Building Arabic Corpora from Wikisource

Imene Bensalem, Salim Chikhi
SCAL team, MISC laboratory
Constantine 2 University
Constantine, Algeria
bens.imene@gmail.com, chikhi@misc-umc.org

Paolo Rosso
Natural Language Engineering Lab. – EliRF
Universitat Politècnica de València
Valencia, Spain
prosso@dsic.upv.es

***Abstract*—This paper describes a new tool that helps extracting clean text from the Arabic Wikisource dump in order to build corpora. The tool purpose is illustrated by the generation of a subcorpus from Wikisource, which is a step towards the building of an evaluation corpus for Arabic intrinsic plagiarism detection.**

Keywords—Arabic Wikisource; tools for building corpora; intrinsic plagiarism detection

I. INTRODUCTION

Building corpora is a time consuming task especially if the source of text is noisy, or if the intended text has specific criteria (e.g. specific genres or topics). Generally, researchers develop tailored or ad hoc codes to compile or preprocess the target text ; by consequence no one else can benefit from their codes since they are developed without the intention to be shared. In fact, tools of building Arabic corpora, particularly, are very few. These are two examples that we found in literature: Khoja [1] developed a tool that helps creating corpora from RSS feeds of blogs ; Meftouh et al. [2] proposed a software that crawls Arabic web pages according to a list of keywords.

II. WHAT IS WIKISOURCE ?

Wikisource [1] is a free web library that contains public domain textual documents, such as, heritage books. Researchers do not have to worry about the copyright issue when using Wikisource as a source of text since it provides only texts without copyright.

Wikisource content cannot be obtained directly as a raw text (as its similar project Gutenberg[2] e.g.). It can be instead downloaded as an XML dump [3] which is noisy with Wiki markups; hence the need of a cleaning step that, on the one hand, discards the markups and on the other hand, extracts relevant information from them, such as, the author of the text and its category. In fact, there exist many tools of extracting text from Wikimedia dumps. However, they are more tailored to Wikipedia (i.e., do not process Wikisource specific markups) and do not consider languages specifications. At first glance, this task seems easy to implement , but actually it is tedious because of the untidy nature of the Wiki markups. We believe that the tool described here will save time and effort of researchers during the process of Arabic corpora building.

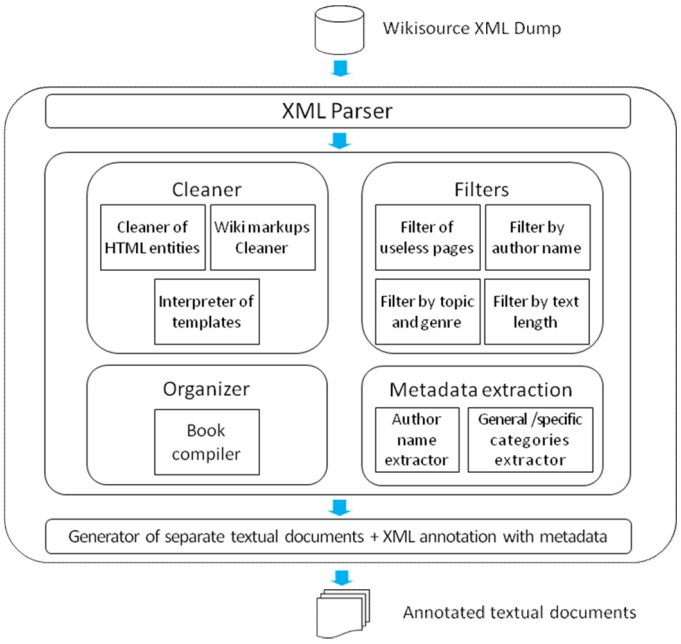

Fig. 1. The tool achitecture

III. DESCRIBING THE TOOL

The tool (See Fig. 1) is based on a Perl script that allows applying to Wikisource dump the following main operations:

1) Cleaning: This task entailed the use of more than 80 regular expressions to delete the noise and substitute the templates using for abbreviation with the appropriate phrases.

2) Filtering: It allows discarding documents or creating subcorpora according to many criteria, namely categories, author names, document lengths and some templates indicating non-content pages (e.g. disambiguation pages and pages that contain only a link to a scanned book)

3) Extracting Metadata: It consists in parsing the Wiki markups to extract the text author name and categories.

4) Organizing: It helps gathering pages of the same book in one document based on similarities in their titles and some Wiki markups.

