. Lavelli
ITC-irst
Trento, Italy

M. E. Califf
Department of Applied Computer Science
Illinois State University, USA

F. Ciravegna
Computer Science Department
University of Sheffield, UK

D. Freitag
Fair Isaac Corporation
San Diego, California, USA

C. Giuliano
ITC-irst
Trento, Italy

N. Kushmerick
Computer Science Department
University College Dublin, Ireland

L. Romano
ITC-irst
Trento, Italy

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.

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. 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 & 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. (Freitag 1998) proposes three different criteria for matching reference instances and extracted instances: *exact*, *overlap*, *contains*. 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 & 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 different 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. We then describe the main reference corpora, their characteristics, how they have been evaluated, etc. Finally, we suggest some directions for future work.

The aim of the paper is to make a proposal to reach an agreement on a widely accepted experimental methodology which future IE evaluations should follow in order to make comparisons between algorithms useful and reliable.

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), the emphasis on domain-independence and portability, and the identification of a number of different tasks which can be evaluated separately.

In particular, the MUC conferences made available annotated corpora for training and testing[1], along with the evaluation software (i.e., the MUC scorer (Douthat 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: e.g., Seminar Announcements (Freitag 1998), Job Postings (Califf 1998). Further specific details about such corpora will be provided in Section "Reference Corpora 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 the last MUC conferences (such as MUC-6 and MUC-7) 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 in-between 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.

In cases where elaborate representations (nested templates, set fills) are required of a system, the task's difficulty may approach that of full NLP. In general, the challenges facing NLP cannot be circumvented in Information Extraction. Some semantic information and discourse-level analysis is typically required. To this are also added subproblems unique to Information Extraction, such as slot filling and template merging.

Problematic Issues in IE Evaluation

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 problems are common to other NLP tasks (e.g., see (Daelemans & 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.

Before proceeding, let us mention the issue of statistical analysis. All too often, IE research—like much of computer science—merely reports numerical performance differences between algorithms, without analyzing their statistical properties. The most important form of analysis is whether some reported numerical difference is in fact statistically significant. While rigourous statistical analysis is clearly important, this issue is beyond the scope of this paper.

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 split *vs.* a 80/20 one) 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

[1] The corpora for MUC-3 and MUC-4 are freely available in the MUC website (http://www.itl.nist.gov/iaui/894.02/related_projects/muc/), while those of MUC-6 and MUC-7 can be purchased via the Linguistic Data Consortium (http://ldc.upenn.edu/).

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 is distinguishing between an algorithm and the features it uses in their contribution to performance. In IE, for instance, some algorithms have employed simple orthographic features, while others use more complex linguistic feature such as PoS tags or semantic labels extracted from gazetteers (Califf 1998; Ciravegna 2001b; Peshkin & Pfeffer 2003).

Apart from those problematic issues mentioned above, there are others that are specific to IE evaluation.

A first 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 & Daelemans 2003) mention such criteria and present the results of their algorithm for all of them.

A second 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 scorers, 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 & 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 & 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)[2]
- 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 $\mathrm{location}$ are "0-1", $\mathrm{speaker}$ 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 each of them 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 Lab.

   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'.

[2] Note that the occurrences considered here are only those that can be interpreted without resorting to any kind of contextual reasoning. Hence, phenomena related to coreference resolution are not considered at all.

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 based on inspecting a sample of the labeled fields, intuition, etc.

Another relevant element concerns 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 final element that makes a sound comparison between different algorithms difficult is the fact that some papers present results only on one of the major reference corpora (e.g., Seminar Announcements, Job Postings, etc.). For example, (Roth & Yih 2001; Chieu & Ng 2002; Peshkin & Pfeffer 2003) report results only on the Seminar Announcements[3] and (De Sitter & Daelemans 2003) only on the Job Postings. On the other hand, (Freitag 1998) presents results on Seminar Announcements, corporate acquisition, and university web page collection, (Califf 1998) on Seminar Announcements, corporate acquisition and also on Job Postings, and (Ciravegna 2001a; Freitag & Kushmerick 2000) on both Seminar Announcements and Job Postings. Related to this aspect, there is also the fact that sometimes papers report only F-measure but not precision and recall, while the tradeoff between precision and recall is a fundamental aspect of performance.

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 concentrate on the prerequisite of defining a precise and reproducible evaluation methodology. 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.

