Travelling through time with corpus annotation software

Paul Rayson
UCREL, Lancaster University

Vast quantities of searchable material are being created in electronic form through large digitisation initiatives currently underway e.g. Open Content Alliance[1], Google Book Search[2], Early English Books Online[3]. These initiatives are largely focussed on historical or more recent out-of-copyright material. As well as image-based digitisation, transcription and OCR-scanning techniques produce text-based materials and these will facilitate new methods for research. Annotation, typically at the part-of-speech (POS) level is carried out on modern corpora for linguistic analysis, information retrieval and natural language processing tasks such as named entity extraction. Increasingly researchers will carry out similar tasks on historical data (Nissim et al, 2004). However, historical data is considered noisy for these tasks. In this talk I will highlight the problems faced when applying corpus annotation tools trained for modern language data to historical data. Annotation tools such as POS taggers are generally robust on modern data across a number of registers and genres (Leech and Smith, 2000), but less is known about their accuracy on historical data. Spelling issues tend to create relatively minor (though still complex) problems for taggers applied to modern text, for example hyphenation and full stops in relation to tokenisation. However, different spelling conventions, compositing practices and morpho-syntactic customs as well as 'misspelling' in historical data can be expected to reduce the accuracy of the same tools when they are applied historically.

In an information retrieval setting, solutions explored so far have typically employed fuzzy searching techniques to improve retrieval (Pilz et al, 2006) and a cross-language approach (Koolen et al, 2006). Corpus linguistics researchers have adopted an approach of adding historical variants to the POS tagger's lexicon, for example in TreeTagger annotation of GerManC (Durrell et al, 2006), or 'back-dating' the lexicon in the Constraint Grammar Parser of English (ENGCG) when annotating the Helsinki corpus (Kytö and Voutilainen, 1995).

In previous research, we have highlighted the requirement to evaluate the coverage of language resources (such as lexicons embedded in annotation tools) both synchronically and diachronically (Piao et al, 2004). In addition to retraining the annotation tools and the lexicons they contain, a further key consideration is altering the taxonomies that are to be employed in an historical context, for example (i) changing the POS tagsets embedded within POS taggers to reflect changes in grammar over time (Britto et al, 1999; Kytö and Voutilainen, 1995) and (ii) changes in meaning over time require careful consideration of the applicability of sense distinctions and hierarchical structures of modern semantic tagsets (Archer et al, 2004).

Our studies have mainly focussed on English corpus annotation tools and in dealing with the problem of spelling variation in historical corpora. In this talk I will highlight our proposed solution which incorporates a corpus pre-processor for detecting historical spelling variants and inserting modern equivalents alongside them (Rayson et al, 2006). This enables retrieval as well as annotation tasks and to some extent avoids the need to retrain each annotation tool that is subsequently applied to the corpus. The modern taggers can then be applied to the modern spelling equivalents that have been inserted rather than the historical variants.

References

Archer, D., Rayson, P., Piao, S., McEnery, T. (2004). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In Williams G. and Vessier S. (eds.) *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004)*, Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Volume III, pp. 817-827.

Britto, H. Galves, C., Ribeiro, I., Augusto, M., and Scher, A. (1999) Morphological Annotation System for Automatic Tagging of Electronic Textual Corpora: from English to Romance Languages. *Proceedings of the 6th International Symposium of Social Communication*. Santiago, Cuba, pp.582-589.

[1] http://www.opencontentalliance.org/

[2] http://books.google.com/

[3] http://eebo.chadwyck.com/home

Durrell, M., Bennett, P., Ensslin, A. (2006). Towards a Methodology for Constructing and Annotating Historical Corpora: Tackling Structural and Lexical Variability in Early Modern German Newspaper Texts, *4th Days of Swiss Linguistics Conference*, Basel, Switzerland, November 2006.

Koolen, M., Adriaans, F., Kamps, J., and de Rijke, M. (2006). A Cross-Language Approach to Historic Document Retrieval, In *Proceedings 28th European Conference on Information Retrieval (ECIR 2006)*, LNCS 3936, pages 407-419, April 2006.

Kytö, M. and Voutilainen, A. (1995). Applying the Constraint Grammar Parser of English to the Helsinki Corpus. *ICAME Journal* 19, pp. 23 – 48.

Leech, G. and Smith, N. (2000). Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging. Accessed 6th March 2007. http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2postag_manual.htm

Nissim, M., Matheson, C. and Reid, J. (2004). Recognising Geographical Entities in Scottish Historical Documents. *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*.

Piao, S. L., Rayson, P., Archer, D., McEnery, T. (2004). Evaluating Lexical Resources for A Semantic Tagger. In *proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 26-28 May 2004, Lisbon, Portugal, Volume II, pp. 499-502.

Pilz, T., Luther, W., Fuhr, N., Ammon, U. (2006). Rule-based search in text databases with non-standard orthography, *Literary and Linguistic Computing*, 21 (2), pp. 179-186.

Rayson, P., Archer, D., Baron, A. and Smith, N. (2006). Tagging historical corpora - the problem of spelling variation. In *proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science*, Schloss Dagstuhl, Wadern, Germany, December 3rd-8th 2006.