Bringing Modern Spell Checking Approaches to Ancient Texts – Automated Suggestions for Incomplete Words

dh2012.uni-hamburg.de/conference/programme/abstracts/bringing-modern-spell-checking-approaches-to-ancient-texts-autom

XML

Büchler, Marco, Leipzig University, Germany, mbuechler@e-humanities.net
Kruse, Sebastian, Leipzig University, Germany, skruse@eaqua.net
Eckart, Thomas, Leipzig University, Germany, teckart@e-humanities.net

One of the most challenging tasks for scholars working with ancient data is the completion of texts that have only been partially preserved. In the current situation, a great deal of scholarly experience and the use of dictionaries such as *Liddell Scott Jones* or *Lewis & Short* are necessary to perform the task of text reconstruction manually. Even though text search tools such as Diogenes or papyri.info exist, scholars still have to work through the results manually and require a very good knowledge about the text, its cultural background and its documentary form in order to be able to decide about the correct reconstitution of the damaged text. Therefore, a 'selective and relatively small scope' especially of younger scholars restricts the set of potential candidates.

To overcome these barriers an unsupervised approach from the field of machine learning is introduced to form a word prediction system based on several classes of spell checking (Kukich 1992; Schierle et al. 2008) and text mining algorithms.

Both spell checking and text completion can be separated into two main tasks: identification of incorrect or incomplete words and the generation of suggestions. While the identification of misspelled words can be a very difficult task when working with modern texts (such as with spell checking support provided by modern word processing suites), existing sigla of the Leiden Conventions (Bodard et al. 2009) can be used when dealing with ancient texts. The second step of the process is then to generate likely suggestions using methods such as:

- **Semantic approaches**: *Sentence co-occurrences* (Buechler 2008) and *document co-occurrences* (Heyer et al. 2008) are used to identify candidates based on different contextual windows (Bordag 2008). The basic idea behind this type of classification is motivated by Firth's famous statement about a word's meaning: '*You shall know a word by the company it keeps'* (Firth 1957).
- **Syntactical approaches**: *Word bi- and trigrams* (Heyer et al. 2008): With this method, the immediate neighbourhood of a word is observed and likely candidates are identified based on a selected reference corpus.
- **Morphological dependencies**: Similar to the *Latin and Greek Treebank of Perseus* (Crane et al. 2009) morphological dependencies are used to suggest words by using an expected morphological code.
- **String based approaches**: The most common class of algorithms for modern texts compares words by their word similarity on letter level. Different approaches like the *Levenshtein distance* (Ottmann & Widmayer 1996) or faster approaches such as *FastSS* (Bocek et al. 2007) are used to compare a fragmentary word with all candidates.
- **Named Entity lists**: With a focus on deletions of inscriptions, existing and extended named entity lists for person names, cities or demonyms like the *Lexicon of Greek Personal Names* (Fraser et al. 1987-2008) or the *Wörterlisten* of Dieter Hagedorn are used to look for names of persons and places and give them a higher probability.
- **Word properties**: When focusing on Stoichedon texts, word length is a relevant property. For this reason the candidate list can be restricted by both *exact length* as well as by *min-max thresholds*.

From a global perspective, every found word in a vocabulary is a potential suggestion candidate. To reduce this list of anywhere from several hundred thousand to several million words to a more reasonable size, the results of all selected algorithms are combined to a normalised score between 0 and 1 (Kruse 2009). In the last working step of this process, the candidates list (ordered by score in descending order) is then provided to the user.

Based on the aforementioned approaches the full paper will explain three different completion strategies:

1. Using only known information about a word (word length and some preserved characters),
2. using only contextual information such as word bigrams, co-occurrences, and classification data,
3. using all available information (combination of strategy a) and b)) of a word.

The main objective of this step by step explanation is to highlight both strengths and weaknesses of such a completely automatized system.

A video demonstration of the current implementation can be viewed at

http://www.e-humanities.net/lectures/SS2011/2011-DigClassSeminar/THATCamp_DevChallenge_BuechlerEckart_TextCompletion.ogv

References

Bocek, T., E. Hunt, and B. Stiller (2007). *Fast Similarity Search in Large Dictionaries.* Department of Informatics, University of Zurich.

Bodard, G., et al. (2009). *EpiDoc Cheat Sheet: Krummrey-Panciera sigla & EpiDoc tags*, 2006-2009. Version 1085, last accessed: Nov., 10th, 2009 [date] URL: http://epidoc.svn.sourceforge.net/viewvc/epidoc/trunk/guidelines/msword/cheatsheet.doc.

Bordag, St. (2008). *A Comparison of Co-occurrence and Similarity Measures as Simulations of Context*, 2008. In *CICLing*, Vol. 4919. Berlin: Springer (Lecture Notes in Computer Science).

Büchler, M. (2008). *Medusa. Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*. Saarbrücken: Vdm Verlag Dr. Müller.

Crane, G., and D. Bamman (2009). *The Latin and Ancient Greek Dependency Treebanks*, 2009. URL: http://nlp.perseus.tufts.edu/syntax/treebank/ last accessed: Nov., 10th 2009.

Firth, J. R., *A Synopsis of Linguistic Theory*. Oxford.

Fraser, P. M. E. Matthews, and M. J. Osborne (1987-2008). *A Lexicon of Greek Personal Names.* (In Greek and English), Vol. 1-5, Suppl. Oxford: Clarendon Press.

Heyer, G., U. Quasthoff, and T. Wittig (2008). *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse*. 2nd edition. Herdecke: W3L-Verlag.

Kruse, S. (2009). *Textvervollständigung auf antiken Texten*. University of Leipzig, Bachelor Thesis. pp 48-49. URL http://www.eaqua.net/~skruse/bachelor, last accessed on Nov., 10th 2009.

Kukich, K. (1992). Technique for Automatically Correcting Words in Text. *ACM Computing Surveys* 24(4).

Ottmann,T., and P. Widmayer (1996). *Algorithmen und Datenstrukturen.* Heidelberg: Spektrum Verlag.

Schierle, M., S. Schulz, and M. Ackermann (2008). From Spelling Correction to Text Cleaning – Using Context Information. In *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V*.