ILLINOIS INSTITUTE OF TECHNOLOGY

Other Tasks: Topic Segmentation and Summarization

CS-585

Natural Language Processing

Sonjia Waxmonsky

Based on slides from Derrick Higgins and Kai Shu

TOPIC SEGMENTATION

Segmentation

Words in a Sentence

The [$_{ORG}$New York Times] reported that the summit would be in [$_{PLACE}$Norway].

```
O B-ORG I-ORG I-ORG  O       O
The New York Times reported that

O     O       O   O  O  B-PLACE O
the summit would be in Norway .
```

Sentences in a Text

The New York Times reported that the summit would be in Norway.

Delegates are to begin arriving next week.

In other news, oil prices saw a sharp drop this week as GDP slowed.

OPEC nations plan an emergency meeting to agree on production targets.

In sports, the Champions League kicked off this week with a victory by German champions Dortmund over PSG.

Topic Segmentation

- **Topic segmentation**: dividing a text into units (groups of sentences) that are semantically related – on the same general topic
- Generally an **unsupervised task** – not common to have a list of known topics to start with and an annotated corpus to learn from
- Other possible segmentation criteria: discourse-functional, speech acts

Topic segmentation

- Probabilistic models, e.g., HMM
  - Transition probabilities are from latent topics to latent topics (instead of latent tag, for example, in POS tagging)
  - Observation probabilities are models of sentence likelihoods given topic (rather than, e.g., word likelihoods given POS tag)
- Generative variants can be trained in either supervised or unsupervised fashion

SUMMARIZATION

Extractive vs. Abstractive Summarization

Extractive

Across the country, college campuses have become ghost towns. Students and professors are hunkered down inside, teaching and learning online. University administrators are tabulating the financial costs of the Covid-19 pandemic, which already exceed the CARES Act's support for higher education. The toll of this pandemic is high and will continue to rise. But another crisis looms for students, higher education and the economy if colleges and universities cannot reopen their campuses in the fall.

Abstractive

Campuses have closed, but another crisis looms for universities and colleges if they cannot reopen in the fall.

Methods for Extractive Summarization (1): Sentence Scoring

Unsupervised Approaches

What makes a good summary?

- *Frequency*: Topic/information gets mentioned a lot
  - For example, assign each sentence a score based on the average probability of its content words
- *Centrality*: Topic/information is close to the center of the semantic space represented by the document
  - For example, sentence scores based on similarity with other sentences in the same document (higher is better)

Supervised Approaches

We can train a model based on human-created (extractive) summaries: which sentences are *actually* included in the summary?