Another relevant issue is the need of providing effectiveness measures (i.e., precision, recall, and F-measure) both on a per-slot basis as well as microaveraged over all slots[4].

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. For further details about the different versions for each corpus, see Section "Reference Corpora for IE".

[3]Although in (Roth & Yih 2002) the results for Job Postings are also included. Moreover, (Chieu & Ng 2002) report also results on Management Succession.

[4]The "all slots" figures are obtained by computing the sums of True Positives, False Positives, and False Negatives over all the slots (what in Text Categorization is called "microaveraging").

A final issue concerning annotations is the fact that different algorithms may need different kinds of annotations: either tagged texts (e.g., BWI (Freitag & 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 solve 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.

Reference Corpora for IE

The datasets used more often in IE[5] are Job Postings (Califf 1998), Seminar Announcements, corporate acquisition and the university web page collections (Freitag 1998). In the following we will describe the main characteristics of the first two of these corpora (set of fields to extract, standard train/test split, ...) together with tables showing the results published so far (precision, recall and F1 on a per-slot basis as well as microaveraged over all slots[6]). We report the results although, as indicated above, the different performances reported are not always reliably comparable. If the experimental conditions in which the results were obtained were different from those described in the original papers, we will describe them.

Seminar Announcements

The Seminar Announcement collection (Freitag 1998) consists of 485 electronic bulletin board postings distributed in the local environment at Carnegie Mellon University[7]. The purpose of each document in the collection is to announce or relate details of an upcoming talk or seminar. The documents were annotated for four fields: *speaker*, the name of seminar's speaker; *location*, the location (i.e., room and number) of the seminar; *stime*, the start time; and *etime*, the end time. Figure 1 shows an example taken from the corpus.

Methodology and Results (Freitag 1998) randomly partitions the entire document collection five times into two sets

[5]Note that here we are not taking into account the corpora made available during the MUC conferences which, because of the complexity of the IE tasks, have been not very often used in IE experiments after the MUC conferences. (Hirschman 1998) provides an overview of such corpora and of the related IE tasks.

[6]See footnote 4.

[7]Downloadable from the RISE repository: http://www.isi.edu/info-agents/RISE/repository.html.

<0.6.1.94.14.16.40.xu+@IUS4.IUS.CS.CMU.EDU (Yangsheng Xu).0>
Type: cmu.cs.robotics
Who: <speaker>Ralph Hollis</speaker>
Senior Research Scientist
The Robotics Institute
Carnegie Mellon University
Topic: Lorentz Levitation Technology:
a New Approach to Fine Motion Robotics, Teleoperation
Haptic Interfaces, and Vibration Isolation
Dates: 15-Jan-94
Time: <stime>3:30 PM</stime> - <etime>5:00 PM</etime>
Place: <location>ADAMSON WING Auditorium in Baker Hall</location>
Host: Yangsheng Xu (xu@cs.cmu.edu)
PostedBy: xu+ on 6-Jan-94 at 14:16 from IUS4.IUS.CS.CMU.EDU (Yangsheng Xu)
Abstract:

RI SEMINAR

WHEN: Friday, Jan 15, 1994; <stime>3:30 pm</stime> - <etime>5:00 pm</etime>
Refreshments will be served starting at 3:15 pm

WHERE: <location>ADAMSON WING Auditorium in Baker Hall</location>

SPEAKER: <speaker>Ralph Hollis</speaker>
Senior Research Scientist
The Robotics Institute
Carnegie Mellon University

TITLE: Lorentz Levitation Technology:
a New Approach to Fine Motion Robotics, Teleoperation
Haptic Interfaces, and Vibration Isolation

Figure 1: An excerpt from the seminar announcement cmu.cs.robotics-1018:0.

of equal size, training and testing. The learners are trained on the training documents and tested on the corresponding test documents for such partition. The resulting numbers are averages over documents from all test partitions. In (Freitag 1997), however, the randomly partitioning is performed ten times (instead of five). Later experiments have followed alternatively one of the two setups: e.g., (Califf 1998; Freitag & Kushmerick 2000; Ciravegna 2001a) follow the ten run setup[8]; (Roth & Yih 2001; Chieu & Ng 2002) follow the five run one; (Peshkin & Pfeffer 2003) do the same as well[9] and provide results on each single slot but showing only F-measure. Finally, (Soderland 1999) reports WHISK performances using 10-fold cross validation on a randomly selected set of 100 texts instead of using the standard split for training and test sets.

