



IE EVALUATION STRATEGY

**Philipp Cimiano[#], Fabio Ciravegna^{*}, Claudio Giuliano⁺,
Alberto Lavelli⁺, Lorenza Romano⁺, Mark Stevenson^{*}**

⁺ ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Povo (TN), Italy

^{*} Department of Computer Science, University of Sheffield, Regent Court, 211
Portobello Street, Sheffield. S1 4DP UK

[#] Institut für Angewandte Informatik und Formale Beschreibungsverfahren,
Universität Karlsruhe, D-76128, Karlsruhe, Deutschland

Abstract

This deliverable defines the evaluation strategies for the Information Extraction tools that are under development within the Dot.Kom project.

Version	1
Number of pages	11
Type	Deliverable
Status	Draft

Work Package	WP3
Period covered	1/7/2003-30/9/2003
WP/ Task responsible	ITC-irst
Project contact point:	F. Ciravegna
EC project officer	Kimmo Rossi
Status	Public
Actual distribution	Consortium-EC
Key words	information extraction, evaluation strategy

1 Introduction.....3
2 Evaluation in IE3
 2.1 Open Issues4
3 Evaluation in Dot.Kom.....6
 3.1 Use of ontology.....8
References9

1 INTRODUCTION

The goal of the deliverable is the definition of the evaluation strategies for the IE tools in the Dot.Kom project.

Two directions are outlined. On the one hand, our tools will be evaluated with a project-specific perspective. On the other hand we will produce evaluation strategies, tools and resources for IE algorithms that will serve the international scientific community. We will focus on a number of publicly available corpora for which now there is no clear methodology.

2 EVALUATION IN IE

Evaluation has a long history in IE, thanks to the MUC conferences, where most of the evaluation methodology (as well as most of the IE methodology as a whole) was developed [Hirschman 1998]. An important part of the MUC contribution was the availability of annotated corpora for training and testing, as well as of the evaluation software (e.g., the MUC scorer [Douthat 1998]). The corpora for MUC-3 and MUC-4 are freely available in the MUC website¹ while those of MUC-6 and MUC-7 can be bought via the Linguistic Data Consortium (<http://ldc.upenn.edu/>).

It should be noticed that MUC evaluation concentrated mainly on IE from free texts, i.e. newswire texts. The RISE repository [RISE 1998] provides a number of corpora for web-related documents, from structured to semi-structured and free ones. They are mainly intended for Machine Learning based IE. The RISE corpora, as well as the MUC corpora, come not only with the correct solutions (e.g., filled templates), but also with the results obtained by other researchers who previously performed the same tasks.

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

Concerning the definition of the different domain-specific tasks, at MUC-7 the following ones were evaluated (description taken from [Hirschman 1998]):

- Named Entity: identification of person (PERSON), location (LOC) and organization (ORG) names, as well as time, date and money expressions. At MUC-6 the highest performing automated Named Entity system was able to achieve a score comparable to human-human interannotator agreement. At MUC-7 the results were lower because of the absence of training data for the satellite launch domain.
- Coreference: identification of coreferring expressions in the text, including name coreference (*Microsoft Corporation* and *Microsoft*), definite reference

¹ http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

(the Seattle-based company) and pronominal reference (*it, he, she*). This was the most experimental of the tasks.

- Template Element: identification of the main entities (persons, organizations, locations), with one template per entity including its name, other “aliases” or shortened forms of the name, and a short descriptive phrase useful in characterizing it. The template elements constituted the building blocks for the more complex relations captured in template relation and scenario template tasks.
- Template Relation: identification of properties of Template Elements or relations among them (e.g., *employee_of* connecting person and organization, or *location_of* connecting organization and location). This task was introduced in MUC-7.
- Scenario Template: extract predefined event information and relate the event information to particular organization, person or artifact entities involved in the event. At MUC-7 the scenario concerned satellite launch events and the event template consisted of 7 slots.

The MUC 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. In [Makhoul et al. 1999] some limitations of F-measure are underlined and a new measure, slot error rate, is proposed. Even if the proposal is interesting, it does not seem to have had any impact on the IE community, which keeps on employing F-measure as the standard way of comparing systems’ performances.

