20th International Council on Archives Congress

"Development of OCR software for minority languages as the way to ensure equal status of native and non-native archives"

ABBYY Speaker:

Maxim Mikhalev, PhD

Business Development Manager

ABBYY 3A

5-10 September 2016

Seoul, Republic of Korea

Development of OCR software for minority languages as the way to ensure equal status of native and non-native archives.

Digital divide

In our days, the need for digitization of the humankind's cultural heritage is widely acknowledged and seldom challenged. It goes without saying that digitization is required for the purposes of preservation and dissemination of culture. To put it another way, it is indispensable for ensuring its availability for the present and future generations. At the same time, it is clear that the underrepresented languages constitute an important part of the world cultural heritage and their existence and development is necessary for safeguarding of the global cultural diversity. They also embody precious original works of science and literature, and are often of significant historical value by themselves.

Yet, the two statements above often appear as oxymorons. When we speak of digitization, we often think only about the cultural assets in major languages. This is actually misleading and even dangerous, since selective digitization makes overall situation with preservation of cultural heritage worse. As a matter of fact, majority of the underrepresented languages across the globe are already suffering from rapidly shrinking number of their active speakers, which subsequently, gradually shrink the sphere of their usage. Under these circumstances, digitization of archives and libraries in major languages only widens and deepens digital divide between those languages and the languages spoken by minorities. As a result, minority languages can become obsolete even faster than it is generally feared. As societies become increasingly dependent on the digital information and digital communication, the inability of certain languages to "perform digitally" turns them into useless and inessential tools.

If this trend continues, one can safely assume that underrepresented languages will soon be unable to withstand the onslaught of the digital revolution, while the priceless heritage that they embody can vanish in the foreseeable future. The best way to prevent this negative scenario from becoming reality and to reduce the "digital divide" between languages is to ensure that the invaluable contents in underrepresented languages are digitized and converted into editable and searchable archives accessible to a broader readership.

The software problem

One can argue that, considering current political paradigm, both the major and minority languages enjoy equal status. Furthermore, linguistic policies are more often than not designed to favor minority languages. Indeed, digitization projects associated with underrepresented languages can sometimes get better financing from the international NGOs and various aid institutions, who tend to favor minorities over majorities. Despite such positive trends, issues still exist. Full-text digital preservation is not only routine or mechanical digitization of archive collections – which may or may not be supported by governments or NGOs – but it also requires an OCR software designed for each specific underrepresented language. Without quality text recognition that such software provides, digitization brings regretfully poor results. Regrettably, the result of such daunting task cannot be used for the creation of fully editable and fully searchable archives suitable for further work and research.

Obstacles

High-quality software suitable for this task is mostly being developed by private enterprises that are driven by market forces and profit motive rather than political or cultural considerations. No wonder that software development companies are seldom willing to allocate their resources into time-consuming development projects such as OCR for the underrepresented languages. However, profitability it is not the only problem. Another challenge is the degree of complexity and obscurity that characterizes many of the world's underrepresented languages. For the private companies with limited connections in the academic circles, finding capable specialists for quality implementation of such project is a daunting prospect. All these combined to make the task of comprehending any rare language a serious undertaking that requires serious investments. As if this was not enough, anyone willing to develop an OCR software for an underrepresented language is confronted with new challenges: that the number of available texts in those languages is also limited; they are hard to find and even harder to get an access to. The availability of the reasonable volume of texts is a key to improving general quality of recognition, as only massive reference data can guarantee better software performance.

Taking into consideration all the above obstacles, one shall not be surprised to discover that so far almost no quality OCR solutions have been created for many of the world's underrepresented languages. Because of this, the threat of their possible extinction in the digital age is getting more and more acute.

