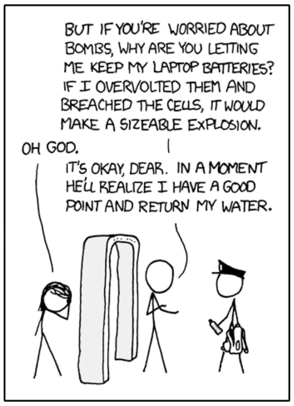


Topic Segmentation

David Kauchak
cs458
Fall 2012

Administrative

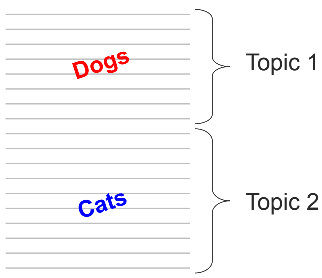
- Schedule
- Readings
- Lunch today!
- HW4 due tomorrow
- Attendance

Today's class

- Blend of introductory material and research talk
  - Problem of topic segmentation
  - Common data to work with
  - Approaches
  - Evaluation
  - Some initial results
- Represents a change in the direction of the course

Topic segmentation

Linearly break text into related regions



- Longer documents such as magazine articles, news articles or books often talk about multiple topics
- Topic segmentation approaches attempt to break up these documents into sub-documents

Topic segmentation for IR

- How is this useful for IR?
  - return sub-document rather than whole document
  - snippet/summary generation
  - IDF calculations
  - co-occurrence statistics
- Many long documents are becoming available digitally
  - Magazine articles
  - News articles
  - Books
    - Project Gutenberg
    - Million books project at CMU
    - Online Books Page at U. Pennsylvania
    - Google books

Data: Broadcast news

Good Morning, Im Judy Simpson. Vermont Congressman Peter Welch got an earful Saturday at two town hall meetings on health care reform. Welch called for a civil discussion -- and this meeting at the Williston Central School remained civil -- but still contentious. People feel strongly on both sides.

Matt Kimball/Burlington: "I feel that as a citizen of this country I should be guaranteed health care -- as a human right." – applause

Joeie Clark/Fairfax: "I'm wondering why our representatives are trying not to let us read the bills before you vote on them." Welch told the crowd that he remains a strong supporter of single payer, or universal government insurance.

Middlebury College now has 10 suspected cases of the H1N1 virus. The students have all been quarantined to try and prevent the spread of swine flu -- but are being allowed to use the cafeteria with other students. UVM and St. Michaels College have also had confirmed cases of swine flu. And at Dartmouth College in New Hampshire -- 175 students have reported influenza like symptoms. Two-thirds are thought to have H1N1.

Identify when the story changes

Data: Expository texts

Magazine/Journal articles

Rudi Spiller, CEO of Germany's Leica Camera AG, has a square jaw and stands as though rooted to the floor. His hair is closely cropped, and behind his rimless glasses his eyes are keen.

In opening remarks at a recent press conference heralding three new Leica products, Spiller described the company's latest triumphs with the confidence of a supervillain proclaiming world domination. After greeting the inhabitants of planet Earth with salutations in a dozen languages, he waxed poetic about Leica's commitment to excellence, and included a simple summation: "Our precision. Is. Perfect."

The Leica M9, which debuted that day last month, is the latest in a line of rangefinder cameras that stretches back half a century. "This camera became an instant classic in 1954, and is now perfectly designed," says Spiller, causing listeners to wonder whether the world will ever need an M10.

It has been three years since the much-heralded release of the M8, the previous model in Leica's M series, which was the first digital in the line (and a year since the M8.2 update). Wired.com visited the Leica factory in Solms, Germany, before the M9 was released and looked behind the scenes to see how Leica builds its renowned cameras.

Identify sub-sections

Data: Narrative texts

Books

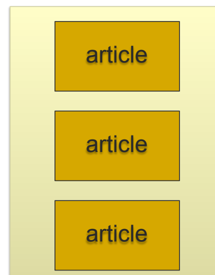
In the early nineteen seventies, a British photo retoucher named Robert Stevens arrived in south Florida to take a a job at the National Enquirer, which is published in Palm Beach County. At the time, photo retouchers for supermarket tabloids used an airbrush (nowadays they use computers) to clarify news photographs of world leaders shaking hands with aliens or to give more punch to pictures of six-month-old babies who weigh three hundred pounds. Stevens was reputed to be one of the best photo retouchers in the business. The Enquirer was moving away from stories like "I Ate My Mother-in-Law's Head," and the editors recruited him to bring some class to the paper. Thy offered him much more than he made working for tabloids in Britain.

