Text summarization

CSC401/2511 – Natural Language Computing – Spring 2026
Lecture 16 Ken Shi
University of Toronto

Text Summarization

Objective: return shortened version of text that includes its main points.

This includes:

- "gisting": just a few words — almost topic classification
- abstracting, e.g., in MS Word
- longer summaries, e.g., 20% of original document size)
- original length (from multiple documents)

Kinds of Summaries

- Text vs. template
- **Perspective**: informative vs. indicative
- **Composition**: extract vs. abstract
- **Orientation**: document vs. query
- **Source**: single vs. multiple document
- **Background**: complete vs. update

Indicative vs. Informative Summaries

Overview summary of Angina

You are at: Angina

Search: ○ all documents ○ within Angina

Get more detailed information on the sections: [variant angina: | what is the treatment? | diagnosis | signs and symptoms | what are the symptoms | treatment]

Synopsis: Treatment is designed to prevent or reduce ischemia and minimize symptoms. Angina that cannot be controlled by drugs and lifestyle changes may require surgery. Angina attacks usually last for only a few minutes, and most can be relieved by rest. Most often, the discomfort occurs after strenuous physical activity or an emotional upset. A doctor diagnoses angina largely by a person's description of the symptoms. The underlying cause of angina requires careful medical treatment to prevent a heart attack. Not everyone with ischemia experiences angina. If you experience angina, try to stop the activity that precipitated the attack.

Highlighted differences between the documents:

- This file (5 minute emergency medicine consult) is close in content to the summary.
- More information on additional topics which are not included in the summary are available in these files (The American Medical Association family medical guide and The Columbia University College of Physicians and Surgeons complete home medical guide). The topics include "definition" and "what are the risks?"
- The Merck manual of medical information contains extensive information on the topic.

Summarization by Extraction

Identify important information, and drop it into summary.

How do we determine importance?

- Position in text, e.g.:
  - first sentence of each paragraph
  - first and last paragraphs of document
  - section headings, captions, etc.
  - varies with genre
  - Hovy-Lin (partial) ordering:
    - WSJ: $T > P1S1 > P1S2 > \cdots$
    - Ziff-Davis: $T > P2S1 > P3S1 > P2S2 > \{P4S1, P5S1, P3S2\} > \cdots$

Summarization by Extraction

Identify important information, and drop it into summary.

How do we determine importance?

- *Indicators*
  - *cues*, e.g.:
    - "in this paper, we show"
    - "in conclusion"
    - "recommend that"
  - *clues* (*bonus words*), e.g.:
    - "significantly"
    - "this paper"
  - *stigma words*, e.g.:
    - "hardly"
    - "incidentally"
    - "supported by a grant"

Summarization by Extraction

Identify important information, and drop it into summary.

How do we determine importance?

- Position in text
- *Indicators*
  - *cues*
  - *clues (bonus words)*
  - *stigma words*
  - content words from title
  - **not** tf.idf

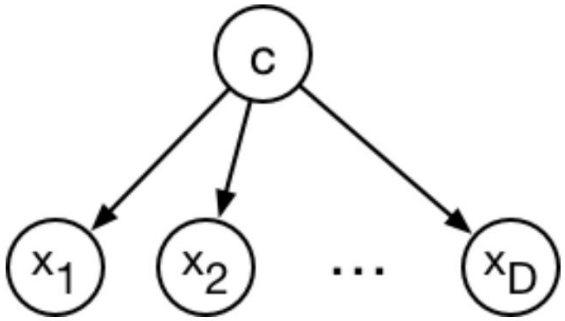
Naïve Bayes and SoftMax

- Broadly, Bayesian probability conceives of probability *not* as frequency of some phenomenon occurring, but rather as an expectation related to our own certainty.
- Given an observation x, **Naïve Bayes** simply chooses the class $c \in C$ that maximizes $\boldsymbol{P(c \mid x)}$.
  - This can be done in many ways.

$$\underset{c}{\operatorname{argmax}} P(c|x) = \frac{P(c)}{\cancel{P(x)}} P(x|c)$$

Estimate the $P(\cdot)$ using Gaussians, or…

Bayesian Classifier



Given features $\mathbf{x} = [x_1, x_2, \cdots, x_D]^T$
want to compute class probabilities using Bayes Rule:

$$\underbrace{p(c|\mathbf{x})}_{\text{Pr. class given feature}} = \frac{\overbrace{p(\mathbf{x}|c)}^{\text{Pr. feature given class}} \quad p(c)}{p(\mathbf{x})}$$

In words,

$$\text{Posterior for class} = \frac{\text{Pr. of feature given class} \times \text{Prior for class}}{\text{Pr. of feature}}$$

To compute $p(c|\mathbf{x})$ we need: $p(\mathbf{x}|c)$ and $p(c)$.