The RISE repository contains a number of disparate corpora, without any specific common aim, being mainly corpora defined by independent researchers for evaluating their own systems and made available to others. Most of them are devoted to implicit relation extraction. As explained in D3.2, this is a task mainly defined by the wrapper induction community, requiring the identification of implicit events and relations. For example Freitag [Freitag 1998] defines the task of extracting speaker, start-time, end-time and location from a set of seminar announcements. No explicit mention of the event (the seminar) is done in the annotation. Implicit event extraction is simpler than full event extraction, but has important applications whenever either there is just one event per texts or it is easy to devise extraction strategies for recognizing the event structure from the document [Ciravegna and Lavelli *forthcoming*]. This kind of evaluation is suitable for the kind of tools developed so far in Dot.Kom and therefore the RISE corpora will be mainly used.

2.1 OPEN ISSUES

As said above, there is an IE methodology derived by MUC conferences with standard measures to evaluate and compare the results (precision, recall, F-measure). Moreover, there are reference annotated corpora with the performances obtained by previous systems. However, the definition of standard evaluation measures and the

availability of reference 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 (see for instance [Daelemans & Hoste 02, Hoste et al. 02, Daelemans et al. 03]): 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. Apart from problematic issues common to Machine Learning based NLP tasks, there are other aspects that are specific to IE evaluation such as the issue of how relevant is the exact identification of the boundaries of the extracted items (a more detailed discussion is presented in Section 3).

Another problematic aspect when comparing the performances of different algorithms is that only some papers present results on all the main reference corpora (e.g., Seminar Announcements, Job Postings) [Califf 98, Freitag 98, Ciravegna 01a, Freitag & Kushmerick 00]. Other papers show their results only on a single corpus, i.e. the Seminar Announcements² [Roth & Yih 01, Peshkin & Pfeffer 03]. So, it is difficult to evaluate how general and successful the proposed solutions are.

The situation is even more problematic for papers on explicit relation extraction, an area where there are no widely accepted reference corpora (see Table 1 for a list of the different corpora used to test the performances of different relation extraction algorithms). This lack of a common ground for evaluation makes it difficult to have a complete and fair comparison of all the approaches known in the literature. Another relevant issue concerns the evaluation methodology adopted for relation extraction. The MUC corpora, for instance, seem to be ideal as testbeds because they are provided not only with the correct solutions (e.g., filled templates), but also with the results obtained by other researchers who previously used them. However, the recent papers on relation extraction that use the MUC corpora have been almost always forced to adopt indirect evaluation measures for two related reasons: (i) MUC annotations are not done within texts but consist of external templates associated to each text; (ii) this makes it very difficult to automatically define a mapping between the learned patterns and the MUC templates. The only notable exception is the recent paper by [Chieu et al. 03].

Another important issue is the influence of the use of a coreference module on the performances of IE tools. Will a coreference module improve the performance of the IE tools? This issue will be coped with not during the first evaluation phase but only in the second evaluation phase.

² Even if in [Roth & Yih 02] also the results for Job Postings are included.

Terrorist events (MUC-4)	[Chieu et al. 03]
Management succession (MUC-6)	[Soderland 97, 99] [Yangarber et al. 00, Yangarber 03] [Chieu & Ng 02]
News articles	[Sudo et al. 01, 03]: Japanese articles [Roth & Yih 02a]: some hundreds articles taken from TREC-9 [Zelenko et al. 03]: 200 news articles from various sources
Aircraft crash (MUC-7 dryrun and training set)	[Català et al. 00, Català 03]
MEDLINE abstracts	[Skounakis et al. 03]

Table 1. Corpora used in different papers describing algorithms for explicit relation extraction.

3 EVALUATION IN DOT.KOM

In Dot.Kom we will adopt the evaluation measures that are standard in the IE community since the MUC conferences, i.e. precision, recall and F-measure.

