

The Markov Expert for Finding Episodes in Time Series

Jimming Cheng
Harvard University
jvcheng@fas.harvard.edu

Michael Mitzenmacher
Harvard University
michaelm@eecs.harvard.edu

We describe a domain-independent, unsupervised algorithm for segmentation of time series data into meaningful episodes, focusing on the problem of text segmentation. The VOTING EXPERTS algorithm of Cohen et al. [1] achieves results with fairly low rates of error by combining two experts that analyze the input's frequency and entropy patterns. The MARKOV EXPERT is a new approach that improves the performance of VOTING EXPERTS by further refining those results with votes from an additional expert.

The new expert applies a method inspired by Teahan et al.'s compression-based approach for Chinese text [2]. Their supervised approach requires a large, correctly segmented training corpus. Segmentation of the input is modeled as a Markov process, with spaces inserted such that the resulting string is smallest under PPM compression with respect to the corpus. In the unsupervised setting, external corpuses are not available. Thus, we draw event pattern data from a new corpus constructed by generating a preliminary segmentation using the original VOTING EXPERTS. Since VOTING EXPERTS finds episode boundaries fairly well (precision and recall around 77% and 75%), the quality of this new corpus is sufficient to allow the MARKOV EXPERT to further improve results significantly. The MARKOV EXPERT votes on possible boundaries by accumulating votes within a sliding window that moves over the input. The context within each window is compared to the corpus using a segmentation utility function. Quality of a particular segmentation is positively correlated to the frequency of the resulting suffixes and prefixes in the corpus, and negatively correlated to instances in which the current context appears intact within a word in the corpus.

Our experiments focus on comparisons between our MARKOV EXPERT and the original VOTING EXPERTS, as Cohen et al. already proved it to be much more effective for this problem than other algorithms, including the SEQUITUR algorithm. The MARKOV EXPERT was consistently superior over several human language inputs, as well as in recordings of high and low-level controllers for a wandering robot. Performance is gauged in F-measure for total boundaries found (harmonic mean of precision and recall), words captured exactly, words captured with false positives within and words lost.

| | F-measure | Precision % | F.P. Rate % | Exact % | Captured % | Lost % |
|----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| Joyce (English) | 76.8 / 82.1 | 73.2 / 85.5 | 19.1 / 21.1 | 44.9 / 59.9 | 51.7 / 72.9 | 5.4 / 2.0 |
| Orwell (English) | 75.3 / 79.8 | 74.8 / 80.9 | 24.2 / 21.3 | 44.4 / 54.2 | 54.7 / 65.5 | 5.1 / 3.6 |
| Song lyrics (Romaji) | 64.2 / 67.6 | 65.3 / 72.7 | 36.9 / 36.9 | 29.6 / 37.2 | 41.3 / 52.3 | 10.8 / 5.0 |
| Goethe (German) | 69.9 / 75.2 | 74.2 / 81.5 | 34.0 / 30.3 | 41.4 / 50.7 | 54.6 / 66.3 | 6.1 / 3.3 |
| Robot Data | 67.5 / 70.8 | 59.9 / 65.8 | 22.7 / 23.4 | 33.2 / 41.3 | 37.0 / 46.6 | 17.3 / 14.9 |

Segmentations optimized for F-measure. VOTING EXPERTS results on the left, MARKOV EXPERT results on the right

This work was supported in part by NSF grants CCR-9983832 and CCR-0121154.

- [1] P. Cohen, B. Heeringa, and N. Adams. An Unsupervised Algorithm for Segmenting Categorical Timeseries into Episodes. *Proceedings of ICDM 2002*, p. 99-106.
- [2] W.J. Teahan, Y. Wen, R.J. McNab, and I.H. Witten. A Compression-based Algorithm for Chinese Word Segmentation. *Computational Linguistics*, 26: 375-393, 2000.