Concept Based Classification (CBC) for Multi-Document Summarization

Asli Celikyilmaz, Dilek Hakkani-Tür

Microsoft Speech Labs

asli@ieee.org, dilek@ieee.org

Summarization

- Distill the most important information from:
  - A single document (e.g., a news story, a voicemail)
  - Multi-documents (e.g., news stories about an event, reviews of a product)
  - Speech (e.g., lectures, meetings)

- Information overload:
  - A variety of sources.
  - May have redundancy.
  - Limited time to process.
  - Information not necessarily in the right ordering.

Image from:
http://blogs.nyit.edu/library/2008/06/25/info-overload-the-problem/

- Text summarization research facilitated by NIST DUC and TAC evaluations.
  - A set of documents on a topic, paired with corresponding human summaries.

Related Work

- Summarization as classification

  (Radev *et al.,* 2004; Nenkova and Vanderwende, 2005)

  - Document sentences are marked as summary and non-summary sentences using similarity measures.
  - Binary classification, with features: Sentence length, position, average term frequency, etc.
  - Issue: word based similarity measures fail to capture semantic similarity.

- Generative models

  (HierSum, Haghighi and Vanderwende, 2009;

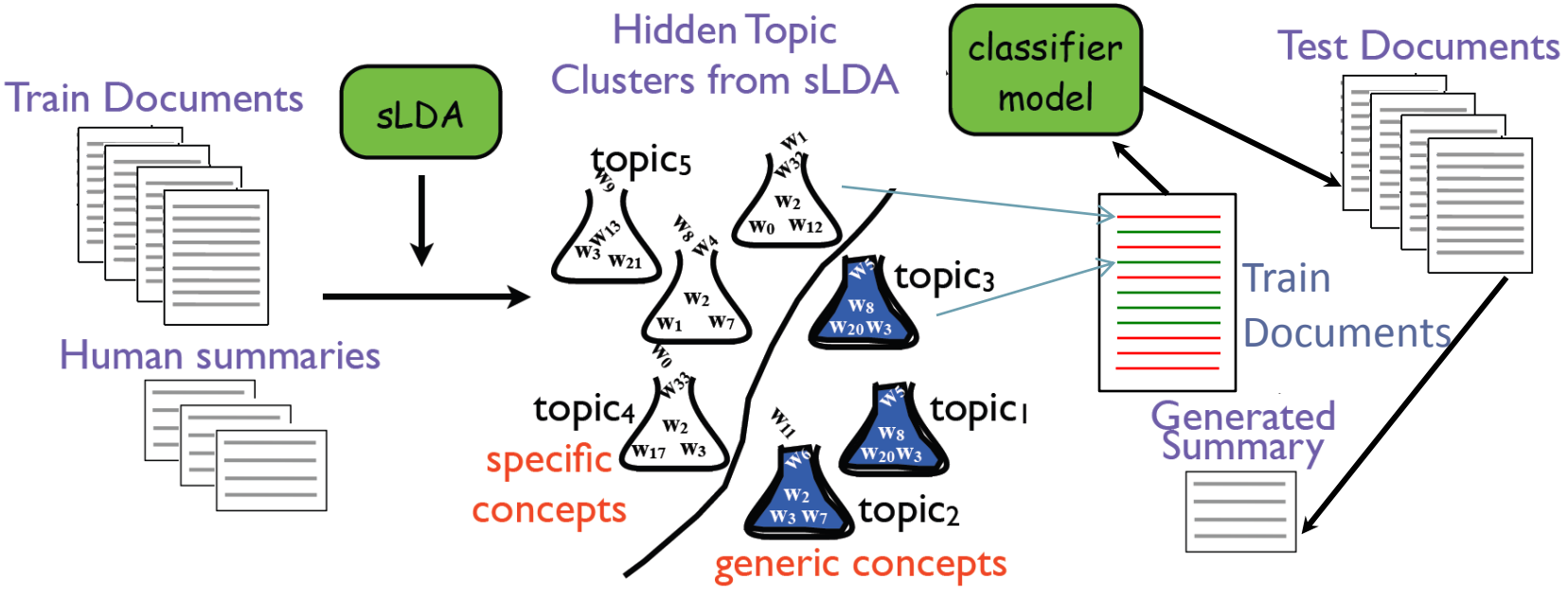
  sLDA, Celikyilmaz and Hakkani-Tür, 2010)

