NLP and the humanities: the revival of an old liaison

Franciska de Jong
University of Twente
Enschede, The Netherlands
fdejong@ewi.utwente.nl

Abstract

This paper present an overview of some emerging trends in the application of NLP in the domain of the so-called Digital Humanities and discusses the role and nature of metadata, the annotation layer that is so characteristic of documents that play a role in the scholarly practises of the humanities. It is explained how metadata are the key to the added value of techniques such as text and link mining, and an outline is given of what measures could be taken to increase the chances for a bright future for the old ties between NLP and the humanities. There is no data like metadata!

1 Introduction

The humanities and the field of natural language processing (NLP) have always had common playgrounds. The liaison was never constrained to linguistics; also philosophical, philological and literary studies have had their impact on NLP , and there have always been dedicated conferences and journals for the humanities and the NLP community of which the journal Computers and the Humanities (1966-2004) is probably known best. Among the early ideas on how to use machines to do things with text that had been done manually for ages is the plan to build a concordance for ancient literature, such as the works of St Thomas Aquinas (Schreibman et al., 2004). which was expressed already in the late 1940s. Later on humanities researchers started thinking about novel tasks for machines, things that were not feasible without the power of computers, such as authorship discovery. For NLP the units of processing gradually became more complex and shifted from the character level to units for which string processing is an insufficient basis. At some stage syntactic parsers and generators were seen as a method to prove the correctness of linguistic theories. Nowadays semantic layers can be analysed at much more complex levels of granularity. Not just phrases and sentences are processed, but also entire documents or even document collections including those involving multimodal features. And in addition to NLP for information carriers, also language-based interaction has grown into a matured field, and applications in other domains than the humanities now seem more dominant. The impact of the wide range of functionalities that involve NLP in all kinds of information processing tasks is beyond what could be imagined 60 years ago and has given rise to the outreach of NLP in many domains, but during a long period the humanities were one of the few valuable playgrounds.

Even though the humanities have been able to conduct NLP-empowered research that would have been impossible without the the early tools and resources already for many decades, the more recent introduction of statistical methods in langauge is affecting research practises in the humanities at yet another scale. An important explanation for this development is of course the wide scale digitisation that is taken up in the humanities. All kinds of initiatives for converting analogue resources into data sets that can be stored in digital repositories have been initiated. It is widely known that ”*There is no data like more data*” (Mercer, 1985), and indeed the volumes of digital humanities resources have reached the level required for adequate performance of all kinds of tasks that require the training of statistical models. In addition, ICT-enabled methodologies and types of collaboration are being developed and have given rise to new epistemic cultures. Digital Humanities (sometimes also referred to as Computational Humanities) are a trend, and digital scholarship seems a prerequisite for a successful research career. But in itself the growth of digi-

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 10–15,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

tal resources is not the main factor that makes the humanities again a good testbed for NLP. A key aspect is the nature and role of metadata in the humanities. In the next section the role of metadata in the humanities and the the ways in which they can facilitate and enhance the application of text and data mining tools will be described in more detail. The paper takes the position that for the humanities a variant of Mercer's saying is even more true. *There is no data like metadata*!

The relation between NLP and the humanities is worth reviewing, as a closer look into the way in which techniques such as text and link mining can demonstrate that the potential for mutual impact has gained in strength and diversity, and that important lessons can be learned for other application areas than the humanities. This renewed liaison with the now digital humanities can help NLP to set up an innovative research agenda which covers a wide range of topics including semantic analysis, integration of multimodal information, language-based interaction, performance evaluation, service models, and usability studies. The further and combined exploration of these topics will help to develop an infrastructure that will also allow content and data-driven research domains in the humanities to renew their field and to exploit the additional potential coming from the ongoing and future digitisation efforts, as well as the richness in terms of available metadata. To name a few fields of scholarly research: art history, media studies, oral history, archeology, archiving studies, they all have needs that can be served in novel ways by the mature branches that NLP offers today. After a sketch in section 2 of the role of metadata, so crucial for the interaction between the humanities and NLP, a rough overview of relevant initiatives will be given. Inspired by some telling examples, it will be outlined what could be done to increase the chances for a bright future for the old ties, and how other domains can benefit as well from the reinvention of the old common playground between NLP and the humanities.

