LREC 2010 Tutorial

ANNOTATION INTERCHANGE

Graham Wilcock, University of Helsinki

The tutorial presents a practical introduction to interchange of linguistic annotations between different NLP frameworks. Concrete examples are based on four widely-used annotation frameworks: GATE, UIMA, OpenNLP and WordFreak. Each framework is demonstrated by annotating a short example text, and the different annotation formats are described and compared. Most of the frameworks use a stand-off XML annotation format in order to support multiple levels of possibly overlapping annotations. The structure of the stand-off annotations is basically similar, but each specific format has its own distinct details. In practice, the differences between the formats are an obstacle to interoperability between different frameworks.

The tutorial shows how to interchange annotations between different frameworks using XML transformations specified by XSLT stylesheets. One reason XML is so widely used as a standard for data interchange is that it is supposed to be easy to change the format details using XSLT. This was summed up about ten years ago in the joke *XML means never having to say you're sorry*. Although the formats of stand-off annotations are more complex than in-line annotations, the tutorial shows that the joke is not wrong: stand-off XML annotations in one specific format can be transformed into stand-off annotations in another specific format using XSLT.

Schedule

The tutorial will have two parts, separated by a coffee break. Each part will include time for questions and discussion.

Part 1: Annotation Frameworks and Formats (with demonstrations)
This part introduces four widely-used annotation frameworks: OpenNLP, WordFreak, GATE, and UIMA. Each framework is demonstrated by annotating a short example text.
OpenNLP produces in-line annotations in a plain text format. The other frameworks produce stand-off XML annotations in three different formats (WordFreak XML, GATE XML and UIMA XML). The different annotation formats are described and compared. The example annotation files are made available for participants to use in Part 2.

Part 2: Annotation Transformations and Interchange (with practical exercises)
This part introduces a set of XSLT stylesheets that transform annotations between the formats described in Part 1. Three ready-to-use stylesheets are explained in detail, and participants use them to transform the example annotations from WordFreak to OpenNLP format, from GATE to WordFreak, and from WordFreak to UIMA. In further practical exercises, parts of existing stylesheets can be merged to produce new transformations, for example to create stylesheets that transform directly between GATE and UIMA.

Requirements

In addition to demonstrations of the frameworks, the tutorial includes practical exercises in which participants can transform the example annotations between different formats using XSLT stylesheets. No previous knowledge of XSLT is required in order to use the supplied stylesheets, but participants will be encouraged to create additional stylesheets for further transformations and XSLT skills would then be useful. Participants who wish to do the practical exercises will need a laptop with Java, and will need to download the examples from the internet.

Tutorial speaker

Graham Wilcock is Adjunct Professor of Language Technology at University of Helsinki (http://www.ling.helsinki.fi/~gwilcock). He has a PhD in computational linguistics from University of Manchester. He previously worked in industry as a software engineer with ICL in England and as a researcher with Sharp Corporation in Japan. His research topics have included machine translation, HPSG, natural language generation, spoken dialogue systems, XML for NLP, OWL ontologies, and linguistic annotation tools. He has been co-organizer and co-chair of the following workshops:

2nd Workshop on Natural Language Processing and XML (COLING-2002 Taipei),
Language Technology and the Semantic Web (EACL-2003 Budapest),
RDF and OWL in Language Technology (ACL-2004 Barcelona),
10th European Workshop on Natural Language Generation (Aberdeen 2005),
Multi-Dimensional Markup in Natural Language Processing (EACL-2006 Trento),
The Linguistic Annotation Workshop (ACL-2007 Prague).

He is the author of a textbook *Introduction to Linguistic Annotation and Text Analytics* (Morgan & Claypool, 2009) that describes the annotation frameworks, XML formats and XSLT stylesheets used in the tutorial.