Stevens was in his early thirties when he moved to Florida. He brought a red Chevy pickup truck, and he put a CB radio in it and pasted an American-flag decal in the back window and installed a gun rack next to the flag.

Identify chapters or sections

Data: Synthetic

concatenate TDT articles



How hard is this problem?

Previous approaches have achieved error rates of 10%-20% on non-narrative data sets

(Hearst, 1994) examined the problem of paragraph identification

7 humans were asked to identify paragraphs

How well do you think people did?

Error rates were ~25%

Data Sets

Broadcast news

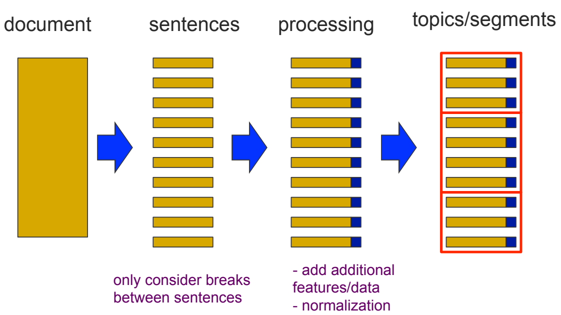
Expository Texts

Narrative texts

Synthetic Texts

How can we do this?

General problem setup



Data Set Cues

Broadcast news (Beeferman et al., 1999)
- Cues at boundaries, "Coming up…"

Synthetic Texts (Brants et al., 2002; Li and Yamashi, 2000)
- Strong topic shifts

Expository Texts (Hearst, 1994)
- Repetition of terms within segments

Narrative Texts (Kauchak, 2005)
- challenging

Previous Methods

Similarity based

Using Lexical Chains

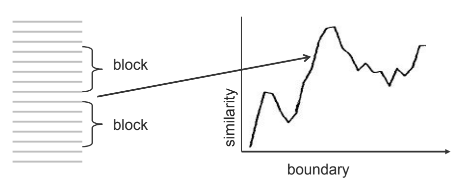
Feature based

Similarity based

Text on the same topic should be more similar to text on a different topic

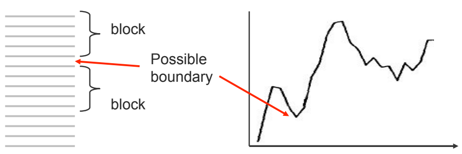
Calculate the similarity between adjacent segments of text and plot

Where are the boundaries?



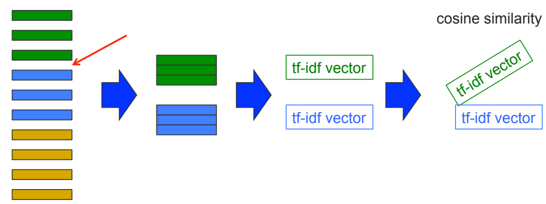
Similarity Based

Segment boundaries are identified by troughs in the similarities



Similarity Measures

Cosine Similarity of word frequency vectors

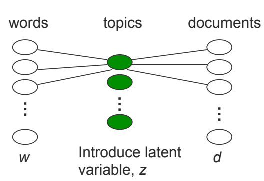


Similarity Measures: PLSA

Probabilistic latent semantic analysis (PLSA)

What is the probability of a word occurring in a particular document? p(w,d)

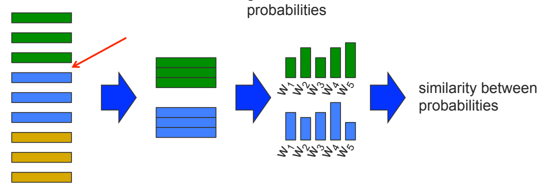
Rather than model directly, associate words with topics and documents as a blend of topics



Similarity Measures: PLSA

for each word, calculate the probability of occurring in each block

gives us a vector of word probabilities



Lexical chains

A lexical chain is a sequence of word occurrences where every word occurs within a predefined distance and each work is connected by a lexicographical relationship

Relationships: synonymy, part/whole, specialization/generalization

?

The dog and the cat are friends.

The cat likes the dog because they play together.

That furry feline loves to play so much.

It frolics around and around and around.

Lexical chains

A lexical chain is a sequence of word occurrences where every word occurs within a predefined distance and each work is connected by a lexicographical relationship

Relationships: synonymy, part/whole, specialization/generalization

The dog and the cat are friends.