From Technology Upskilling ML Software Foundations by Juhan Bae and En-Shiun Annie Lee

Independence Assumption

- Naive assumption: The features x_i are conditionally independent given the class c.
- Allows us to decompose the joint distribution:

$$p(c, x_1, \ldots, x_D) = p(c)\, p(x_1|c) \cdots p(x_D|c).$$

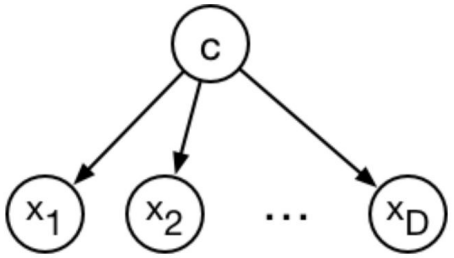
- Compact representation of the joint distribution.
  - Prior probability of class:

$$p(c = 1) = \pi$$

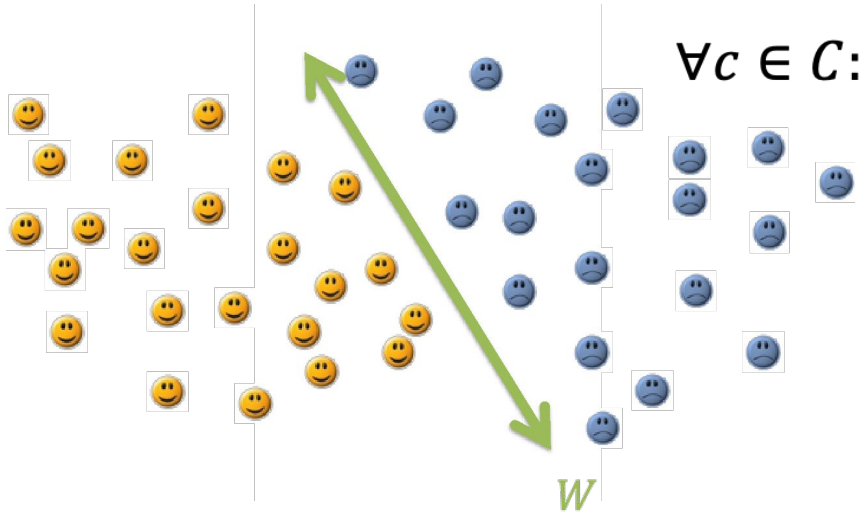
  - Conditional probability of feature given class:

$$p(x_j = 1|c) = \theta_{jc}$$

Naïve Bayes and SoftMax



- Assume $x \in \mathbb{R}^d$, learning a linear decision boundary is tantamount to learning $W \in \mathbb{R}^{C \times d}$.



P(Class|features) = P(features|Class)*P(Class)

$$\forall c \in C: \boldsymbol{f_c} = W[c, \cdots] \cdot x = \sum_{i=1}^{d} W[c, i] \cdot x[i]$$

Uh oh – $\boldsymbol{f_c}$ can be negative and we want something on $[0,1]$, to be a probability.

Solution: Just raise it with an exponent

Softmax:

$$P(y|x) = \frac{\exp(f_y)}{\sum_{c=1}^{C} \exp(f_c)}$$

Naive Bayes: https://www.youtube.com/watch?v=O2L2Uv9pdDA
SoftMax: https://www.youtube.com/watch?v=8ps_JEW42xs
Example on Text: https://www.youtube.com/watch?v=temQ8mHpe3k
Naive Bayes on Spam: https://youtu.be/M59h7CFUwPU
Why Naive Bayes are Cool: https://www.youtube.com/watch?v=8NEfN3JbINA

Naive Bayes Properties

- An amazingly cheap learning algorithm!
- **Training time**: Estimate parameters using maximum likelihood.
  - Compute co-occurrence counts of each feature with the labels.
    I Requires only one pass through the data!
- **Test time**: Apply Bayes' Rule.
  - Cheap because of the model structure. For more general models, Bayesian inference can be very expensive and/or complicated.
- Analysis easily extends to prob. distributions other than Bernoulli.
- Less accurate in practice compared to discriminative models due to its "naive" independence assumption.

