# Active Document Enrichment using Adaptive Information Extraction from Text

**Fabio Ciravegna[1], Alexiei Dingli[1] and Daniela Petrelli[2]**
[1]Department of Computer Science [2]Department of Information Studies,
{F.Ciravegna|A.Dingli}@dcs.shef.ac.uk, D.petrelli@shef.ac.uk
University of Sheffield, Regent Court, 211 Portobello Street,
S1 4DP Sheffield, UK

The traditional process of document annotation for knowledge identification and extraction in the Semantic Web (SW) is complex and time consuming, as it requires manual annotation by domain experts. There is currently a strong interest in Text Mining technologies (and in particular in Human Language-based Technologies), for reducing the burden of text annotation for Knowledge Management (KM)[Maybury2001]. In this poster we present Melita, an annotation interface that uses Adaptive Information Extraction from texts for reducing the burden of text annotation. In Melita, adaptation starts with the definition of a scenario, including, among other information, a tag set for annotation (possibly organized as an ontology) and a corpus to be annotated. 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 [Day1997, Cunningham2001], Melita actively supports corpus annotation using Amilcare, an adaptive Information Extraction (IE) tool based on the $(LP)^2$ algorithm [Ciravegna2001]. While users annotate texts, the 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. If the IE based annotation becomes very reliable, the user can decide to let the IE system proceed automatically for further annotation. Melita provides non-intrusive and just in time support for annotation. It comes just in time because training is performed while user annotates the text. It is non-intrusive because user can fully customize the level of support the interface provides (pervasive, very active, active, lazy or very lazy).

In some experiments we have simulated the user annotation of two manually tagged corpora used for testing IE systems (the CMU seminar announcement [Freitag 1998] and the Austin Jobs corpus [Califf 1998]). Melita showed to be able to drastically reduce the quantity of user tagging. Using less than 30 texts for training it was able to reproduce correctly about 90% of the annotation including domain specific time expressions, with 30 texts specific location names, etc.

The work is carried on in the framework of the AKT project (http://www.aktors.org), an Interdisciplinary Research Collaboration (IRC) sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/N15764/01). AKT involves the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University (www.aktors.org). A description of Melita and a link to all its related publications can be found at www.dcs.shef.ac.uk/~alexiei/Melita.htm

## References

[Califf 1998] Mary E. Califf, Relational Learning Techniques for Natural Language IE, *Ph.D. thesis*, Univ. Texas, Austin, www.cs.utexas.edu/users/mecaliff

[Ciravegna 2001] F. Ciravegna: "Adaptive Information Extraction from Text by Rule Induction and Generalisation" in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001.

[Cunningham 2001] H. Cunningham, D. Maynard, V. Tablan, C. Ursu, K. Bontcheva: "*Developing Language Processing Components with GATE*", www.gate.ac.uk

[Day 1997] 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 Applied Natural Language Processing, Washington, 1997.

[Freitag 1998] Dayne Freitag, `Information Extraction from HTML: Application of a general learning approach', *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[Maybury 2001] M. Maybury (ed.): Proc. of the 2001 *EACL/ACL Workshop on Human Language Technology and Knowledge Management*, at the at the 39th meeting of the ACL, July 6-11, 2001, Toulouse, France