Possibly also label sentences for relative importance

Methods for Abstractive Summarization

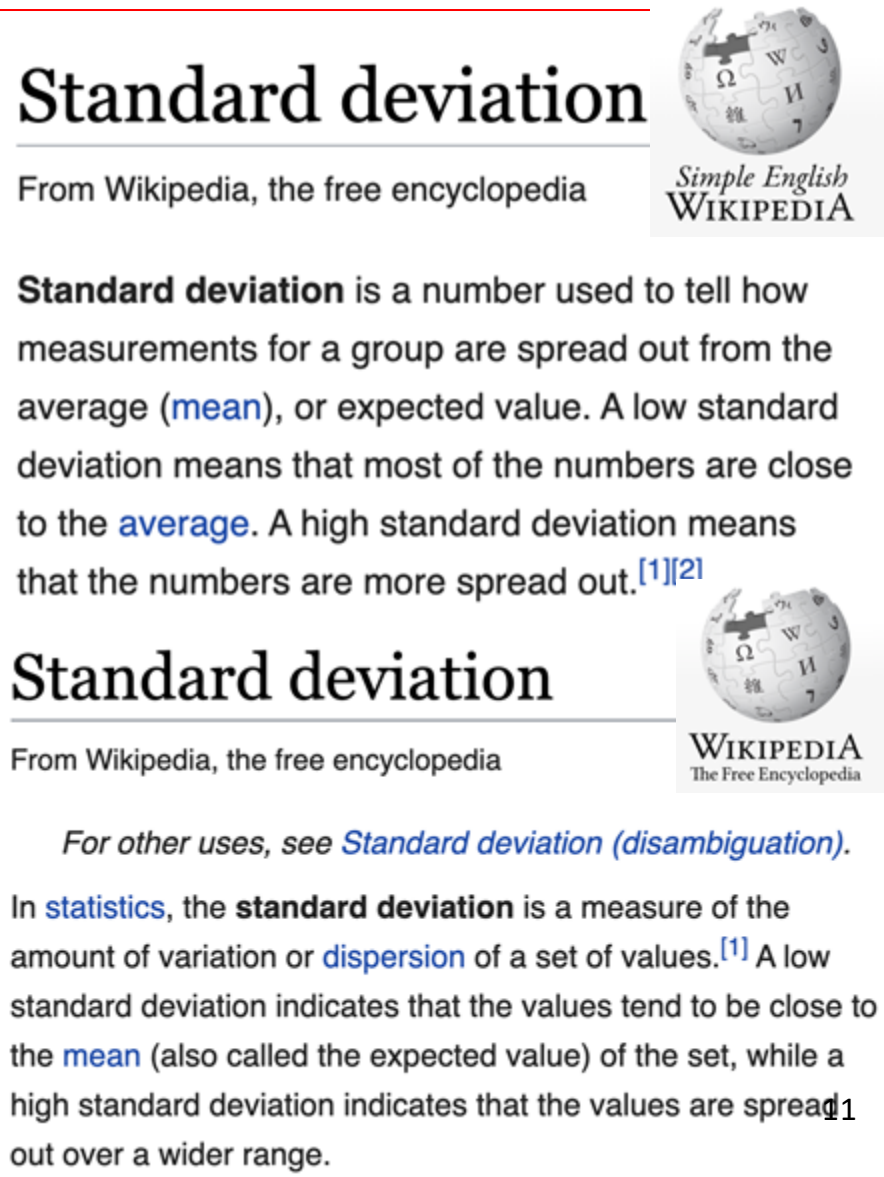
- Sequence-to-sequence neural network models have great potential for summarization
- Just as in MT, seq-to-seq models can learn from bilingual corpora, abstractive neural summarization models aim to learn from collections of text with summaries (or abstracts, for example)
- Challenges
  - Training data not as readily available
  - Need to process long documents (and consequent issues with memory capacity, vanishing gradients)
  - Problems with OOV words
  - Capturing global information about salience and topicality

Evaluation of Text Summarization Models

- Similar to machine translation, in that ground-truth evaluation involves manual review and is very costly
- Multiple criteria:
  - Coverage: Including all key information
  - Diversity: Lack of redundancy across sentences included
  - Fluency: Readability as a standalone text
  - Brevity: Omission of information that is not critical to summary
- Similar to MT (BLEU score), researchers use easily-calculated metrics to approximate true evaluations, given a human-created summary (or set of summaries)
  - ROUGE-N: Precision, Recall and F-measures for ngram overlap between system and reference summaries
  - Slightly different from MT/BLEU – brevity penalty does not apply

Summarization: related tasks

- Text Simplification
  - Produce a version of a text with simpler, more accessible language
  - Cf., e.g., “Simple Wikipedia”
  - Useful applications: educational support, simplifying legal jargon
- Multi-document summarization
  - Given a *set* of documents about the same document/event, produce a summary that incorporates key information from all of them



Standard deviation

From Wikipedia, the free encyclopedia

Simple English Wikipedia

Standard deviation is a number used to tell how measurements for a group are spread out from the average (mean), or expected value. A low standard deviation means that most of the numbers are close to the average. A high standard deviation means that the numbers are more spread out.[1][2]

Standard deviation

From Wikipedia, the free encyclopedia

Wikipedia The Free Encyclopedia

For other uses, see Standard deviation (disambiguation).

In statistics, the **standard deviation** is a measure of the amount of variation or dispersion of a set of values.[1] A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.