[1] http://wikisource.org
[2] http://www.gutenberg.org
[3] http://dumps.wikimedia.org

This work is the result of the collaboration between the two universities in the framework of the bilateral research project on Application of Natural Language Processing to the Need of the University (AECID-PCI AP/043848/11).

The corpus generation is the last task: It converts the XML dump to a set of textual files. Each textual file is associated with an XML file that contains metadata, namely the title and the author name. The scheme used for annotation is the same used in PAN[4] corpora for plagiarism detection [3].

IV. Towards a Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Intrinsic plagiarism detection consists in uncovering plagiarism on the basis of changes in the writing style among fragments of the same text. Arabic intrinsic plagiarism detection is still an unexplored area due to the lack of an evaluation corpus [4]. In such a corpus, each document must be written by only one author but contains fragments borrowed from other documents written by other authors.

In this section we describe a collection of documents that has been built from Arabic Wikisource in order to be used as target documents for plagiarism (i.e., documents where to insert plagiarism). This represents the first step in the process of creating a complete corpus of the evaluation of intrinsic plagiarism detection.

We choose to extract text from Wikisource for two main reasons: (i) most of Wikisource texts are books. These are generally written by one author and do not contain many reused fragments, which is not the case of Wikipedia and newswire texts for example. (ii) Wikisource is the only resource that clearly provides Arabic content (with the desired criteria mentioned in the former point) without copyrights which is an important issue since we are planning to make the corpus publicly available.

In addition to the step of cleaning, the process that we explain below was applied to Wikisource dump using the developed tool in order to obtain a corpus with the desired features.

1) Deletion of religion books, poetry pages, legal and language's texts: we believe that these types of text are not relevant to build a corpus for intrinsic plagiarism detection for the following reasons:

- Arabic religion books contain usually many citations from Quran and Hadith which may alter the study of the author's writing style.
- Poetry is another type of writing. We think that it does not make sense to use poems as target documents for plagiarism i.e., inserting plagiarism from another text type (e.g. novels) into poems. In fact, plagiarism in poetry should be evaluated using a separate corpus where the target documents and the inserted fragments of plagiarism are both poems.
- Legal texts may be written by many authors.
- Language books, especially dictionaries, have a peculiar structure. We choose to not consider them for the same reason of poetry.

2) Deletion of pages without author and topic: authors names and topics are crucial information for the step of insertion of plagiarism. The text author must be known to avoid inserting fragments of plagiarism whose author is the same of the target document's author. The topic tag is important to be able to simulate the real plagiarism by inserting fragments from source documents that have the same topic of the target document.

3) Deletion of pages very short (< 1 page): Indeed experiences in English text have shown that the writing style analysis become unreliable with short texts [5].

4) Compilation of separate pages of some books in one document: The aim of this step is to create long documents in order to have a corpus with different documents lengths.

Statistics on the obtained corpus are provided in Table I.

TABLE I. Selected Statistics of the Generated Corpus

Number of documents	1008
Number of authors	114
Number of topics	15

V. Conclusion and Future work

The main goal of this paper was to shed light on a free of copyright source of Arabic text for corpora building, namely Arabic Wikisource. One of the main disadvantages of Wikisource is the noise in texts, hence the need of a processing in order to extract raw text from it. Our main contribution is the development of an automatic tool for generating clean plain text corpora (or subcorpora) from Wikisource XML dump.

This work is part of the main project of building a corpus for the evaluation of Arabic intrinsic plagiarism detection. The creation of such a corpus needs mainly two steps which are: (i) building the collection of document where to insert plagiarism and (ii) inserting plagiarism. In this paper we illustrate the first step using Arabic Wikisource as a source of text.

References

[1] S. Khoja, "An RSS Feed Analysis Application and Corpus Builder," Proceedings of the Second International Conference on Arabic Language Resources and Tools, pp. 115–118, 2009.

[2] K. Meftouh, K. Smaili, and M. T. Laskri, "Constitution d'un corpus de la langue Arabe à partir du Web," Colloque International sur le Traitement Automatique de la Langue Arabe (CITALA'07), pp. 1–7, 2007.

[3] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An Evaluation Framework for Plagiarism Detection," in Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10), 2010, pp. 997–1005.

[4] I. Bensalem, P. Rosso, and S. Chikhi, "Intrinsic Plagiarism Detection in Arabic Text: Preliminary Experiments," in II Spanish Conference on Information Retrieval (CERI'12), 2012.

[5] B. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," Language Resources and Evaluation, vol. 45, no. 1, pp. 63–82, Jan. 2010.

[4] http://pan.webis.de