Adaptive Information Extraction for Document Annotation in Amilcare

Fabio Ciravegna, Alexiei Dingli and Yorick Wilks
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello Street,
S1 4DP, Sheffield, UK +44-114-2221814
{fabio|alexiei|Yorick}@dcs.shef.ac.uk

Daniela Petrelli
Department of Information Studies,
University of Sheffield,
Regent Court, 211 Portobello Street,
S1 4DP, Sheffield, UK, +44-114-2222683
D.Petrelli@ sheffield.ac.uk

Categories & Subject Descriptors: I.2.7 [Artificial Intelligence**]: Natural Language Processing –** ***text analysis.***

General Terms: Management, Performance, Design, Experimentation, Human Factors, Theory

Amilcare is a tool for Adaptive Information Extraction (IE) designed for supporting active annotation of documents for the Semantic Web (SW). It can be used either for unsupervised document annotation or as a support for human annotation. Amilcare is portable to new applications/domains without any knowledge of IE, as it just requires users to annotate a small training corpus with the information to be extracted. It is based on $(LP)^2$, a supervised learning strategy for IE able to cope with different texts types, from newspaper-like texts, to rigidly formatted Web pages and even a mixture of them[1][5].

Adaptation starts with the definition of a tag set for annotation, possibly organized as an ontology. Then users have to manually annotate a small training corpus. Amilcare provides a default mouse-based interface called Melita, where annotations are inserted by first selecting a tag from the ontology and then identifying the text area to annotate with the mouse. Differently from similar annotation tools [4, 5], Melita actively supports training corpus annotation. While users annotate texts, Amilcare runs in the background learning how to reproduce the inserted annotation. Induced rules are silently applied to new texts and their results are compared with the user annotation. When its rules reach a (user-defined) level of accuracy, Melita presents new texts with a preliminary annotation derived by the rule application. In this case users have just to correct mistakes and add missing annotations. User corrections are inputted back to the learner for retraining. This technique focuses the slow and expensive user activity on uncovered cases, avoiding requiring annotating cases where a satisfying effectiveness is already reached. Moreover validating extracted information is a much simpler task than tagging bare texts (and also less error prone), speeding up the process considerably. At the end of the corpus annotation process, the system is trained and the application can be delivered. MnM [6] and Ontomat annotizer [7] are two annotation tools adopting Amilcare's learner.

In this demo we simulate the annotation of a small corpus and we show how and when Amilcare is able to support users in the annotation process, focusing on the way the user can control the tool's proactivity and intrusivity. We will also quantify such support with data derived from a number of experiments on corpora. We will focus on training corpus size and correctness of suggestions when the corpus is increased.

1. REFERENCES

[1] F. Ciravegna: "Adaptive Information Extraction from Text by Rule Induction and Generalisation" in *Proceedings of 17th IJCAI*, Seattle, August 2001.

[2] F. Ciravegna (2001): "Challenges in Information Extraction from Text for Knowledge Management", *IEEE Intelligent Systems and Their Applications* 16(6) 88-90.

[3] F. Ciravegna (2001c): "$(LP)^2$, an Adaptive Algorithm for Information Extraction from Web-related Texts" in *Proc. of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, Seattle

[4] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson and M. Vilain *Mixed-initiative development of language processing systems*. In Proc. of the Fifth Conference on ANLP, Washington, 1997.

[5] H. Cunningham, D. Maynard, V. Tablan, C. Ursu, K. Bontcheva: "Developing Language Processing Components with GATE", www.gate.ac.uk

[6] J.B. Domingue, M. Lanzoni, E. Motta, M. Vargas-Vera and F. Ciravegna: "MnM: Ontology driven semi-automatic or automatic support for semantic markup", submitted paper.

[7] S. Handschuh, S. Staab and F. Ciravegna: "S-CREAM - Semi-automatic CREAtion of Metadata", submitted paper.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.