A Text Denormalization Algorithm Producing Training Data for Text Segmentation

Kilian Evang
University of Groningen
k.evang@rug.nl

Valerio Basile
University of Groningen
v.basile@rug.nl

Johan Bos
University of Groningen
j.bos@rug.nl

As a first step of processing, text often has to be split into sentences and tokens. We call this process *segmentation*. It is often desirable to replace rule-based segmentation tools with statistical ones that can learn from examples provided by human annotators who fix the machine's mistakes. Such a statistical segmentation system is presented in Evang et al. (2013).

As training data, the system requires the original raw text as well as information about the boundaries between tokens and sentences within this raw text. Although raw as well as segmented versions of text corpora are available for many languages, this required information is often not trivial to obtain because the segmented version differs from the raw one also in other respects. For example, punctuation marks and diacritics have been normalized to canonical forms by human annotators or by rule-based segmentation and normalization tools. This is the case with e.g. the Penn Treebank, the Dutch Twente News Corpus and the Italian PAISÀ corpus. This problem of missing alignment between raw and segmented text is also noted by Dridan and Oepen (2012).

We present a heuristic algorithm that recovers the alignment and thereby produces standoff annotations marking token and sentence boundaries in the raw test. The algorithm is based on the Levenshtein algorithm and is general in that it does not assume any language-specific normalization conventions. Examples from Dutch and Italian text are shown.

References

Dridan, R. and Oepen, S. (2012). Tokenization: Returning to a long solved problem. a survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382.

Evang, K., Basile, V., Chrupała, G., and Bos, J. (2013). Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426.

Presentation type preference: poster. Track submission: none