The cat likes the dog because they play together.

That furry feline loves to play so much.

It frolics around and around and around.

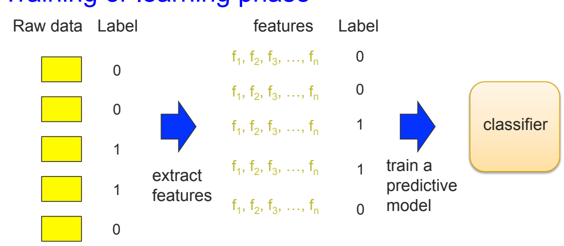
Lexical chains

How might we use lexical chains to identify topic boundaries?

- Boundaries are located where there is a high density of chain beginnings and endings
- Boundaries have few lexical chains crossing them

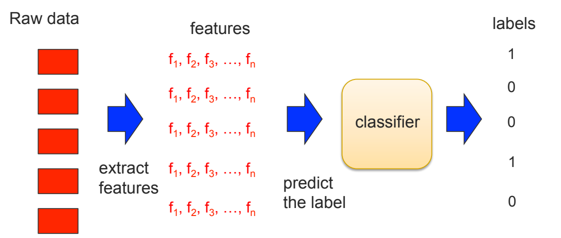
Feature based: Binary classifier setup

Training or learning phase



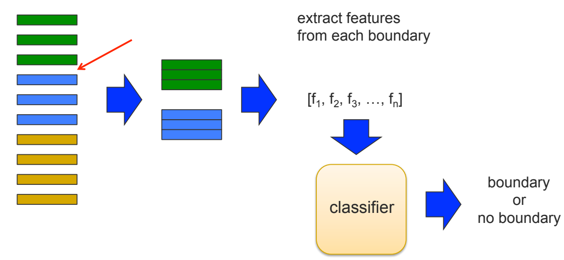
Feature based: Binary classifier setup

Testing or classification phase



Feature based

Learn a "boundary" classifier



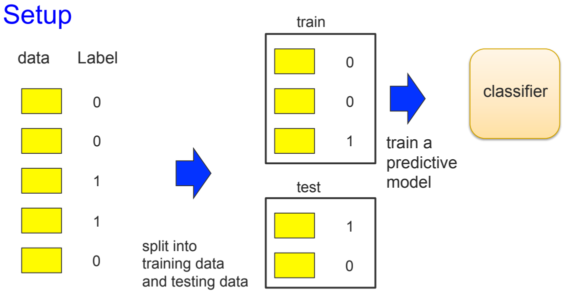
Features?

What types of things would be indicative of a shift in topic (or alternatively of staying within a topic)?

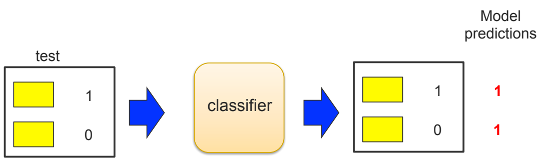
- Particular words
  - Word groups (e.g. synonyms)
- Pronouns within 5 words from beginning
- Lexical chain features
- Part of a conversation?
- Number of named entities
- Number of synonyms to the right and left
- Full name
- numbers
- …

Evaluation

Setup



Evaluation



Need to quantify the results by comparing predicted labels vs. “true” labels

Evaluation metrics

actual label	predicted label
0	0
0	1
0	1
1	0
0	0
0	0
1	1
0	0
0	0

How can we quantify the performance?

Evaluation metrics

Precision? Recall? – not so good ☹

Word error probability (Beeferman et al., 1999)
- Probability that a pair of words *k* words apart is inconsistently classified

Sentence error probability (Beeferman et al., 1999)
- Probability that a pair of sentences *s* words apart is inconsistently classified

WindowDiff (Pevzner and Hearst, 2002)
- Sliding window over the data
- Difference between number of hypothesized segments and the number of actual segments

Example data set

Two narrative, non-fiction books
- Contain author identified boundaries at a finer level than chapters, which are taken as segment boundaries
- *Biohazard* by Ken Alibek
  - 5858 possible boundaries, where boundaries are between sentences
  - 213 actual boundaries
- *The Demon in the Freezer* by Richard Preston
  - 4466 possible boundaries, where boundaries are between sentences
  - 120 actual boundaries