2 Metadata in the Humanities

Digital text, but also multimedia content, can be mined for the occurrence of patterns at all kinds of layers, and based on techniques for information extraction and classification, documents can be annotated automatically with a variety of labels, including indications of topic, event types, authorship, stylistics, etc. Automatically generated annotations can be exploited to support to what is often called the semantic access to content, which is typically seen as more powerful than plain full text search, but in principle also includes conceptual search and navigation.

The data used in research in the domain of the humanities comes from a variety of sources: archives, musea (or in general cultural heritage collections), libraries, etc. As a testbed for NLP these collections are particularly challenging because of the combination of complexity increasing features, such as language and spelling change over time, diversity in orthography, noisy content (due to errors introduced during data conversion, e.g., OCR or transcription of spoken word material), wider than average stylistic variation and cross-lingual and cross-media links. They are also particularly attractive because of the available metadata or annotation records, which are the reflection of analytical and comparative scholarly processes. In addition, there is a wide diversity of annotation types to be found in the domain (cf. the annotation dimensions distinguished by (Marshall, 1998)), and the field has developed modelling procedures to exploit this diversity (McCarty, 2005) and visualisation tools (Unsworth, 2005).

2.1 Metadata for Text

For many types of textual data automatically generated annotations are the sole basis for semantic search, navigation and mining. For humanities and cultural heritage collections, automatically generated annotation is often an addition to the catalogue information traditionally produced by experts in the field. The latter kind of manually produced metadataa is often specified in accordance to controlled key word lists and metadata schemata agreed for the domain. NLP tagging is then an add on to a semantic layer that in itself can already be very rich and of high quality. More recently initiatives and support tools for so-called social tagging have been proposed that can in principle circumvent the costly annotation by experts, and that could be either based on free text annotation or on the application of so-called folksonomies as a replacement for the traditional taxonomies. Digital librarians have initiated the development of platforms aiming at the integration of the various annotation processes and at sharing tools that can help to realise an infrastructure for distributed annotation. But whatever the genesis is of annotations capturing the semantics of an entire document, they are a very valuable source for the training of automatic classifiers. And traditionally, textual resources in the humanities have lots of it, partly because the mere art of annotating texts has been invented in this domain.

2.2 Metadata for Multimedia

Part of the resources used as basis for scholarly research is non-textual. Apart from numeric data resources, which are typically strongly structured in database-like environments, there is a growing amount of audiovisual material that is of interest to humanities researchers. Various kinds of multimedia collections can be a primary source of information for humanities researchers, in particular if there is a substantial amount of spoken word content, e.g., broadcast news archives, and even more prominently: oral history collections.

It is commonly agreed that accessibility of heterogeneous audiovisual archives can be boosted by indexing not just via the classical metadata, but by enhancing indexing mechanisms through the exploitation of the spoken audio. For several types of audiovisual data, transcription of the speech segments can be a good basis for a time-coded index. Research has shown that the quality of the automatically generated speech transcriptions, and as a consequence also the index quality, can increase if the language models applied have been optimised to both the available metadata (in particular on the named entities in the annotations) *and* the collateral sources available (Huijbregts et al., 2007). ‘Collateral data is the term used for secondary information objects that relate to the primary documents, e.g., reviews, program guide summaries, biographies, all kinds of textual publications, etc. This requires that primary sources have been annotated with links to these secondary materials. These links can be pointers to source locations *within the collection*, but also links to related documents from *external sources*. In laboratory settings the amount of collateral data is typically scarce, but in real life spoken word archives, experts are available to identify and collect related (textual) content that can help to turn generic language models into domain specific models with higher accuracy.

2.3 Metadata for Surprise Data

The quality of automatically generated content annotations in real life settings is lagging behind in comparison to experimental settings. This is of course an obstacle for the uptake of technology, but a number of pilot projects with collections from the humanities domain show us what can be done to overcome the obstacles. This can be illustrated again with the situation in the field of spoken document retrieval.

For many A/V collections with a spoken audio track, metadata is not or only sparsely available, which is why this type of collection is often only searchable by linear exploration. Although there is common agreement that speech-based, automatically generated annotation of audiovisual archives may boost the semantic access to fragments of spoken word archives enormously (Goldman et al., 2005; Garofolo et al., 2000; Smeaton et al., 2006), success stories for real life archives are scarce. (Exceptions can be found in research projects in the broadcast news and cultural heritage domains, such as MALACH (Byrne et al., 2004), and systems such as SpeechFind (Hansen et al., 2005).) In lab conditions the focus is usually on data that (i) have well-known characteristics (e.g, news content), often learned along with annual benchmark evaluations,[1] (ii) form a relatively homogeneous collection, (iii) are based on tasks that hardly match the needs of real users, and (iv) are annotated in large quantities for training purposes. In real life however, the exact characteristics of archival data are often unknown, and are far more heterogeneous in nature than those found in laboratory settings. Language models for realistic audio sets, sometimes referred to as *surprise data* (Huijbregts, 2008), can benefit from a clever use of this contextual information.

