Benchmarking Textual Annotation Tools for the Semantic Web

Diana Maynard

Dept of Computer Science, University of Sheffield, Sheffield, UK {diana}@dcs.shef.ac.uk

Abstract content

This paper investigates methods and results for benchmarking textual annotation tools. We define first some criteria for benchmarking, including both performance and usability issues, and examine those factors which are particularly important for a user to be able to determine which is the most suitable tool for their use. We then perform a series of experiments on a set of annotation tools, and discuss the results, finally drawing some conclusions about the future of annotation tools.

Textual annotation tools are used to populate ontologies with instances from text, and/or to annotate text with conceptual information from an ontology. This task forms an important part of ontology creation and management, by enabling us to combine and associate existing ontologies, perform more detailed analysis of the text, and to extract deeper and more accurate knowledge. This in turn leads to the development or enhancement of many different kinds of applications such as semantically-enhanced information retrieval, question answering, data gathering, business intelligence, and so on. However, there exists a variety of annotation tools which have largely been developed for specific purposes within research projects. It is difficult for a user to understand the differences between these tools and to decide which – if any – of them is most appropriate to his or her needs. This paper outlines some research we have performed in the context of the KnowledgeWeb Network of Excellence, in order to determine some guidelines for benchmarking annotation tools and to aid the user in finding the most appropriate tool for their needs.

The tools we have chosen to examine here are the following: MnM (Motta et al., 2002), OntoMat (Handschuh et al., 2002), GATE (Cunningham et al., 2002), KIM (Popov et al., 2004), and Magpie (Domingue et al., 2004). These have been chosen for a number of reasons. First, they all perform in some way annotation of textual data with respect to an ontology, and are all XML-based. Second, they are all open source, readily available and do not require extensive training to use. Third, they have also been chosen for their diversity: MnM is a very basic tool which was developed some years ago largely as proof of concept. It is no longer maintained, so is a good reflection of the initial state-of-the-art, and in some way can act as a baseline. GATE and KIM are quite generic tools, which are actively maintained and developed, and are used as the basis of many other annotation systems. OntoMat and Magpie were both developed for quite specific tasks rather than for just general annotation, so it is interesting to compare them with the more generic tools.

We first discuss the problem of benchmarking such tools, and the different kinds of requirements that a user might have. We can break this down into issues concerning performance, scalability, usability, and interoperability. It is important to note that, contrary to most research projects which aim simply to produce tools with as high an accuracy as possible, in real life other factors such as usability and interoperability may be more important. The nature of natural language processing tasks means that the vast majority are on the one hand intrinsically difficult for a machine to perform, and yet on the other hand time-consuming and tedious for a human to carry out, especially on a large scale. Thus the most acceptable solution for a user of an annotation tool may simply be to have a tool to aid the human to perform the task and reduce time and effort, but which may need some manual post-editing or intervention. This means that factors such as usability and suitability for the task, time taken to learn to use and/or adapt the tool, and interoperability with other tools, may be higher priority than the actual performance of annotation. In this paper, we investigate such issues in more detail, using examples from real annotation tools and situations.

Second, we investigate the problem of performance of annotation tools. While traditional information extraction tools can be evaluated quite adequately with metrics such as Precision and Recall, this is not the case for ontology-based tools which may offer solutions with varying degrees of correctness. For example, classifying a Researcher as a Professor is not necessarily as wrong as classifying a Lecturer as a Location, so some credit could be given for a ”near miss”. In this paper we propose the use of an evaluation metric, the BDM (Maynard, 2005; Maynard et al., 2006), which takes into account the ontological similarity of the key (gold standard) and response (system) result of the annotation, in order to better evaluate its performance. We discuss the appropriateness of this metric and compare it with other metrics such as a straightforward distance-based metric and a flat metric such as traditional Precision and Recall. We compare the performance of various annotation tools using such a metric and show how it provides a better benchmark than traditional metrics.

Finally, we investigate the issues of interoperability and scalability of annotation tools, which have an important bearing on the usefulness of such tools with respect to their use in a real world industrial setting rather than simply in a research environment, i.e. where huge volumes of data and/or large ontologies are required, and where the tool is not used as standalone but may be required to integrate at least its results with that of other semantic web tools.