  - Aim to capture hidden concepts in documents, using Latent Dirichlet Allocation (LDA) -based models.

sLDA for Summarization Overview

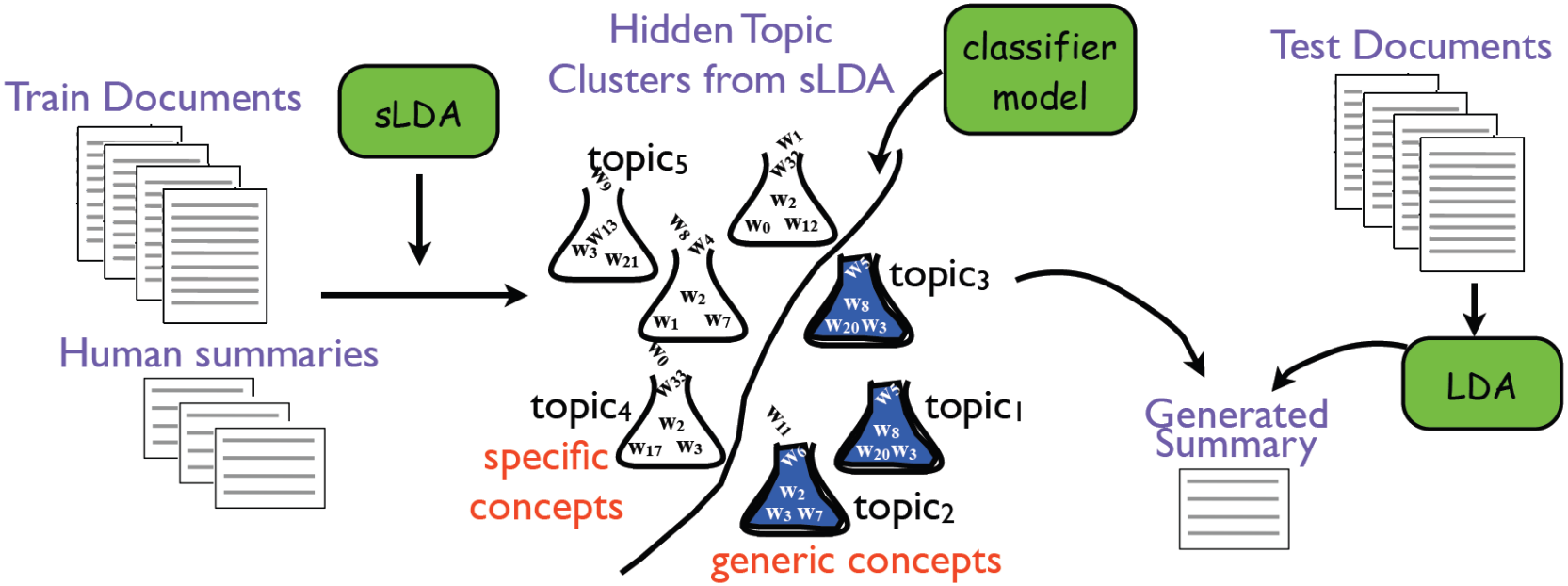
(Celikyilmaz and Hakkani-Tur, Interspeech 2010)

- Semi-supervised extractive summarization model based on latent concept clustering and sentence classification.
- Based on:
  - two types of hidden concepts (specific and generic) being mentioned in documents.
  - Generic concepts are the ones that are included in the human summaries.