In Table 1 we list the results obtained by different systems on Seminar Announcements.

Different Versions During their experiments using Seminar Announcements, Fabio Ciravegna and Leon Peshkin produced their own "improved" versions of the corpus. These two versions have been used as a starting point to produce a new revised version, which will be soon made publicly available on the web site of the Dot.Kom project (http://www.dot-kom.org). Such version mainly fixes obvious annotation errors. E.g., errors in the inexact identification of stime and etime boundaries; usually, a missing final dot "." at the right boundary: the following piece of text of the RISE version (will be given at <stime>10:45 a.m</stime>., Tuesday,) was corrected as follows (will be given at <stime>10:45 a.m.</stime>, Tuesday,). Moreover, three further changes have been done: (1) file names are compliant with Windows conventions; (2) all <sentence> and <paragraph> tags have been stripped from the corpus; (3) the documents are XML-legal.

Moreover, there is also the Seminar Announcements corpus with associated templates produced by Mary Elaine Califf to run RAPIER.

Finally, (Peshkin & Pfeffer 2003) created a derivative dataset in which documents are stripped of headers and two extra fields are sought: *date* and *topic*.

Job Postings

The Job Posting collection (Califf 1998) consists of a set of 300 computer-related job postings from the Usenet newsgroup austin.jobs[10]. The information extraction task is to identify the types of information that would be useful in creating a searchable database of such jobs, with fields like *message-id* and the posting *date* which are useful for maintaining the database, and then fields that describe the job itself, such as the job *title*, the *company*, the *recruiter*, the *location*, the *salary*, the *language*s and *platform*s used, and required years of experience and degrees. Some of these slots can take only one value, but for most of the slots a job posting can contain more than one appropriate slot-filler. There are a total of 17 different slots for this task. Figure 2 shows an example taken from the corpus. Note that, differently from the Seminar Announcements, the annotations

[8](Califf 1998; Freitag & Kushmerick 2000) use exactly the same partitions as Freitag

[9]What is written in their paper is not completely clear but they have confirmed to us that they have adopted the five run setup (p.c.).

[10]Downloadable from the RISE repository: http://www.isi.edu/info-agents/RISE/repository.html. The collection we refer to in the paper is the following: http://www.isi.edu/info-agents/RISE/Jobs/SecondSetOfDocuments.tar.Z.

	SRV			RAPIER			WHISK			BWI		
Slot	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)
speaker	54.4	58.4	56.3	80.9	39.4	53.0	52.6	11.1	18.3	79.1	59.2	67.7
location	74.5	70.1	72.3	91.0	60.5	72.7	83.6	55.4	66.6	85.4	69.6	76.7
stime	98.6	98.4	98.5	96.5	95.3	95.9	86.2	100	92.6	99.6	99.6	99.6
etime	67.3	92.6	77.9	95.8	96.6	96.2	85.0	87.2	86.1	94.4	94.9	94.6
All slots			77.1			77.3			64.9			83.9
	$(LP)^2$			SNoW			ME_2			BIEN		
Slot	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)
speaker	87.0	70.0	77.6	83.3	66.3	73.8			72.6			76.9
location	87.0	66.0	75.1	90.9	64.1	75.2			82.6			87.1
stime	99.0	99.0	99.0	99.6	99.6	99.6			99.6			96.0
etime	94.0	97.0	95.5	97.6	95.0	96.3			94.2			98.8
All slots			86.0						86.9			

Table 1: Results obtained by different systems on CMU seminar announcements.

From: spectrum@onramp.net
Newsgroups: austin.jobs
Subject: US-TX-Austin - VISUAL BASIC Developers $50K to $70K
Date: Sat, 23 Aug 97 09:52:21
Organization: OnRamp Technologies, Inc.; ISP
Lines: 65
Message-ID: <NEWTNews.872347949.11738.consults@ws-n>
NNTP-Posting-Host: ppp10-28.dllstx.onramp.net
Mime-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII
X-Newsreader: NEWTNews & Chameleon – TCP/IP for MS Windows from NetManage
Xref: cs.utexas.edu austin.jobs:119473

US-TX-Austin - VISUAL BASIC Developers $50K to $70K

POSTING I.D. D05

Major corporations have immediate openings for Visual Basic programmers. 2-5 years experience; Oracle or SQL Server helpful. Windows 95 and Windows NT programming a plus. Please contact Bill Owens at (972) 484-9330; FAX (972) 243-0120 at Resource Spectrum.