We conclude the paper with a summary of the tools and their appropriateness for different users and situations. Having investigated various aspects of non-performance-related issues such as general usability, accessibility, interoperability and so on, we find that we cannot draw any hard and fast rules, as clearly it depends quite precisely on the user’s requirements as to which tool best fulfils their needs according to this aspect. For example, GATE is probably the most fully-featured tool in terms of accessibility, allowing the user to have control over fonts, colours, text size. However, its graphics are quirky and unclear, and actions are almost entirely mouse-driven, forcing users to perform nearly all functions using a mouse rather than keyboard or alternative input device. On the other hand if a very simple, easy-to-learn annotation tool is required, then something like Magpie might be more appropriate. Some problems can of course be overcome if time and effort is available to be invested: for example it might be possible to adapt a tool to use a different ontology format (Magpie and MnM in the versions we tested were not compatible with OWL, for example).

With respect to evaluating performance, it is important to note, as shown in the full paper, that there is a variety of ways in which to measure performance, and the choice of measure may be as dependent on the tool itself as on the aim of the evaluation. For example, although MnM annotates texts with instances from an ontology (as do the other textual annotation tools) it is almost impossible to run – and therefore to evaluate it – on an ontology of any significant size (with more than a very small handful of concepts). As the results from the performance experiments show, however, we can see quite different results on the same tools depending on which kind of measure we use. For example, we see with GATE and KIM that using the BDM metric rather than standard Precision and Recall gives us a much better idea of the strengths and weaknesses of the tools. The difference between their performance is smaller with the BDM than with traditional metrics, because KIM actually finds many entities correctly but misclassifies them. Since partial credit is given for minor misclassifications, KIM’s scores is therefore not only better with the BDM, but more similar to GATE’s than with the traditional metrics, as GATE has respectively fewer misclassifications but more missing and spurious entities. In terms of quality of performance on the dataset used, we find that GATE outperforms the other tools when using various Machine Learning algorithms, with KIM a close second, while MnM struggles to perform on any kind of unstructured text. Even when using datasets and ontologies most appropriate to the tool in question, we still find that the other tools are some way behind in performance. We reiterate, however, that performance is not the only consideration, and also that some tools may be designed to be used in conjunction with human annotation than solely for automatic extraction.

It is clear from the interoperability investigation that most annotation tools are not really designed with interoperability in mind, especially as far as the Semantic Web is concerned. Most current annotation tools are based on legacy information extraction tools which do not take ontologies into account, and have in fact been adapted to perform ontology-based annotation. For this reason, most of them were not developed with interoperability concerns in mind, and the effort to make this possible can be quite substantial. In fact, only GATE and KIM were found to be truly interoperable in terms of OWL import and export, and even then, some problems were found.

As far as scalability goes, most of the tools examined here were designed originally as small scale research prototypes rather than tools for large scale annotation. Of these, Magpie, GATE and KIM are the only tools which have really been designed for general purpose use and/or have been adapted for dealing with large scale real world applications. Both GATE and KIM perform reasonably well when used with large ontologies and data sets, although GATE can be quite slow with large datasets or with complex applications (for example, there are sometimes problems when using massive gazetteer lists).

Finally, we also take a look at the future of annotation tools, focusing on the difficulties and problems still to be surmounted. For example, while annotation tools have shown much success in real world applications such as Del.ici.ous, Flickr, digital libraries such as Perseus, Garlik (which mines data about consumers present in various sources including the web), Fizzback (which provides real-time customer feedback from SMS and email feeds) and so on, there remain several reasons why semantic annotation is not more widespread. For example, it is still difficult and time-consuming to produce annotations in open domains. We propose a solution combining human annotation with automatic methods. This in itself is far from new, but what is required (and currently lacking) is a clear statement of how to specify and implement new annotation tasks, especially those oriented towards non-HLT experts.

1. References

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

J. Domingue, M. Dzbor, and E. Motta. 2004. Magpie: Supporting Browsing and Navigation on the Semantic Web. In N. Nunes and C. Rich, editors, *Proceedings ACM Conference on Intelligent User Interfaces (IUI)*, pages 191–197.

S. Handschuh, S. Staab, and F. Ciravegna. 2002. S-CREAM — Semi-automatic CREAtion of Metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 358–372, Siguenza, Spain.

D. Maynard, W. Peters, and Y. Li. 2006. Metrics for evaluation of ontology-based information extraction. In *WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON)*, Edinburgh, Scotland.

D. Maynard. 2005. Benchmarking ontology-based annotation tools for the semantic web. In *UK e-Science Programme All Hands Meeting (AHM2005) Workshop "Text Mining, e-Research and Grid-enabled Language Technology"*, Nottingham, UK.

E. Motta, M. Vargas-Vera, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. 2002. MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 379–391, Siguenza, Spain.

B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. 2004. KIM – Semantic Annotation Platform. *Natural Language Engineering*.