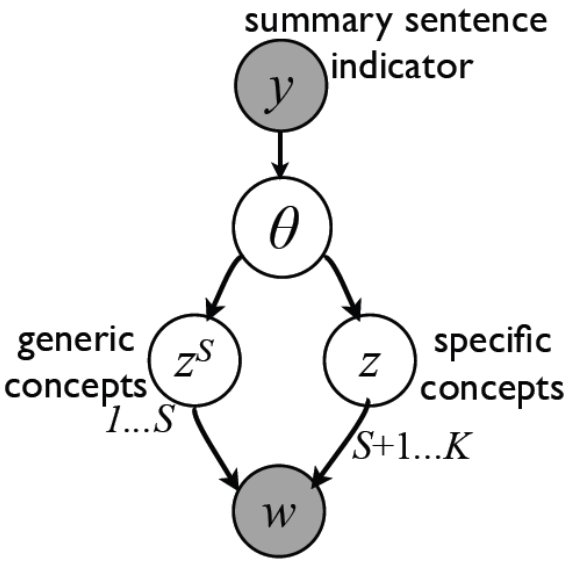
CBC for Summarization Overview

- Semi-supervised extractive summarization method based on latent concept clustering and topic classification.
- Based on:
  - two types of hidden concepts (specific and generic) being mentioned in documents.
  - Generic concepts are the ones that are included in the human summaries.



sLDA – Summary Focused LDA

- Extension of LDA for the summarization task.
- A generative model. Sets of observations can be explained by unobserved groups that explain why some parts of the data are similar.
- Each document is a mixture of a small number of topics.
- Each word's creation is attributable to one of the document's topics.
- If a word exists in a human summary, then the summary-topic mixing variables θ for topics k=1…S allocated for the generic topics are updated.



y – indicates our prior belief on a given n-grams likelihood of being generated from a specific or generic concept.

1…S- topics are designated for generic topics and the rest for non-summary topics.

CBC Features

- Given a document set, most frequent n-grams are captured: $v_{freq}^{uni} = \{w_i^*\}_{i=1}^{f_1}, v_{freq}^{bi} = \{w_i^{**}\}_{i=1}^{f_2}$
- Given a hidden concept z_i^m, the topic-word frequency is:

$$x_{im} = \hat{p}(z_k^m | w_i^*) * p(w_i^*)$$

- Given a concept, we measure the percentage of (frequent) unigrams that are considered specific or generic:

$$x_{im} = \frac{1}{f_1} \sum_{i=1}^{f_1} \mathcal{I}\left[\hat{p}(z_k^m | w_i^*) \geq \mathcal{T}\right]$$

- Maximum Entropy Classification

Inference

- When a new document set is given, we:
  - Build standard LDA model on each set.
  - Use the CBC model to predict the labels of the *K* topics as generic or specific.
  - Calculate sentence scores for each document sentence based on the frequency of predicted hidden “generic” topics they contain.
  - Merge the scores based on CBC, with scores based on term frequencies, following previous work (Nenkova and Vanderwende, 2005).

Sentence Scoring

- We predict the class of each hidden topic in each sentence to score each sentence:

$$r(s_m^{CBC}) = \gamma_1 * \frac{1}{K} \sum_{k=1}^{K} \hat{p}(y_{z_k}^{s_m} = 1) + \gamma_2 * r(s_m^{sLDA})$$

- Using the expected topic-word likelihood in a given sentence, sentence rank score is determined based on frequency of hidden generic topics it contains.

$\hat{p}(y_{z_k}^{s_m} = 1)$: Predicted score for sentence s_m, indicating that the topic it includes is a generic concept.

$r(s_m^{sLDA})$: Predicted score via original sLDA.

Sentence Scoring-2

- Ranking salient sentences based on interpolated sentence score:

$$r(s_m) = A * r(s_m^{CBC}) + B * r(s_m^{uni}) + C * r(s_m^{bi})$$

  - $r(s_m^{CBC})$: Predicted score from CBC.
  - $r(s_m^{uni})$: Rank score of unigrams in a sentence that have a high likelihood of appearing in a summary. Normalized total count of high frequency unigrams that the sentence, s_m, contains.
  - $r(s_m^{bi})$: Rank score of bi-grams in a sentence that have a high likelihood of appearing in a summary.

Redundancy Elimination

- Greedy search when forming the system summary:
  - Start from the highest scoring sentence
  - Add sentence to the summary only if:
    - Word overlap between the sentence and summary is low.
    - Otherwise skip the sentence

    Until the summary length is satisfied.