To review several hundred positions with similar requirements please visit our web site at www.spectrumm.com. Please reference Posting ID and position title when contacting us. Qualified, experienced people from all over the world will be considered. You must speak and write English fluently. You must be a US citizen, a Permanent Resident, and meet all job requirements.

YOUR RESUME MUST BE SENT IN ASCII Text and then it will be stored digitally in our system. You will have a MUCH BETTER CHANCE OF being notified when a CAREER OPPORTUNITY presents itself by transmitting via E-Mail. MS-Word, WordPerfect, etc., will all convert a file from their normal format to a "Text Only" format. (ASCII text)

Resource Spectrum
5050 Quorum Dr., Ste 700
Dallas, Texas 75240

Internet Address: spectrum@onramp.net (We prefer this transmission)
Fax: (972) 243-0120
Voice (972)484-9330
Contact: Bill Owens

---

computer_science_job
id: NEWTNews.872347949.11738.consults@ws-n
title: Developers
salary: $50K to $70K
company:
recruiter: Resource Spectrum
state: TX
city: Austin
country: US
language: VISUAL BASIC
platform: Windows NT Windows 95
application: SQL Server Oracle
area:
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 23 Aug 97

Figure 2: An excerpt from the job posting job119473 together with its associated template.

of the Job Postings in (RISE 1998) are provided as separate templates associated with each text.

Methodology and Results (Califf 1998) performs experiments randomizing the collection, dividing it into 10 parts and doing 10-fold cross-validation; she also trained RAPIER on subsets of the training data at various sizes in order to produce learning curves. (Freitag & Kushmerick 2000; Roth & Yih 2002) adopt the same 10-fold cross-validation methodology. (Ciravegna 2001a) randomly partitions ten times the entire document collection into two sets of equal size, training and testing. (Soderland 1999) reports WHISK performances using 10-fold cross validation on a randomly selected set of 100 texts instead of using the standard split for training and test sets. Moreover, he reports only the overall figures for precision and recall and not the figures for the single slots.

(De Sitter & Daelemans 2003) use a Job Posting collection which is different from the one described above and consists of 600 postings[11] As a matter of fact, this version includes 600 postings with templates associated, while the tagged postings are 300 only and they are exactly those of the Job Postings collection available in RISE. De Sitter and Daelemans perform their evaluation using 10-fold cross-validation.

In Table 2 we list the results obtained by different systems on Job Postings. We do not list systems that either did not report results slot by slot but only overall figures (Soderland 1999) or reported results only on few slots (Freitag & Kushmerick 2000).

Different Versions Given the fact that some IE algorithms need a tagged corpus (rather than an external annotation as provided by the version of Job Postings available in the RISE repository), some researchers produced their own tagged version: we have found out three different versions produced by Mary Elaine Califf, Fabio Ciravegna, and Scott Wen-tau Yih. The creation of a standard "tagged" version is rather complex and its preparation will need some time.

Conclusions and Future Work

The "ideal" long-term goal would be to provide a flexible unified tool that could be used to recreate many of the previous algorithms (e.g., BWI (the original C version, or TIES, the Java reimplementation carried on at ITC-irst), RAPIER, $(LP)^2$, etc); along with standard code for doing test/train splits, measuring accuracy, etc. In short, we envision a sort of "Weka for IE". However, this goal is very challenging because it would involve either integrating legacy code written in different programming languages, or reimplementing published algorithms, whose details are subtle and sometimes not described in complete detail.

The work reported in this paper addresses a more practical mid-term goal: to elaborate a clear and detailed experimental methodology and propose 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 (http://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 issue 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 split *vs.* a 80/20 one) 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.

Definition of preprocessing tasks Some of the preparation subtasks (e.g., tokenization) may influence the performances of the algorithms. Therefore, when possible, we will provide an annotated version of the corpora with, for example, tokens, PoS tagging, gazetteer lookup and named entity recognition in order to allow fair comparison of the different algorithms. This will also facilitate the comparison of the impact of different features in the learning phase.