Naive Bayes Classification

We can treat summarization as a *sequence* of *binary* classification problems: every sentence is either *in* or *out*.

Bayes decision rule: choose outcome that is most probable in given context of features:

$$\max\{\ P(s \in \text{Summary}|f_1 \ldots f_k),\ P(s \notin \text{Summary}|f_1 \ldots f_k)\}$$

$P(o|f_1 \ldots f_k)$ is hard to measure, so we use Bayes's rule:

$$P(o|f_1 \ldots f_k) = \text{what?}$$

Naive Bayes Classification

$P(o|f_1 \dots f_k)$ is hard to measure, so we use Bayes's rule:

$$P(o|f_1 \dots f_k) = \frac{P(f_1 \dots f_k|o)P(o)}{P(f_1 \dots f_k)}$$

The *Naive Bayes Assumption*: all features of context are conditionally independent. Thus:

$$P(f_1 \dots f_k|o) \doteq \prod_{1 \leq j \leq k} P(f_j|o)$$

And we can use relative frequency in annotated corpora for these:

$$P(f_j|o) = \frac{C(f_j, o)}{C(o)}$$

Disadvantages of Summ. by Extraction

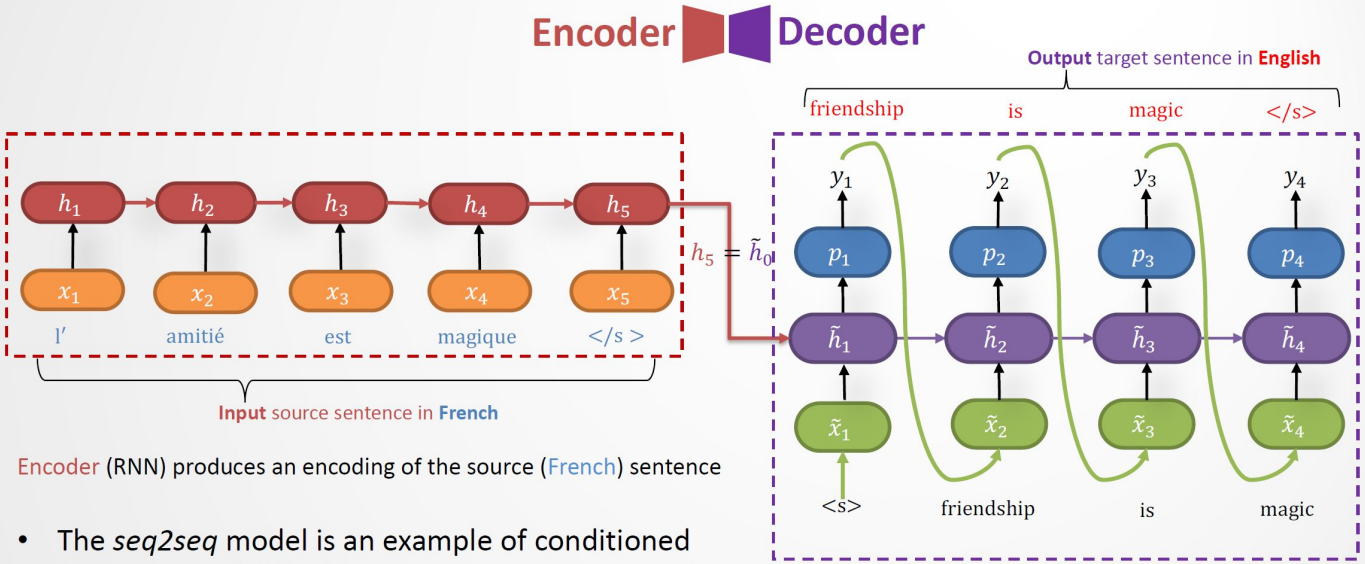
- Hard to read, misleading, and/or incoherent, e.g.:
  - lost pronoun antecedents
  - discourse/argument connectives no longer appropriate
- Parts of extracted sentences may be unimportant
  - negation (of clues and stigma words)
  - granularity of sentence-sized extracts

Improvements on Summ. by Extraction

- Abstraction
- Use argument structure (???) to determine importance and conservatively synthesize new text
- Summarize multiple documents/background collection and use comparisions to boost confidence in importance.
- Task-based evaluation: determine how well summaries work in context. How do people use summaries?