Grolier’s encyclopedia articles
- 1000 articles, containing major and minor headings, which are taken as segment boundaries
- 102,116 possible boundaries

Performance of some approaches

Similarity
- Within segment similarity is similar to across boundary similarity
- PLSA and TextTiling (cosine similarity) perform similarly to random

Cue based
- No words occur significantly at both training and testing boundaries

Lexical chains
- Lexical chain occurrences are not correlated with boundaries

Narrative document properties ☹

Segment Similarity

Examine adjacent block similarities and compare within segment similarities and similarities crossing segment boundaries

PLSA	Average	Standard Deviation
Within segment	0.903	0.074
Across boundary	0.914	0.041

Narrative document properties ☹

Vocabulary

25% of the content words in the test set do not occur in the training set

33% of the content words in the test set occur two times or less

Narrative document properties ☹

Boundary words

474 terms occur in first sentences of training boundaries

103 of these words occur at test boundaries

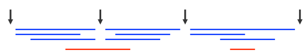
Only 9 *significant* words that occur in training occur in the test

No words occur significantly at both training and testing boundaries

Narrative document properties

☹

Lexical chains



- Created chains using synonymy and repetition
- 219 chains
- 2 begin and 1 ends at a boundary
- 20% of the chains cross boundaries
- Average segment length is 185
- Average distance to closest beginning chain is 39 words
- Average distance to closest ending chain is 36 words
- About 4 chains per segment

Biohazard and *Demon in the Freezer* (SVM)

	Word Error	Sentence Error	Window Diff	Sent. Error improv.
Biohazard random (sent.) random (para.)	 0.488 0.481	 0.485 0.477	 0.539 0.531	 (baseline)
Biohazard exp1 → test exp2 → test 3x cross validation	 0.367 0.344 0.355	 0.357 0.325 0.322	 0.427 0.395 0.404	 **25%** **32%** **24%**
Train *Biohazard* Test *Demon*	 0.387	 0.364	 0.473	 **25%**

Grolier's results

	Word Error	Sent. Error	Window Diff
random	0.482	0.483	0.532
Cosine Sim	0.407	0.412	0.479
PLSA	0.420	0.435	0.507
features (stumps)	**0.387**	**0.400**	*0.495*
features (SVM)	**0.395**	**0.398**	*0.503*

All methods given same number of segments (expected number of segments)

Concluding comments

Difficult problem: still room for improvement

Topic granularity level: how do you predict the number of segments in the text?

Hand crafting features is time consuming

Sequential data needs sequential models

References

- D. Beeferman, A. Berger and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34:177-210
- T. Brants, F. Chen and I. Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of CIKM*, 211-218.
- M.A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *ACL*, pg. 9-16.
- H. Kozima and T. Furugori. 1994. Segmenting narrative text into coherent scenes. In *Literary and Linguistic Computing*, 9:13-19.
- H. Kozima. 1993. Text segmentation based on similarity between words. In *ACL*, pg. 286-288.
- H. Li and K. Yamanishi. 2000. Topic analysis using a finite mixture model. In *EMNLP*, pg. 35-44.
- J. Reynar. 1999. Statistical models for topic segmentation. In *ACL*, pg. 357-364.
- N. Stokes, J. Carthy, A. Smeaton. 2002. Segmenting broadcast news streams using lexical chains. In *STAIRS*, PG. 145-154/

Discussion

Google announces new search tools

http://www.cnn.com/2012/08/08/tech/web/google-search-tools/index.html

Good additions?

Siri?

Do you want to see your e-mail in search results?

An aside… 30 trillion unique URLs… crawls 20 billions sites per day (231,481 sites/second) ☺

Analysis of features

holdout: 74 actual boundaries and 2086 possible boundaries

	boundary	non-boundary
Paragraph	74	621
Entity groups	44	407
Word groups	39	505
Numbers	16	59
Full Name	2	109
Conversation	0	510
Pronoun	8	742
Pronoun ≤ 5	1	330

Analysis of features

Perfect recall

	boundary	non-boundary
Paragraph	74	621
Entity groups	44	407
Word groups	39	505
Numbers	16	59
Full Name	2	109
Conversation	0	510
Pronoun	8	742
Pronoun ≤ 5	1	330

Example reweighting

- Substantially more negative examples than positive
- Simply classifying all examples as negative results in reasonable classification performance
- Must reweight positive vs. negative examples
- Iteratively change weighting
  - Train
  - Test
  - Stop when expected number of segments based on the training data is approximately found in the test set