We will perform an evaluation of the tools developed in the project using a rigorous approach as it is usual in the IE community. Considering the kind of tools developed so far in Dot.Kom, we will work on implicit relation identification. This is an area where neither the task nor the evaluation methodology is very well specified. Different researchers have used different strategies and therefore it is quite complex to compare systems and algorithms. What Dot.Kom intends to develop is therefore a standardization of the evaluation methodology for a number of corpora for implicit relation extraction. Seminar Announcements and Job Postings are examples of corpora we have already used in the past and on which we intend to invest, considering that they are among the most widely used scientific testbeds for Information Extraction. Another corpus that we will use in our experiments is that of Rental Ads [Soderland 99]. Some of these corpora contain a significant amount of errors and we will try to correct them as much as possible. This will be done partially manually and partially semi-automatically. In this respect an approach based on boosting, such as TIES, could be useful because boosting naturally allows for a form of “data cleaning”. As a matter of fact, while producing the classifiers, boosting also produces a ranking of examples in terms of how hard it was to classify them correctly, which basically corresponds to their probability of their being misclassified examples. We will investigate the use of this feature to help cleaning the data.

We will define a set of guidelines for evaluation that will be also distributed to the scientific community with the appropriate tools and resources for the evaluation when

possible. We will create a number of web pages where these guidelines and resources will be made available.

They include:

- *n*-fold cross validation experiments, where *n* should be generally equal to 10. All the experiments will have to be performed using a corpus partition. We will formally define the partitions of the corpora to be used in the evaluation. 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 select the documents (e.g., random split repeated *n* times vs. *n*-fold cross-validation). In addition, as it is well known different partitions can affect the system results, therefore we will establish the partitions to be used for the tests.
- Fragment evaluation. Errors in extraction can be considered 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. [Freitag 98] proposes three different criteria for matching reference instances and extracted instances. [DeSitter & Daelemans 03] present results for their algorithm for all the three criteria. We will define standard evaluation strategies to be used in the experiments.
- 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. We will contact MITRE, the owner of the software copyright, in order to allow the distribution of the code, if possible. Otherwise we will look for other scorers (e.g. the Gate's scorer).
- Definition of preprocessing tasks. Some of the preparation subtasks are important in determining the results of the algorithms. For example tokenization is often considered obvious and non problematic but it is not so [Habert et al. 98]. 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. We plan to use Annie, Gate's shallow IE system to provide such annotation³. This will also allow comparing 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 enough. The study of the learning curve is very important. Therefore all the evaluations will be carried on tracing the learning curve. We will define the strategy to be used for determining the learning curve for each corpus.

Dot.Kom can and will define the above guidelines and resources, but it seems clear to us that agreement in the international community is needed on them, therefore we are currently discussing the details with some well-known researchers (Nick Kushmerick, Ion Muslea, Dayne Freitag and Mary Elaine Califf among others). A paper that critically surveys the above-mentioned issues is planned for submission to LREC 2004.

³ www.gate.ac.uk

Concerning Dot.Kom specific corpora, we will make reference to those mentioned in deliverable D4.1 and establish the same criteria mentioned for the publicly available corpora.

An evaluation of relation extraction is not planned for the first release of the Dot.Kom IE tools, so we will not deal with such aspect of evaluation in this deliverable. We plan to produce a Dot.Kom working paper with a survey of the situation of the evaluation of relation extraction, which will be useful in later stages of the project.

3.1 USE OF ONTOLOGY

We should also take into consideration the fact that in Dot.Kom our aim is to evaluate IE tools in the context of KM applications. And in the interaction with the KM tools it would be useful if IE tools were able to exploit the information contained in the ontologies defined for KM purposes.

The use of ontologies will obviously have an impact on the evaluation measures and on the evaluation of how precisely (with regard to the concept hierarchy) instances are tagged by the IE tools. [Hahn & Schnattinger 1998, Resnik 1999] proposed two different procedures to measure the learning accuracy of a system with respect to a concept hierarchy. [Hahn & Schnattinger 1998] presents a measure of learning accuracy which considers various path distances within the concept hierarchy. The Concept Learning Accuracy (CLA) defined in [Maedche & Staab 00] is very similar in spirit to the Learning Accuracy presented in [Hahn & Schnattinger], but its definition is more concise as it does not distinguish between correct and incorrect 'predictions'.