ABBYY

Thankfully, it's not all gloom and doom. ABBYY, one of the world leaders in the development of linguistic software and technologies, is stepping in with the good news. This privately owned company has been engaged in developing and marketing of its linguistic-related products since 1989. ABBYY's current product line-up consists of a several dozens of linguistic-related software. Among the best known are "Lingvo" electronic dictionaries and the comprehensive solutions based on semantic analysis of metadata such as "InfoExtractor" and "Smart Classifier", that were developed with the help of its pioneering "Compreno" technology. The company's flagship products – including highly acclaimed "FineReader", "Recognition Server" and "FlexiCapture" – have always been OCR solutions for full-text digitization and data capture. ABBYY is well positioned to license its technologies to software developers and the company has been offering a number of SDKs that allows its partners to create custom solutions aligned to the needs of their clients. One should note that ABBYY counts among its partners several libraries, national archives and scientific content holders around the globe – the National Library of Latvia, CNKI of China and National Assembly Library in Korea among others.

Case study: National Library of Latvia

Some of the digitization projects with the participation of ABBYY involved underrepresented languages. A good example worth mentioning is co-operation between ABBYY and National Library of Latvia. Although Latvian language is the official language of the country and by no means a minority one, but its use of old gothic script classifies it as an unrepresented language because of the presence in it of several unique letters, which are not present in other European

gothic scripts.

Due to such difference, ABBYY developed a special edition of its OCR for the Latvian gothic to achieve better recognition results. Thanks to the generous financial support from the EU and general enthusiasm of the people involved in the project both in ABBYY and at the National Library of Latvia, this task was successfully completed. As a result, it was technically possible to preserve four million pages of precious documents from the library collection and make them available at www.periodika.lv. The project revealed the complexity of the process and highlighted all those obstacles that have been mentioned above. At the same time, the undisputable benefit that such projects can bring to society inspired the management of the company to take development of such cultural heritage related projects into serious consideration.

Case study: Burmese OCR

Another successful example of supporting underrepresented languages was recent development of Burmese OCR completed with the support from one of the governmental agencies in Singapore. The product itself did not have any obvious commercial potential; nevertheless, its application allowed enormous heritage-related data kept in old and modern Burmese to be preserved and to make it accessible to the scholars worldwide. Despite being eventually successful project, it, however, also highlighted some of the above-mentioned problems. First of all, it was only by sheer luck that ABBYY's chief programmer got so "addicted" to this relatively rare language, that he decided to learn it in his spare time and was eventually able to comprehend its structure. Otherwise, the project would have been impossible to complete, as there are very few specialists in Burmese living in Russia. Second, when the project was finally completed, it became close to impossible to put it to use for the purpose of culture preservation, as, ironically, cultural institutions in Myanmar expressed very little interest in using Burmese OCR, even when it was offered to them for free. The reason was simple: due to financial and organizational constraints, they themselves were not ready for any massive digitization project.

Culture heritage initiative

Despite all these problems, ABBYY is now planning to launch a global initiative in the area of cultural heritage preservation.

As one of the few companies in the world capable of creating professional software that can be used for full-text digitization, ABBYY is pioneering the development of a series of OCR products designed especially for the underrepresented languages around the world. This will include languages of those of the minorities living in the larger states (for example, Tibetan in China) and those of the smaller nations that cannot afford commercial OCR software (for example, Afghanistan). This initiative is undertaken in the hope that it will help preserve and protect a large number of valuable texts composed in those often obscure languages and make them available for people across the globe.

Call for co-operation

The main purpose of this paper is to draw attention of key players and decision-makers in the sphere of the preservation of cultural heritage to the problem of digitally underrepresented languages and to the main obstacles to the bridging digital divide between minor languages and the major ones. It is important to understand that this divide is gradually deepening as more and more major languages are digitized or can acquire exclusively-developed quality OCR software. For its part, ABBYY has the ability and willingness to invest its experience and expertise into development of OCR software required for the survival of underrepresented languages in the digital age. In order to fulfil this mission, however, software developers need to co-operate closely not only with the holders of content, that is, with archives and museums, and with scientific community, but also with the governmental and the non-governmental organizations in the sphere of heritage preservation including UNESCO. Content holders can provide the necessary linguistic and textual expertise that is difficult to find otherwise, while governmental and non-governmental organizations can use their authority and resources to ensure that necessary support for these socially and culturally important global projects is provided.