Remember this?

NMT: the seq2seq model



Encoder (RNN) produces an encoding of the source (French) sentence

- The *seq2seq* model is an example of conditioned language model (LM)
- Many variants exists. The classical (vanilla) seq2seq model outlined here
- NMT directly calculates $y^* = \text{argmax}_y P(y|x)$
- I.e. with our formulation:

$$E^* = \text{argmax}_E P(E|F)$$

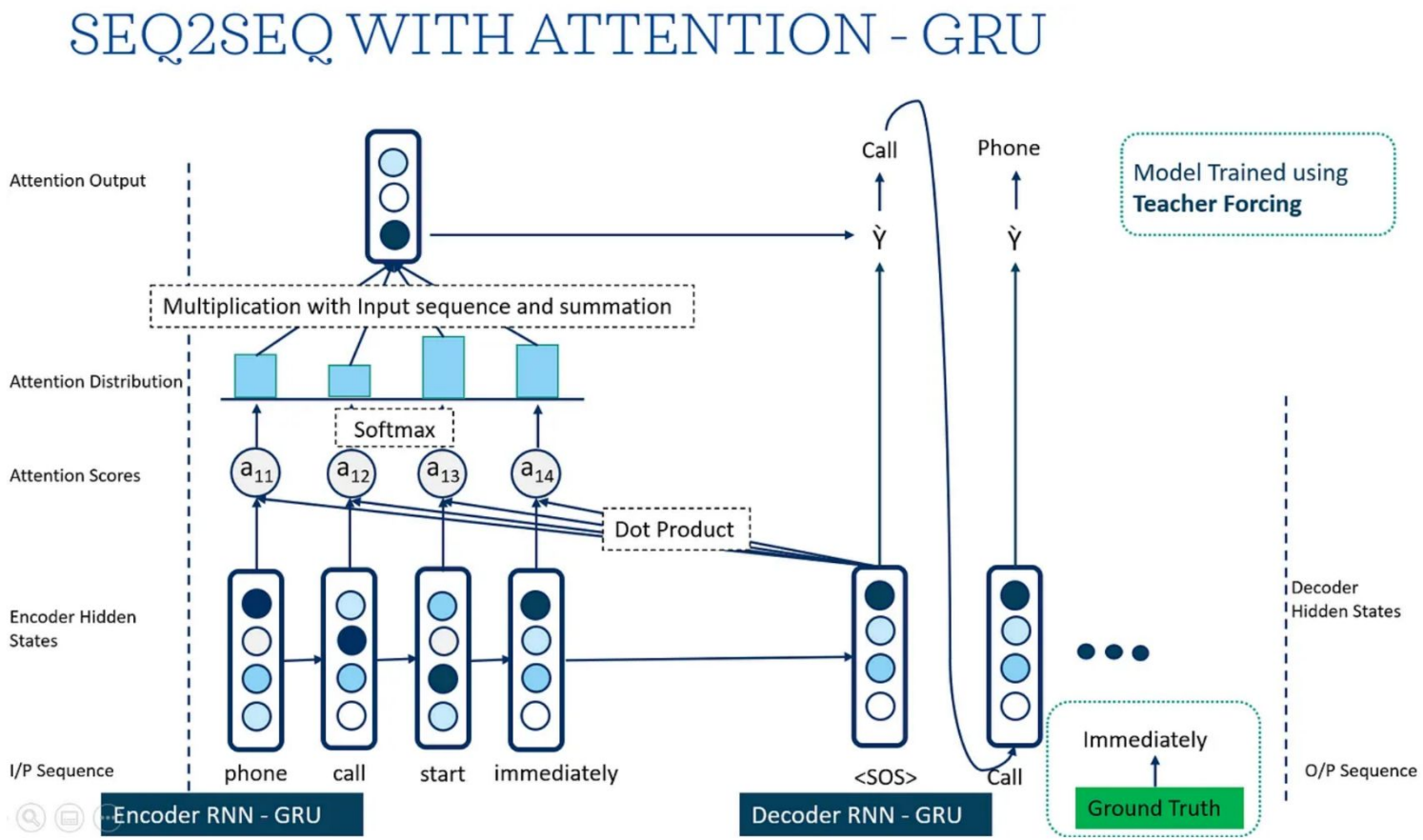
Decoder (RNN) generates target sentence (in English), conditioned on the encoding