Data Sets and Evaluation

- Rouge-1, Rouge-2, Rouge-SU4: measures overlap of unigrams, bigrams and gappy bi-grams between human and system summaries.
- Training data:
  - DUC 2005 & 2006 data sets
  - 100 documents sets, each containing 25 news articles (~80K sentences)
- Test Data
  - DUC 2007
  - 45 document sets, each containing 25 news articles (~25K sentences)
- Task: Create 250 word summaries for each document set.

Results and Discussion

- Baseline: Cosine similarity instead of sLDA.
- PYTHY (Toutanova et al., 2007): summarization as classification. Top performer of DUC-2007.
- HierSum: (Haghighi and Vanderwende, 2009): generative method with hierarchical models and KL divergence.

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline	41.0	9.3	15.2
PYTHY	42.6	**11.9**	16.8
HierSum	42.4	11.8	16.7
sLDA	45.4	11.5	17.3
CBC	**46.1**	11.7	**17.5**

Conclusions and Future Work

- Learning summary content distributions from document sets provided human summaries as supervision.
- Shown improvements in terms of ROUGE scores.
- Improve the sentence selection/redundancy elimination, for example using the Integer Linear Programming framework (Gillick et al; 2009) proposed earlier.
- Hierarchical topic models proposed earlier.

Thank you!

sLDA – Summary Focused LDA

- Original summary documents are assigned a category (summary vs. non-summary) according to the summary topics they include.
- Maximum entropy classification with word frequency features to determine if a sentence should be included in the summary or not.
- Instead CBC learns to categorize topics as summary versus non-summary topics.

Summarization Challenges

- Finding meaning similarity is still unsolved.
- Human evaluation is expensive, automatic evaluation is less accurate.
- Few annotated data sets.
- Most extraction methods do not consider coherence.

Related Work:
Unsupervised Methods

- Maximum Marginal Relevance (MMR) (Carbonell and Goldstein, 1998)

$$score_{i,t} = \lambda \times sim_1(sent_i, query) - (1-\lambda) \times sim_2(sent_i, summary_t)$$

- Topic clusters and centroids (Radev et al., 2004)
  - Centroid: a pseudo document that contains terms that have a count*IDF score higher than a threshold
  - Sentences that contain words from the centroid are more indicative of the cluster.
  - Sentences are scored based on their position, length and similarity to centroid, etc.

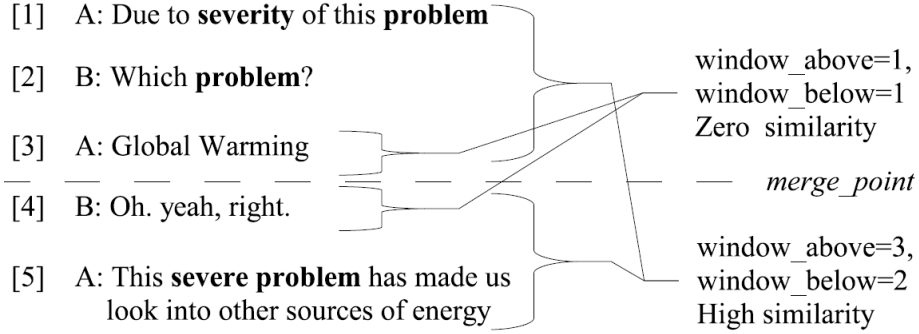
- Advantages:
  - No training data required (except for parameter tuning)
  - Domain independent
  - Can model redundancy
- Disadvantages:
  - Original versions did not use labeled data
  - Mainly greedy solutions (and can stuck at local optima)

