Assessing the detection of text-reuse

Boot, P.

2015

Assessing the detection of text-reuse

Peter Boot
Huygens ING
peter.boot@huygens.knaw.nl

One of the central questions in the history of literature and the history of ideas is always: to which other texts is the text that I am studying indebted? And which other texts has it influenced? The detection of quotation is a problem that has been addressed in a number of digital humanities projects, among others projects focusing on Latin poetry (Coffee et al., 2012), medieval encyclopedias (Mews et al., 2010), 18th-century French philosophy (Olsen et al., 2011), 19th-century French literature (Ganascia et al., 2014), and early US newspapers (Smith et al., 2013). A variety of tools has been developed, applying a variety of algorithms. Some of these tools started their life as tools for the detection of plagiarism (Kane and Tompa, 2011).

Most tools are based on the detection of matching word n-grams, and can be fine-tuned using a number of parameters, for instance allowing non-matching words between the matching words, by applying lemmatization, or by removing stop-words. Different parameter settings may be required for different languages (based on e.g. the amount of flexion in the language) and text collections (dependent on e.g. consistency of spelling, text quality (OCR vs. transcribed text) and genre).

The problem that this paper and demo will address is how to assess the suitability of a certain tool and parameter setting for specific textual corpora. Running a tool that looks for textual correspondences between two textual corpora results in a list of corresponding text places. The hits are true positives, false positives, or something in between. True correspondences that a tool does not detect remain invisible. Re-running the tool with different parameter settings results in an overlapping set of hits, and as there is a large number of different parameter settings, the researcher easily gets lost. Since the true amount of quotation is unknown and subject to interpretation, standard information retrieval measures to detect the best settings do not apply.

I my paper, I argue that what we need is (i) a database that stores all relevant information for each run of the detection tool: the tool version, the corpora properties, the applied settings and the results, (ii) an application that displays parameters and results and allows flexible navigation through the results, ad (iii) the facility to annotate the database at all levels: the program run, the processed texts, and the detected correspondences.

The demo will show a prototype for such a tool. I developed the prototype to help me investigate intertextuality in the 17th-century emblem, a genre in which intertextuality played an important role. I used Text::Pair (https://code.google.com/p/text-pair/) to find textual relations between the emblem corpus digitized in the Emblem Project Utrecht (http://emblems.let.uu.nl/) and a number of other text

corpora. The demo will show how the prototype helps assess the suitability of different parameter settings when using Text::Pair on these corpora.

References

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R. & Jacobson, S. L. 2012. The Tesserae Project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing,* 28**,** 221-228.

Ganascia, J.-G., Glaudes, P. & Del Lungo, A. 2014. Automatic Detection of Reuses and Citations in Literary Texts. *Literary and Linguistic Computing,* 29**,** 412-421.

Kane, A. & Tompa, F. W. 2011. Janus: the intertextuality search engine for the electronic Manipulus florum project. *Literary and Linguistic Computing,* 26**,** 407-415.

Mews, C. J., Zahora, T., Nikulin, D. & Squire, D. 2010. The Speculum morale (c. 1300) and the study of textual transformations: a research project in progress. *Vincent of Beauvais Newsletter,* 35**,** 5-15.

Olsen, M., Horton, R. & Roe, G. 2011. Something borrowed: sequence alignment and the identification of similar passages in large text collections. *Digital Studies/Le champ numérique,* 2.

Smith, D. A., Cordell, R. & Dillon, E. M. Infectious texts: Modeling text reuse in nineteenth-century newspapers. Big Data, 2013 IEEE International Conference on, 2013. IEEE, 86-94.