D-EE2.8 Development and Use of Computational Lexica for OCR And IR on Historical Documents – A Cross-Language Perspective | EE2

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Jesse de Does	21 January 2012	Created
1.0	Final	“	1 February 2012	Comments integrated from internal reviewer and language partners

Approvals

Version	Date of approval	Name	Role in project	Signature
0.1	26 January 2012	EE3 language partners	WP members	OK
0.1	27 January 2012	Rafael Carrasco	Internal reviewer	OK
1.0	23 March 2012	Max Kaiser	SP EE leader	OK
1.0	23 March 2012	Hildelies Balk	Project Director	OK

Distribution

Version	Date of sending	Name	Role in project
0.1	21 January 2012	EE3 partners, Rafael Carrasco	WP members and internal reviewer
1.0	8 February 2012	Max Kaiser, Hildelies Balk	SP EE leader, Project Director
1.0	6 April 2012	Liina Munari	EC Project Officer

Abstract

In the wake of current mass digitization projects in many libraries around the world, huge amounts of historical books and documents are brought to the web. In the scientific community, the difficult metamorphosis from historical paper documents to searchable electronic documents has drawn considerable attention. Though many important projects such as Google Books have already been on their way for some time, from a scientific and technical point of view the path from historical books in paper format to searchable documents in digital libraries is characterized by two core problems that still have not been solved in a fully satisfactory way.

A first difficult step is the conversion from paper to electronic form using optical character recognition (OCR). In spite of recent and ongoing improvements, the quality of OCR’ed historical texts is often still low. This is due to several reasons. Historical fonts often differ per book and are difficult to read. The quality of the paper and the images of historical documents is often suboptimal due to distinct forms of noise and geometric distortion. Furthermore, linguistic components and resources of current OCR systems are often not ‘aware’ of the kind of language variants found in historical texts. Each input text typically comes with its own specific mixture of features and problems, which explains why the quality of OCR results for historical documents may range from excellent to hardly acceptable.

Even if historical texts are recognized in a perfect way, a second general problem is caused by historical spelling and language variation found in the documents. Most users of digital libraries are not familiar with historical language and

want to use modern spelling to search in historical documents. Due to historical language changes and the lack of standardization of orthography in earlier centuries, any modern word may occur in many different spellings in historical documents. When using the current standard techniques in Information Retrieval (IR), these hits are missed, resulting in low recall. To obtain satisfactory answers to search queries, the gap between modern and old spelling needs to be bridged using appropriate methods.

From a cross-language perspective, an additional major challenge in resolving this issue is the fact that for each language the point of departure is different. The level of experience with digitization of historical text material and the amount of reliable text in original historical spelling available is different for different languages (countries). There are also significant differences as to availability of suitable OCR technology (e.g. some Cyrillic characters used in 19th century Bulgarian are not supported by standard OCR engines). For some languages (English, French, German, Spanish) preliminary lexical support for historical language is available in commercial OCR engines, for others this is completely lacking. There are similar differences in the availability of lexical resources for supporting information retrieval on historical documents. Even finer levels of granularity have to be taken into account, because of distinct spellings and alphabet conventions in different periods for different languages.

Language work in IMPACT addresses 9 European languages: Bulgarian, Czech, Dutch, English, French, German, Polish, Slovene, Spanish. We have endeavoured to deal with the aforementioned problems by developing different strategies for lexicon development, taking distinct points of departure into account. These strategies have been implemented in a set of tools to support the construction of historical lexica starting from digital corpora, lexica and electronic historical dictionaries. Using these tools, both OCR and IR lexica have been developed for all 9 languages.

In addition to this, a set of tools to exploit the lexical data in both OCR and IR has been developed. To implement lexicon-supported OCR, we have developed a module enabling the use of these lexica with the well-known ABBYY FineReader OCR engine. A retrieval layer on top of the widely-used Lucene search engine has been developed to exploit the historical lexica of all IMPACT languages in IR.

A problem often underestimated is the question of how to measure progress. In computational linguistics, the importance of good evaluation methods and gold standard data is widely understood, and data sets are available. In digitization, the situation is different. One of the important contributions of IMPACT is the development of an extensive ground truth set in a unified format for 9 languages, both for evaluation of OCR and IR technologies. Apart from the data, a major achievement has been the development of a unified framework for evaluation for 9 languages.

Our final evaluation shows that we have reached significant improvement for all languages. To briefly summarize the outcomes of the evaluation, in OCR we see an improvement of the word recognition rate ranging from 10 to 30%. For IR, we see significant improvement in the recall of historical word forms using modern lemmata as a search key.
The practical progress of the language work in IMPACT is already reflected in take-up of project results in various settings. A commercial provider purchased the Dutch historical OCR lexicon; the Slovene library partner currently deploys the Slovene IR lexicon in its search engine.

Tools and datasets will be made available at www.digitisation.eu, the website of the IMPACT Centre of Competence. A detailed publication of the above mentioned work is envisaged for the first half of 2013.