Computational support for corpus analysis work flows: The case of integrating automatic and manual annotations

Richard Eckart de Castilho[1], Mônica Holtz[2], and Elke Teich[2]

[1] Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt, Germany
eckartde@tk.informatik.tu-darmstadt.de

[2] Institut für Sprach- und Literaturwissenschaft
Technische Universität Darmstadt, Germany
<lastname>@linglit.tu-darmstadt.de

Corpus-based linguistic research relies to a considerable degree on automatic methods of text processing (e.g., sentence segmentation, tokenization) and annotation (e.g., part-of-speech (PoS) tagging, syntactic phrase detection/categorization). While a corpus annotated at shallow levels of linguistic organization (such as PoS or syntactic phrases) is a very valuable resource for many tasks of linguistic analysis (e.g., collocations, word lists, PoS distributions etc), in many contexts it is desirable to have available explicit functional-grammatical or semantic information as well (cf. Teich (2009)). Since there exist no sufficiently reliable automatic methods for annotation in terms of such more abstract linguistic features, typically annotation must be carried out manually, supported by special-purpose annotation tools (e.g., MMAX2 (Müller & Strube, 2006), ExMaralda (Schmidt, 2005), UAM Corpus Tool (O'Donnell, 2008), RST-Tool (O'Donnell, 2000)).

Due to an increased interest in more sophisticated corpus processing, linguists as well as computer scientists build up processing pipelines for their analysis tasks. This, in turn, raises the issue of integrating/harmonizing different types of annotation that have possibly been produced by different tools (cf. frameworks such as GATE (Cunningham et al., 2002) or Apache UIMA (Ferrucci & Lally, 2004)). However, the issue of integrating automatic and manual annotations has, to our knowledge, not been explicitly addressed.

In this paper, we present a computationally supported work flow for integrating automatic and selective manual annotation. The work flow proceeds in the following steps. Given a corpus containing a set of documents, step (1) performs a basic automatic analysis (tokenization, lemmatization, PoS-tagging, etc). Based on the results of this analysis, step (2) selects candidate units for further, manual annotation by means of query. Step (3) extracts these units from the different source documents included in the corpus and aggregates them into a single document convenient for manual annotation. Step (4) merges the manually annotated units back into the original corpus.

The computational basis for this work flow is provided by AnnoLab (Eckart, 2006; Eckart & Teich, 2007), a modular extensible framework for managing text corpora annotated at multiple levels of linguistic organization, so called multilayer annotations. Each layer is represented in an XML document and the different layers are connected to the text data via stand-off references. It uses Apache UIMA to orchestrate linguistic processing pipelines. We have developed additional plug-ins to AnnoLab to export an automatically annotated corpus to external query tools and manual annotation tools as well as merge manually created annotations back into the corpus. To ensure a correct merging, stand-off information is maintained during the whole process by automatically adding stand-off information as extra annotations in the external tools. In case the stand-off anchors become invalid, e.g., because errors in the corpus have been corrected while manual annotation was in progress, we use a simple string-searching approach to locate the annotated sentence in the document.

In our talk, we present this work flow as well as the relevant parts of the AnnoLab system and show its application in a concrete corpus analysis scenario. The application is register analysis (cf. Halliday (1985a,b); Halliday & Hasan (1989)) of a 17 million words corpus of English scientific texts from different domains (Teich & Fankhauser, to appear; Teich & Holtz, in press). Here, we first use AnnoLab to run a processing pipeline that extracts the text from the corpus source files (PDF and HTML), creates PoS and lemma annotations employing TreeTagger (Schmid, 1994) and exports the annotated corpus to IMS-CWB (Christ, 1994). Then we use IMS-CWB's query tool CQP to locate and extract units for manual analysis. Finally, using two AnnoLab plug-ins, we convert the query results (typically a set of sentences) into a project for the UAM Corpus Tool, which is employed for selective manual analysis of functional-grammatical features, and merge the manually created annotations back into the corpus.

Integrating automatic and manual (selective) annotation is an issue in many contexts of corpus-based linguistic research. The method we have developed exploits automatic analysis tools and querying to quickly locate, aggregate and annotate candidate linguistic units for manual analysis. A consequent stand-off approach maintaining stand-off information across various tools allows to merge manually created annotations back into the corpus. A fallback simple string-searching strategy was suggested to handle changes to the corpus. The method can be improved by using a more sophisticated fallback strategy, e.g., employing edit distance.

Acknowledgements

The project LingPro (Linguistische Profile interdisziplinärer Register) is funded by the *Deutsche Forschungsgemeinschaft* (DFG) under grant TE 198/1-1.

Bibliography

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd Conference on Computational Lexicography and Text research (COMPLEX 94)*, pp. 23–32. Budapest, Hungary.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (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*.

Eckart, R. (2006). Towards a modular data model for multi-layer annotated corpora. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 183–190. Sydney, Australia: Association for Computational Linguistics. Available from: `http://www.aclweb.org/anthology/P/P06/P06-2024`.

Eckart, R. & Teich, E. (2007). An XML-based data model for flexible representation and query of linguistically interpreted corpora. In Rehm, G., Witt, A., & Lemnitzer, L., (Eds.), *Data Structures for Linguistic Resources and Applications – Proceedings of the Biennial GLDV Conference 2007*, pp. 327–336. Tübingen, Germany: Gunter Narr Verlag Tübingen.

Ferrucci, D. & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Halliday, M. A. K. (1985a). *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. A. K. (1985b). *Spoken and Written Language*. Victoria: Deakin University.

Halliday, M. A. K. & Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Müller, C. & Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *English Corpus Linguistics, Vol.3*, pp. 197–214.

O'Donnell, M. (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pp. 253–256. Mitzpe Ramon, Israel.

O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the ACL-08:HLT Demo Session (CompanionVolume)*, pp. 13–16. Columbus, Ohio: Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schmidt, T. (2005). EXMARaLDA und Datenbank ‚Mehrsprachigkeit' – Konzepte und prakische Erfahrungen. In *Interdisciplinary studies on information structure*, pp. 21–42. SFB 632, Universität Potsdam.

Teich, E. (2009). Linguistic computing. In Halliday, M. & Webster, J., (Eds.), *Companion to Systemic Functional Linguistic*, Chapter 7. London: Continuum.

Teich, E. & Fankhauser, P. (to appear). Exploring a corpus of scientific texts using data mining. In Gries, S., Davies, M., & Wulff, S., (Eds.), *Selected Papers from the American Conference on Corpus Linguistics (AACL) 2008, Provo, Utah*. Amsterdam: Rodopi.

Teich, E. & Holtz, M. (in press). Scientific registers in contact: An exploration of the lexico-grammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics*.