Making Semantic Topicality Robust Through Term Abstraction

Paul M. Heider
pmheider@buffalo.edu
Department of Linguistics, University at Buffalo, The State University of New York

Rohini K. Srihari
rohini@cedar.buffalo.edu
Janya, Inc. (http://www.janyainc.com)

The Problem

Topic Segmentation:
- Can we divide a document into smaller coherent units?

Halliday and Hasan (1976)'s Answer:
- Yes. Abstract semantic units are realized through sentences that can be grouped based on their internally strong cohesion and externally weak cohesion.

Our Contribution

We present a topic segmenter which breaks from earlier work by focusing on semantic cues to define coherent conversations within a larger, multi-party document. Using word’s thesaurus entry as a proxy for its underlying semantic provides a domain-neutral metric for distinguishing conversations.

Methods

The Task: use semantic cues to divide each log into natural conversation topics

The Corpus: chat logs taken from the #linux channel on freenode.net

The Gold Standard: hand-annotated segmentations done by six university students (Elsner and Charniak, 2008)

Is the final line of text below (taken from the corpus) a better match for conversation topic A or B?

Comparison Metrics

1-to-1 Measure
- What alignment of conversations maximizes the global overlap in two annotations?

loc3 Measure
- How much agreement is there for each window of three utterances?

Many-to-1 Measure
- Was one annotation done at a finer level of analysis but with the same major topics?

Similarity Measures

Headword similarity

The headword profile vector for each line of text is compared with likewise constructed vectors for each potential conversation topic.

| Headword | Present | Test | Avg? | Above | Below | No |
|-----------|---------|------|------|-------|-------|----|
| Base      |         |      | +1   | +0.5  | -0.1  |    |
| No        |         |      | +0.5 | +1    | -0.05 |    |
| Min       |         |      | -1   | -0.5  | +0.001|    |

The table at left shows the values used to construct a profile vector.

Classifier Inputs

Same Interlocutor
- A speaker tends to continue in the conversations she has already participated in

Interlocutor's Name Used in Text
- A speaker can overtly affiliate with the conversation of another user by mentioning their name (see O’Neill and Martin, 2003)

Thesaural Headword Similarity
- Semantically related sentences contain tokens that share many of the same headwords

Local agreement (loc3) is the most useful metric.

Headword | Token
|-----------|---------|
| 477       | EVADE   |
| 544       | ESCAPE  |
| 623       | FEIGN   |
| 773       |         |
| 927       |         |
| 940       |         |

Evidence of Oversplitting

Our model has a strong predilection to split a document into smaller conversations.

| E&C Annotators | Our Model | E&C's Best Baseline |
|----------------|-----------|---------------------|
| Mean           | 81.33     | 128                 |
| Avg. Length    | 10.6      | 16.0                |
| Conversations  | 80        | 133                 |
| E&C Annotators | 81.33     | 128                 |
| Mean           | 16.0      | 16.0                |
| Min            | 6.2       | 5.2                 |

A relatively high many-to-1 score supports the idea that our model agrees with human annotators on the major boundaries.

Future Work

- Use ranked similarity (vs. winner-takes-all)
- Adjust the granularity to a more human level
- Disambiguate word sense before introducing thesaurus headwords

Final Notes

References

Micha Elsner and Eugene Charniak. You talking to me? A corpus and algorithm for conversation disentanglement. In The Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008), Columbus, Ohio, 2008.

Michael Halliday and Ruqaiya Hasan. Cohesion in English. Longman Group, New York, 1976.

Jacks O’Neill and David Martin. Text chat in action. In GROUP ’03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, pages 40–49, New York, NY, USA, 2003. ACM Press.

Peter Mark Roget, editor. Roget’s Thesaurus. Project Gutenberg, 1991. URL http://www.gutenberg.org/etext/22.

Grady Ward, editor. Moby Thesaurus List. Project Gutenberg, 2002. URL http://www.gutenberg.org/etext/3202.

Funding

The first author was funded through a fellowship from the SUNY at Buffalo Department of Linguistics and through a research assistantship at Janya, Inc. (Air Force Grant Nos FA8750-07-C-0077 and FA9750-07-D-0019, Task Order 0004).