Surprise data sets are increasingly being taken into account in research agendas in the field focusing on multimedia indexing and search (de Jong et al., 2008). In addition to the fact that they are less homogenous, and may come with links to related documents, real user needs may be available from query logs, and as a consequence they are an interesting challenge for cross-media indexing strategies targeting aggregated collections. Sur-

[1] E.g., evaluation activities such as those organised by NIST, the National Institute of Standards, e.g., TREC for search tasks involving text, TRECVID for video search, Rich Transcription for the analysis of speech data, etc. `http://www.nist.gov/`

prise data are therefore an ideal source for the development of best practises for the application of tools for exploiting collateral content and metadata. The exploitation of available contextual information for surprise content and the organisation of this dual annotation process can be improved, but in principle joining forces between NLP technologies and the capacity of human annotators is attractive. On the one hand for the improved access to the content, on the other hand for an innovation of the NLP research agenda.

3 Ingredients for a Novel Knowledge-driven Workflow

A crucial condition for the revival of the common playground for NLP and the humanities is the availability of representatives of communities that could use the outcome, either in the development of services to their users or as end users. These representatives may be as diverse and include e.g., archivists, scholars with a research interest in a collection, collection keepers in libraries and musea, developers of educational materials, but in spite of the divergence that can be attributed to such groups, they have a few important characteristics in common: they have a deep understanding of the structure, semantic layers and content of collections, and in developing new road maps and novel ways of working, the pressure they encounter to be cost-effective is modest. They are the first to understand that the technical solutions and business models of the popular web search engines are not directly applicable to their domain in which the workflow is typically knowledge-driven and labour-intensive. Though with the introduction of new technologies the traditional role of documentalists as the primary source of high quality annotations may change, the availability of their expertise is likely to remain one of the major success factors in the realisation of a digital infrastructure that is as rich source as the repositories from the analogue era used to be.

All kinds of coordination bodies and action plans exist to further the field of Digital Humanities, among which The Alliance of Digital Humanities Organizations `http://www.digitalhumanities.org/` and HASTAC (`https://www.hastac.org/`) and Digital Arts an Humanities `www.arts-humanities.net`, and dedicated journals and events have emerged, such as the LaTeCH workshop series. In part they can build on results of initiatives for collaboration and harmonisation that were started earlier, e.g., as Digital Libraries support actions or as coordinated actions for the international community of cultural heritage institutions. But in order to reinforce the liaison between NLP and the humanities continued attention, support and funding is needed for the following:

Coordination of coherent platforms (both local and international) for the interaction between the communities involved that stimulate the exchange of expertise, tools, experience and guidelines. Good examples hereof exist already in several domains, e.g., the field of broadcast archiving (IST project PrestoSpace; `www.prestospace.org/`), the research area of Oral History, all kinds of communities and platforms targeting the accessibility of cultural heritage collections (e.g., CATCH; `http://www.nwo.nl/catch`), but the long-term sustainability of accessible interoperable institutional networks remains a concern.

Infrastructural facilities for the support of researchers and developers of NLP tools; such facilities should support them in finetuning the instruments they develop to the needs of scholarly research. CLARIN (`http://www.clarin.eu/`) is a promising initiative in the EU context that is aiming to cover exactly this (and more) for the social sciences and the humanities.

Open access, source and standards to increase the chances for inter-institutional collaboration and exchange of content and tools in accordance with the policies of the *de facto* leading bodies, such as TEI (`http://www.tei-c.org/`) and OAI (`http://www.openarchives.org/`).

Metadata schemata that can accommodate NLP-specific features:

- automatically generated labels and summaries
- reliability scores
- indications of the suitability of items for training purposes

Exchange mechanisms for best practices e.g., of building and updating training data, the use of annotation tools and the analysis of query logs.

Protocols and tools for the mark-up of content, the specification of links between collections, the handling of IPR and privacy issues, etc.

Service centers that can offer heavy processing facilities (e.g. named entity extraction or speech transcription) for collections kept in technically modestly equipped environments hereof.

