Text Analysis with GATE

Presenter: Dr Diana Maynard, University of Sheffield
Address: Dept of Computer Science, Regent Court, 211 Portobello, Sheffield S1 4DP
Email: d.maynard@sheffield.ac.uk
Website: http://www.dcs.shef.ac.uk/~diana
Phone: +44 114 222 1938

Type of tutorial: half day

Abstract: This tutorial takes a detailed view of key text mining tasks (introduction to NLP, linguistic pre-processing, entity and relation recognition, semantic annotation, indexing and multi-paradigm search) of textual content. It will cover both the latest state-of-the-art research and selected established methods and tools. It will show how text analysis tools and techniques can be used to assist with semantic search by providing extra information that is not explicit within the text itself. Finally, some real world applications of the technology will be demonstrated, showcasing the entire pipeline of data collection, annotation, indexing, search and visualization.

Target Audience: The target audience will consist of attendees from any background looking to perform textual analysis on their data or interested to learn how incorporating NLP techniques might help them. No previous knowledge of NLP, GATE, programming or text analysis is required.

Learning Outcomes: This tutorial will introduce the attendees to text analytics and outline key approaches to various linguistic processing tasks. It will cover both the latest state-of-the-art research and also selected established methods and tools. The attendees will learn what text analytics can provide and why it might be useful to them, and will get an insight not only into the available tools and techniques, but also into how these can be easily used to create new applications, with real-life examples based on a variety of domains including climate change, politics, journalism and bioinformatics. They will also learn the basics of the GATE toolkit. Note that in a half-day session it is not really practical for the participants to practise using GATE during the tutorial itself, but adequate information will be given for them to use it themselves later, including hands-on materials (sample applications, corpora etc.) and links to existing step-by-step tutorials in getting started with GATE. Some tried and tested live demonstrations will be given, however, and the participants will be able to experiment with some of the online demonstration tools during the session, which are built on top of GATE.

The tutorial aims to:

- Provide attendees with practical and up-to-date knowledge of NLP and its use in applications that require the analysis of textual content;
- Provide attendees with knowledge about the inherent problems involved in analysing textual data, including problems specific to social media, and potential solutions to these problems;
- Give attendees demonstrations of methods and tools for semantic annotation and related activities, using an established toolkit, GATE;
- Stimulate discussions on what other NLP tools and techniques they might find helpful in their work.

Tutorial Description:
Free text makes up a large proportion of the vast amounts of information generated by modern society, and search engines are often exceptionally good at finding, indexing and searching this. However, the rise of the Semantic Web and the publishing of increasingly large amounts of structured and interlinked data now means that useful information is distributed across multiple sources and in a variety of formats, which cannot be easily reconciled by these search engines as it is not amenable to free text search. Hence, questions which we may wish to ask of society's collective knowledge cannot be easily answered if this information is not explicitly mentioned in the document, or if the information needs to be combined from multiple documents or knowledge sources.

Automatic approaches to text analytics can help to overcome the knowledge acquisition bottleneck, by extracting relevant information from the Web and other documents, and transforming it into a machine-processable representation. This is crucial if information retrieval is to be an intelligent process that goes beyond the keyword-based approach. A semantic approach to search enables information which is not explicitly mentioned in the document to be made available. The field of NLP has matured over the last decade to a point at which robust and scalable applications are not only possible, but crucial for use in such applications.

GATE Mimir enables indexing and search of documents not only as a keyword-based approach, but also by means of semantic annotations and linked open data. The resulting multi-paradigm index allows us to search across multiple information sources in order to answer questions which are either infeasible or impossible to answer using current web search engines. On top of this, GATE Prospector enables complex data visualisations to be incorporated.

The tutorial will introduce the key concepts of text analytics, and will cover tools and techniques for linguistic pre-processing (tokenisation, sentence splitting, lemmatisation, morphological analysis, etc.), Named Entity recognition, relation extraction, semantic annotation and semantic indexing and search. We will demonstrate each element with practical examples using GATE, a widely used, open source and freely available toolkit for text analytics. Semantic indexing and search techniques will be introduced and demonstrated using GATE Mimir and Prospector, and examples of visualisations on top of the annotated data will be shown (for example, how we can examine the different political topics discussed most frequently in different parts of the country).

Tutorial Outline:
The tutorial will be divided into 3 sections:

1. **Introduction:** We will first introduce the concept of text analytics and why it is useful for semantic search. We will also introduce GATE, the toolkit we shall be using as an example throughout the tutorial.
2. **Linguistic Processing and Semantic Annotation with GATE:** In this section, we will describe the various linguistic components typically required in an Information Extraction system, and show how to gradually build up a system, using GATE. We will also introduce the concept of semantic annotation using information from ontologies and linked open data. Finally, we will discuss some typical problems associated with text analysis tools and how these might be overcome.
3. **Semantic Indexing and Search:** In this section, we will show how the semantic annotations produced can be used to index the documents and how this can then be used for more complex multi-paradigm searching and visualisation using GATE Mimir and Prospector.

Tutorial Logistics/Materials: The tutorial will be taught using a combination of slides and live demonstration using the GATE toolkit and other web-based applications. The attendees will be given a copy of the slides and will also have access to hands-on materials, should they wish to try out some tools or demos themselves later. They will be able to try out the demos themselves also during the tutorial. Supporting material such as references and useful links will also be provided on a dedicated webpage for the tutorial.

Bio of presenter: Dr. Diana Maynard is a Research Fellow at the University of Sheffield, UK. She received a PhD from Manchester Metropolitan University in 2000 on the topic of automatic term extraction, and for the last 16 years has been leading the linguistic development of Sheffield's open-source multilingual Information Extraction tools, leading research teams in a number of UK and EU projects. Her main interests are in text mining, Information Extraction, opinion mining, social media analysis and Semantic Web technology. She is currently involved in the development of the social media analysis tools in GATE, has developed a number of opinion mining tools, and has both published widely and given a number of invited talks and tutorials on these topics. She is co-chair of the annual GATE training courses, teaching modules on Basic and Advanced Information Extraction, Semantic Annotation, Opinion Mining and Social Media Analysis. She was co-chair of the ISWC Semantic Web Challenge from 2010-2012, has been area chair for NLP at numerous semantic web conferences, organised a number of national and international conferences, workshops and tutorials, given keynote speeches at international conferences, invited talks, tutorials, lectures and courses on a number of NLP and Semantic Web-related topics. She has recently given tutorials at ISWC, EKAW, LREC, STIL, IADS, and the Sentiment Analysis Symposium, and has taught courses on GATE and text mining at a variety of summer schools. In addition to regular consultancy work, she is also technical advisor to several projects and companies, specialising in use of NLP, social media analytics and opinion mining.