[Resnik 1999] presents a measure of semantic similarity in an IS-A taxonomy based on the notion of shared information content rather than using the traditional edge-counting approach. However, it is important to mention that in order to make use of the measure proposed by Resnik, we would need to compute a probability for each concept in the ontology, which may turn out not to be a trivial task in itself. Resnik accomplishes this by calculating the relative frequency for a certain WordNet synset based on the occurrence of the corresponding words in a corpus.

Alternative methods for computing semantic similarity have been presented by Jiang and Conrath [1997], Lin [1998], Leacock and Chodrow [1998] and Hirst and St-Onge [2002].

4 CONCLUSIONS

In this deliverable we have outlined the evaluation strategy for IE in Dot.Kom. The plan presented above is quite ambitious and it will not be completed in few months. We expect that the full definition, resources and tools will be available by the end of the project. During the first evaluation cycle, we will focus on a restricted number of experiments based on a subset of the abovementioned resources and tools.

REFERENCES

- [Califf 98] Califf, M.E., *Relational Learning Techniques for Natural Language Information Extraction*, PhD Thesis, University of Texas, Austin, August 1998.
- [Català et al. 00] Neus Català, Nùria Castell, Mario Martín: “ESSENCE: a Portable Methodology for Acquiring Information Extraction Patterns”. In *Proceedings of 14th European Conference on Artificial Intelligence (ECAI-2000)*, pages 411-415, Berlin, Germany, August 2000.
- [Català 03] Neus Català: “Acquiring Information Extraction Patterns from Unannotated Corpora”. PhD Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, July 2003.
- [Chieu & Ng 02] Chieu, Hai Long, Hwee Tou Ng: “A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text” in *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI 2002)*, Edmonton, Canada, 2002.
- [Chieu et al. 03] Chieu, Hai Long, Hwee Tou Ng, Lee, Yoong Keok: “Closing the Gap: Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods” in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, 2003.
- [Ciravegna 01a] Ciravegna, F. “Adaptive information extraction from text by rule induction and generalisation.” In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, Washington, 2001.
- [Daelemans & Hoste 02] Daelemans, W., Hoste, V.: “Evaluation of Machine Learning Methods for Natural Language Processing Tasks” In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain, 2002.
- [Daelemans et al. 03] Daelemans, W., Hoste, V., De Meulder, F., Naudts, N.: “Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language” In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, Cavtat-Dubrovnik, Croatia, 2003.
- [De Sitter & Daelemans 03] De Sitter, A., Daelemans, W: “Information Extraction via Double Classification” In *Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining (ATEM 2003)*, Cavtat-Dubrovnik, Croatia, 2003.
- [Douthat 1998] Douthat, A. “The Message Understanding Conference scoring software user’s manual”. In *Proceedings of the 7th Message Understanding Conference* (1998). Available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manua.html.

- [Freitag 98] Freitag, D. *Machine Learning for Information Extraction in Informal Domains*, PhD Thesis, CMU, Pittsburgh, 1998.
- [Freitag & Kushmerick 00] D. Freitag, N. Kushmerick "Boosted Wrapper Induction". In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas, 2000.
- [Habert et al. 98] Habert, B., Adda, G., Adda-Decker, M., Boula de Mareuil, P., Ferrari, S., Ferret, O., Illouz, G., Paroubek, P.: "Towards Tokenization Evaluation". In *Proceedings of 1st International Conference on Language Resources and Evaluation (LREC-98)*, Granada, Spain, 1998.
- [Hahn & Schnattinger 1998] Hahn, U.; Schnattinger, K, "Towards text knowledge engineering", In *Proceedings of the 15th National Conference on Artificial Intelligence & 10th Conference on Innovative Applications of Artificial Intelligence (AAAI'98 / IAAI'98)*. Madison, Wisconsin, July 26-30, 1998. Menlo Park, CA; Cambridge, MA: AAAI Press / MIT Press, pp. 524-531, 1998.
- [Hirschman 1998] Lynette Hirschman, "The Evolution of Evaluation: Lessons from the Message Understanding Conferences", *Computer Speech and Language*, 12, pp. 281-305, 1998.
- [Hirst and St-Onge 2001] "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures" In *Proceedings of the Workshop on WordNet and Other Lexical Resources at Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA. 2001.
- [Hoste et al. 02] Hoste, V., Hendrickx, I., Daelemans, W., van den Bosch, A.: "Parameter Optimization for Machine-Learning of Word Sense Disambiguation" *Natural Language Engineering*, 8(4), pp. 311-325, 2002.
- [Jiang and Conrath 1997] Jiang, J. and Conrath, D. "Semantic similarity based on corpus statistics and lexical taxonomy" In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan. 1997.
- [Leacock and Chodrow 1998] Leacock, C. and Chodrow, M. "Combining local context and WordNet similarity for word sense identification", In *WordNet: An electronic lexical database*, MIT Press, 1998.
- [Lin 1998] Lin, D. "An information-theoretic definition of similarity" In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.
- [Maedche & Staab 2000] Maedche, Alexander and Staab, Steffen "Discovering Conceptual Relations from Text" In *Proceedings of 14th European Conference on Artificial Intelligence (ECAI-2000)*, Berlin, Germany, August 2000.
- [Makhoul et al. 1999] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., "Performance Measures for Information Extraction", In *Proceedings of the DARPA Broadcast News Workshop*. Herndon, Virginia, February 28 – March 3, 1999. Available at <http://www.nist.gov/speech/publications/darpa99/pdf/dir10.pdf>
- [MUC 1998] Message Understanding Conference Proceedings (MUC-7). 1998. Available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