User Interfaces that can flexibly meet the needs of scholarly users for expressing their information needs, and for visualising relationships between interactive information elements (e.g., timelines and maps).

Pilot projects in which researchers from various backgrounds collaborate in analysing a specific digital resource as a central object in order to learn to understand how the interfaces between their fields can be opened up. An interesting example is the the project Veteran Tapes (`http://www.surffoundation.nl/smartsite.dws?id=14040`). This initiative is linked to the interview collection which is emerging as a result for the Dutch Veterans Interview-project, which aims at collecting 1000 interviews with a representative group of veterans of all conflicts and peace-missions in which The Netherlands were involved. The research results will be integrated in a web-based fashion to form what is called an *enriched publication*.

Evaluation frameworks that will trigger contributions to the enhancement en tuning of what NLP has to offer to the needs of the humanities. These frameworks should include benchmarks addressing tasks and user needs that are more realistic than most of the existing performance evaluation frameworks. This will require close collaboration between NLP developers and scholars.

4 Conclusion

The assumption behind presenting these issues as priorities is that NLP-empowered use of digital content by humanities scholars will be beneficial to both communities. NLP can use the testbed of the Digital Humanities for the further shaping of that part of the research agenda that covers the role of NLP in information handling, and in particular those avenues that fall under the concept of mining. By focussing on the integration of metadata in the models underlying the mining tools and searching for ways to increase the involvement of metadata generators, both experts and ‘amateurs’, important insights are likely to emerge that could help to shape agendas for the role of NLP in other disciplines. Examples are the role of NLP in the study of recorded meeting content, in the field of social studies, or the organisation and support of tagging communities in the biomedical domain, both areas where manual annotation by experts used to be common practise, and both areas where mining could be done with aggregated collections.

Equally important are the benefits for the humanities. The added value of metadata-based mining technology for enhanced indexing is not so much in the cost-reduction as in the wider usability of the materials, and in the impulse this may bring for sharing collections that otherwise would too easily be considered as of no general importance. Furthermore the evolution of digital texts from ‘book surrogates’ towards the rich semantic layers and networks generated by text and/or media mining tools that take all available metadata into account should help the fields involved in not just answering their research questions more efficiently, but also in opening up grey literature for research purposes and in scheduling entirely new questions for which the availability of such networks are a *conditio sine qua non*.

Acknowledgments

Part of what is presented in this paper has been inspired by collaborative work with colleagues. In particular I would like to thank Willemijn Heeren, Roeland Ordelman and Stef Scagliola for their role in the genesis of ideas and insights.

References

W. Byrne, D.Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W-J. Zhu. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4).

F. M. G. de Jong, D. W. Oard, W. F. L. Heeren, and R. J. F. Ordelman. 2008. Access to recorded interviews: A research agenda. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 1(1):3:1–3:27, June.

J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. 2000. The TREC SDR Track: A Success Story. In *8th Text Retrieval Conference*, pages 107–129, Washington.

J. Goldman, S. Renals, S. Bird, F. M. G. de Jong, M. Federico, C. Fleischhauer, M. Kornbluh, L. Lamel, D. W. Oard, C. Stewart, and R. Wright. 2005. Accessing the spoken word. *International Journal on Digital Libraries*, 5(4):287–298.

J.H.L. Hansen, R. Huang, B. Zhou, M. Deadle, J.R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul. 2005. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Transactions on Speech and Audio Processing*, 13(5):712–730.

M.A.H. Huijbregts, R.J.F. Ordelman, and F.M.G. de Jong. 2007. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of SAMT 2007*, volume 4816 of *Lecture Notes in Computer Science*, pages 78–90, Berlin. Springer Verlag.

M.A.H. Huijbregts. 2008. *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled.* Phd thesis, University of Twente.

C. Marshall. 1998. Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems (HYPERTEXT '98)*, pages 40–49, Pittsburgh, Pennsylvania.

W. McCarty. 2005. *Humanities Computing*. Basingstoke, Palgrave Macmillan.

S. Schreibman, R. Siemens, and J. Unsworth (eds.). 2004. *A Companion to Digital Humanities*. Blackwell.

A.F. Smeaton, P. Over, and W. Kraaij. 2006. Evaluation campaigns and trecvid. In *8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR2006)*.

J. Unsworth. 2005. *New Methods for Humanities Research. The 2005 Lyman Award Lecture.* National Humanities Center, NC.