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

F. Ciravegna, C. Giuliano, A. Lavelli and L. Romano are supported by the IST-Dot.Kom project (www.dot-kom.org), sponsored by the European Commission as part

[11] Downloadable from ftp:/ftp.cs.utexas.edu/pub/mooney/job-data/job600.tar.gz.

	RAPIER			$(LP)^2$			SNoW			DeSitter - AO			DeSitter - OBD		
Slot	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)	Prec	Rec	F(1)
id	98.0	97.0	97.5	100.0	100.0	100.0	99.7	99.7	99.7	97	98	97	99	96	97
title	67.0	29.0	40.5	54.0	37.0	43.9	62.0	45.9	52.7	31	43	36	35	35	35
company	76.0	64.8	70.0	79.0	66.0	71.9	89.7	65.1	75.4	45	78	57	26	74	38
salary	89.2	54.2	67.4	77.0	53.0	62.8	89.3	61.6	72.9	56	70	62	62	72	67
recruiter	87.7	56.0	68.4	87.0	75.0	80.6	89.4	81.5	85.3	40	79	53	44	74	55
state	93.5	87.1	90.2	80.0	90.0	84.7	91.7	91.8	91.7	77	97	86	93	95	94
city	97.4	84.3	90.4	92.0	94.0	93.0	90.1	87.9	89.0	84	95	89	90	92	91
country	92.2	94.2	93.2	70.0	96.0	81.0	95.6	95.4	95.5	92	98	95	91	94	92
language	95.3	71.6	80.6	92.0	90.0	91.0	83.5	81.6	82.5	25	27	26	33	34	33
platform	92.2	59.7	72.5	81.0	80.0	80.5	74.4	73.8	74.1	31	34	32	35	38	36
application	87.5	57.4	69.3	86.0	72.0	78.4	84.7	47.5	60.9	32	29	30	31	30	30
area	66.6	31.1	42.4	70.0	64.0	66.9	63.5	43.4	51.6	16	17	16	16	18	17
req-years-e	80.7	57.5	67.1	79.0	61.0	68.8	90.2	78.4	83.9	50	80	62	72	81	76
des-years-e	94.6	81.4	87.5	67.0	55.0	60.4	75.3	83.1	79.0	33	55	41	36	66	47
req-degree	88.0	75.9	81.5	90.0	80.0	84.7	86.3	80.9	83.5	29	43	35	41	51	45
des-degree	86.7	61.9	72.2	90.0	51.0	65.1	81.0	48.8	60.9	28	45	35	29	37	33
post date	99.3	99.7	99.5	99.0	100.0	99.5	99.0	99.3	99.2	84	99	91	99	97	98
All slots	89.4	64.8	75.1			84.1									

Table 2: Results obtained by different systems on Job Postings.

of the Framework V (grant IST-2001-34038). N. Kushmerick is supported by grant from 101/F.01/C015 from Science Foundation Ireland and grant N00014-03-1-0274 from the US Office of Naval Research. We would like to thank Leon Peshkin for kindly providing us his own corrected version of the Seminar Announcement collection and Scott Wen-Tau Yih for his own tagged version of the Job Posting collection.

References

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

Califf, M. E. 1998. *Relational Learning Techniques for Natural Language Information Extraction*. Ph.D. Dissertation, University of Texas at Austin.

Chieu, H. L., and Ng, H. T. 2002. Probabilistic reasoning for entity and relation recognition. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2002)*.

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

Ciravegna, F. 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*.

Daelemans, W., and Hoste, V. 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)*.

De Sitter, A., and Daelemans, W. 2003. Information extraction via double classification. In *Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining (ATEM 2003)*.

Douthat, A. 1998. The message understanding conference scoring software user's manual. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

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

Freitag, D. 1997. Using grammatical inference to improve precision in information extraction. In *Proceedings of the ICML-97 Workshop on Automata Induction, Grammatical Inference, and Language Acquisition*.

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

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

Peshkin, L., and Pfeffer, A. 2003. Bayesian information extraction network. In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*.

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*.

Roth, D., and Yih, W. 2001. Relational learning via propositional algorithms: An information extraction case study. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*.

Roth, D., and Yih, W. 2002. Relational learning via propositional algorithms: An information extraction case study. Technical Report UIUCDCS-R-2002-2300, Department of Computer Science, University of Illinois at Urbana-Champaign.

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