Related Work:
Supervised Classification-Based Methods

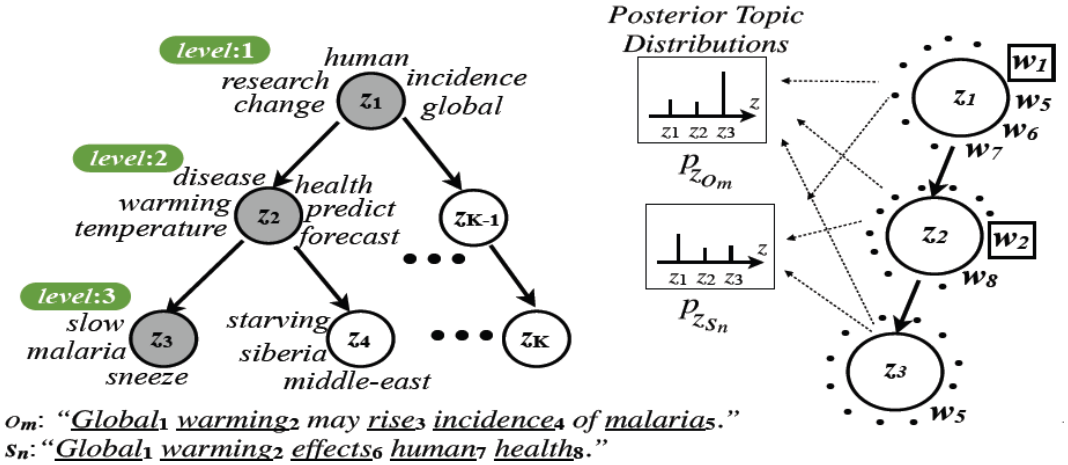
- e.g., (Kupiec et al., 1995; Teufel & Moens, 1997; among others)
- Sentences that should be included in summary are marked (using human summaries).
- Features, such as sentence similarity to the original document and sentence position in the document are extracted.
- Advantages:
  - can learn from labeled data sets
  - simple and effective
- Disadvantages:
  - require labeled training data
  - difficult labeling task (low agreement)
  - domain dependent
  - sparse data for some genre (e.g., meetings, lectures)
  - modeling of redundancy difficult in classification

Other Summarization Work

- Graph-based Methods for Meeting Summarization (N. Garg, K. Riedhammer, B. Favre, and D. Hakkani-Tur)



- Topic modeling (LDA and hLDA) for multi-document summarization (A. Celikyilmaz and D. Hakkani-Tur)



Outline

- Multi-Document Summarization (MDS)
- Concept-Based Classification for MDS
- Inference and Sentence Scoring
- Experiments
- Conclusions

Summarization

- Extractive:
  - Select sentences from the original set.
- Abstractive:
  - Distill information with new wording.

- Query, question, or topic focused:
  - Distill information with respect to a user's information need.
    - *What happened in Chile?* vs. *What is the capital of Chile?*
- Generic:
  - Generate a summary appropriate for a general audience.

Example: Multi-Document Summarization

QUERY: Describe the earthquake in Chile focusing on the following tsunami

SOURCE 1:
A day after Chile's great earthquake, the ground is still shaking, remnants of the resulting tsunami still slosh across the Pacific -- and there is not a geologist alive who is truly surprised that the catastrophe happened where it did.

Chile is right on the "ring of fire" -- the fault lines, all around the perimeter of the Pacific Ocean, that make for some of the most violent and frequent earthquakes on the planet.

The one that struck on Saturday morning will …

SOURCE 2:
A massive 8.8-magnitude earthquake struck Chile early Saturday, killing at least 78 people, collapsing buildings and setting off a tsunami.

Chilean TV showed devastating images …

Tsunami warnings were issued over a wide area, including South America, Hawaii, Australia and New Zealand, Japan, the Philippines, Russia and many Pacific islands.

…

SOURCE 3:
On Saturday, Feb. 27, a massive earthquake hit the coast of Chile at 3:34 a.m. and caused a tsunami that came in three waves and surged over 200 meters high, according to an article on nydailynews.com.

As of Tuesday, 795 people were reported dead and over 1.5 million homes damaged or destroyed after the 8.8-magnitude quake shook the earth.

…

…

EXTRACT:
A massive 8.8-magnitude earthquake struck Chile early Saturday, killing at least 78 people, collapsing buildings and setting off a tsunami. …

ABSTRACT:
An 8.8-magnitude earthquake that hit Santiago, Chile on February 27, 2010, caused a tsunami. …