StreamSage Unsupervised ASR-Based Topic Segmentation

Phil Rennert(phil.rennert@streamsage.com)

Abstract

StreamSage, Inc. is a pure ASR (automatic speech recognition) group; we've developed an unsupervised method for topic segmentation using ASR information only. For purposes of TRECVID, we partnered with the Dublin City University, a mostly video group, to generate all the required runs. We describe here our ASR-only topic segmentation. Our method was developed for in-depth story segments, and doesn't handle well the very short stories (sometimes only a sentence or two) found in broadcast news.

1.0 Description of task

This is done elsewhere (see other papers in this notebook), so we'll omit it.
Our purpose is to divide a document into topical segments.

2.0 Our segmentation method

Our method is totally unsupervised, independent of hand-tagged training data. While it was developed for the English language, it uses only word frequency and co-occurrence information, POS (part-of-speech) tagging and lemmatizing (stemming), so it could in theory be applied to another language. We have a corpus composed of three years of the New York Times which we've lemmatized, POS tagged, and parsed; from this we compute centrality-weighted frequency and mutual information for all words which appear..
There are three segmentation algorithms and a method for combining them.

2.1 Noun-link

The principle is to identify repeated nouns as topic words and link sentences containing them. We then mark segment boundaries in places which break no links.
The text is POS tagged and lemmatized, and divided into sentences. Then each sentence has non-content words removed. A content word is any noun which doesn't appear more frequently than once every 3.3 stories in the NYT; in addition, we add capitalized adjectives and translate certain closely-associated adjectives to the corresponding nouns, e.g. acidic --> acid. Then we define a link distance (dependent on average sentence length) and establish a full link between any two sentences within that distance containing the same content word. If the two sentences contain different content words with high mutual information, that counts as a partial link; multiple partial links add up to full links.
Then topic boundaries are assigned where no full links are broken (when a boundary location spans more than one sentence, mutual information is used as a tie-breaker).

2.2 Boc-Choi (bundles of content - Choi)

This is based on the work of Freddy Choi [Choi 2000]; it computes a similarity measure over sentences, forms the similarity matrix over a moving windows of five sentences, and looks for similarity minima between one window and the next. The boundaries are placed where the window similarities are deeper local minima.

2.3 Induced N-grams

The idea is to inductively determine the topic boundary markers common in news broadcast, such as "For XXX news, this is ...". The noun-link and boc-choi algorithms are used to determine candidate boundaries; then all 1-, 2-, and 3-grams found near these boundaries are listed as potential markers.
For each, percentage of all occurrences which are near candidate boundaries is computed; then the list is sorted and the top ones chosen as boundary markers. Although the candidate boundaries at this stage are not extremely accurate, the markers which emerge tend to be

extremely reliable boundary predictors, such as names of newscasters and other formal end-of-segment indicators.

2.4 Combination algorithm

These three algorithms are combined to produce final boundaries. The total list of boundaries from all three tends to have too many boundaries: good recall, but so-so precision, so it needs to be pruned. The induced n-gram markers tend to have excellent precision but less good recall, so they're believed whenever they indicate a boundary. If noun-link and boc-choi boundaries fall together or within a sentence or two, they're believed. Otherwise, boc-choi is used as a check on a tentative noun-link boundary and vice versa. A boundary confidence measure is computed at each sentence for the noun-link and boc-choi methods (obviously it's highest where the method indicates a boundary), and the confidence of the other method is used as the discriminant where only one method indicates a boundary.

3.0 TRECVID results and conclusions

Results were disappointing; our system clearly didn't find enough boundaries. With the tuning we used (aimed at segments at least 12 sentences long), many of the shorter segments were invisible; however, when we tuned for very short stories, the larger ones got cut into too many pieces. If we want to apply our system to broadcast news, it seems clear we'll have to add a short segment detection module, perhaps using video information such as Dublin City University's anchorperson detection technique [DCU 2003].

4.0 References

[Choi 2000] Freddy Y. Y. Choi. Advances in independent linear text segmentation. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pages 26Q33, Seattle, May 2000.

[DCU 2003] Paul Browne, Csaba Czirjek, Georgina Gaughan, Cathal Gurrin, Hyowon Lee, Seán Marlow, Kieran Mc Donald, Noel Murphy, Noel E. O’Connor, Neil O’Hare, Alan F. Smeaton, Jiamin Ye. Dublin City University Video Track Experiments for TREC 2003