- [Peshkin & Pfeffer 03] Peshkin, L., Pfeffer, A.: "Bayesian Information Extraction Network" In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, Acapulco, Mexico, 2003.
- [Resnik 1999] Philip Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", *Journal of Artificial Intelligence Research (JAIR)*, 11, pp. 95-130, 1999.
- [RISE 1998] *A Repository of Online Information Sources Used in Information Extraction Tasks*, <http://www.isi.edu/info-agents/RISE/index.html>, University of Southern California, Information Sciences Institute.
- [Roth & Yih 02a] Roth, D., Wen-tau Yih: "Probabilistic Reasoning for Entity and Relation Recognition" In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taiwan, August 2002.
- [Roth & Yih 02b] Roth, D., Wen-tau Yih: 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, 2002.
- [Roth & Yih 01] Roth, D., Wen-tau Yih: Relational Learning via Propositional Algorithms: An Information Extraction Case Study" In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, Washington, August 2001.
- [Skounakis et al. 03] Skounakis, M., Craven, M., Ray, S.: "Hierarchical Hidden Markov Models for Information Extraction" In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, Acapulco, Mexico, 2003.
- [Soderland 97] Soderland, S., *Learning text analysis rules for domain specific natural language processing*. PhD Thesis, University of Massachusetts, Amherst, Massachusetts, February 1997.
- [Soderland 99] Soderland, S. "Learning information extraction rules for semi-structured and free text" *Machine Learning* 34, 1 (1999), 233–272.
- [Sudo et al. 01] Sudo, K., Sekine, S., Grishman, R.: "Automatic Pattern Acquisition for Japanese Information Extraction" In *Proceedings of First International Conference on Human Language Technology Research (HLT 2001)*, San Diego, California, 2001.
- [Sudo et al. 03] Sudo, K., Sekine, S., Grishman, R.: "An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition" In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, 2003.
- [Sundheim 1995] Sundheim, B. "Overview of the results of the MUC-6 evaluation". In *Proceedings of the Sixth Message Understanding Conference*. 1995.

- [Will 1993] Will, C. A. "Comparing human and machine performance for natural language information extraction: results for English microelectronics from the MUC-5 evaluation". In *Proceedings of the Fifth Message Understanding Conference*. 1993.
- [Yangarber et al. 00] Yangarber, R., Grishman, R., Tapanainen, P. and Huttunen, Silja: "Automatic Acquisition of Domain Knowledge for Information Extraction" In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, 2000.
- [Yangarber 03] Yangarber, R.: "Counter-Training in Discovery of Semantic Patterns" In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, 2003.
- [Zelenko et al. 03] Zelenko, D., Aone, C., Richardella, A.: "Kernel Methods for Information Extraction" *Journal of Machine Learning Research*, 3, pp. 1083-1106, 2003.