Decoder is predicting the next word of the target sentence y

Prediction is **conditioned** on the source sentence **x**

$$P(y|x) = P(y_1|x)P(y_2|y_1, x) \dots P(y_T|y_1, \dots y_{(T-1)}, x)$$

Attention for Summarization



(towardsdatascience.com)

Is Document Summarization About Centrality?

The trial for one of two men accused in the beating death of University of Wyoming student Matthew Shepard will begin with jury selection March 24.

Defendant **Charges**

Shepard died following a brutal beating near Laramie in October that police said was motivated in part because he was gay.

Why

Officials said about 30 people would be questioned.

Investigator

Henderson and the 21-year-old McKinney both face the death penalty if convicted in a case that has become a central focus for gay rights activists and others seeking stronger bias crime legislation.

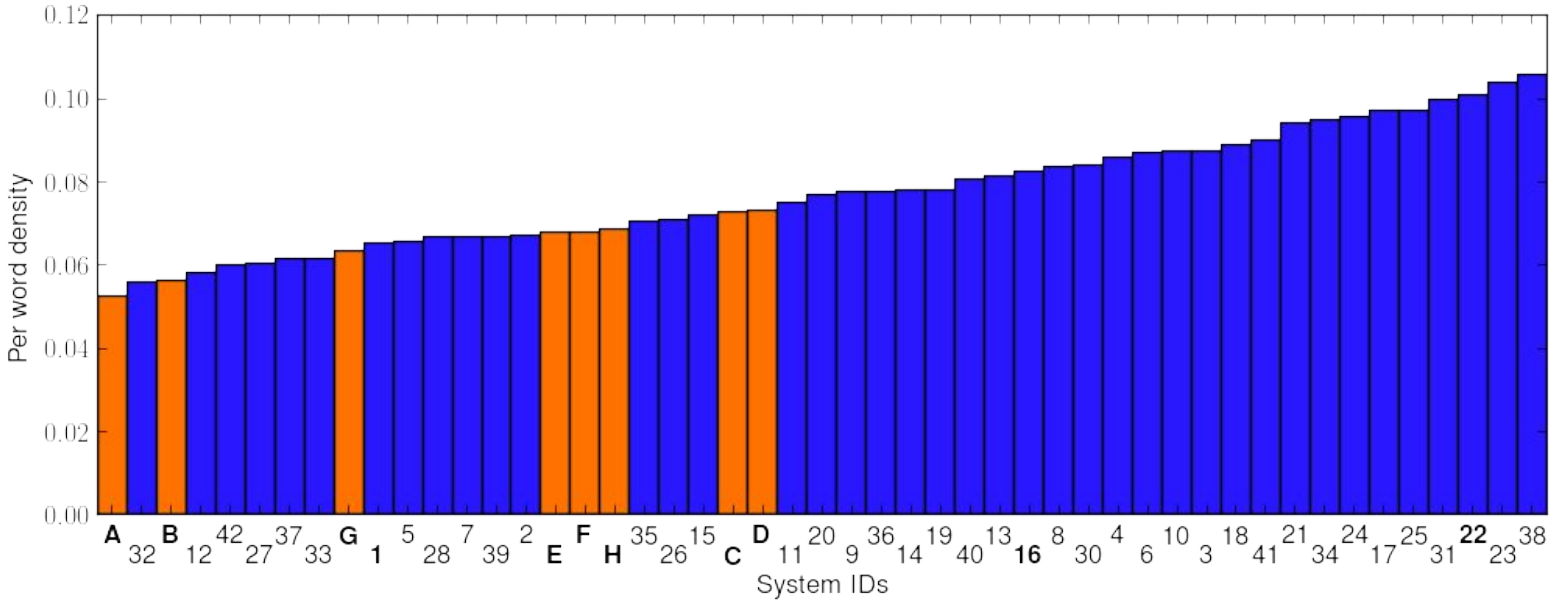
Sentence

"There is no guarantee that these laws will stop hate crimes from happening."

Broader Consequences

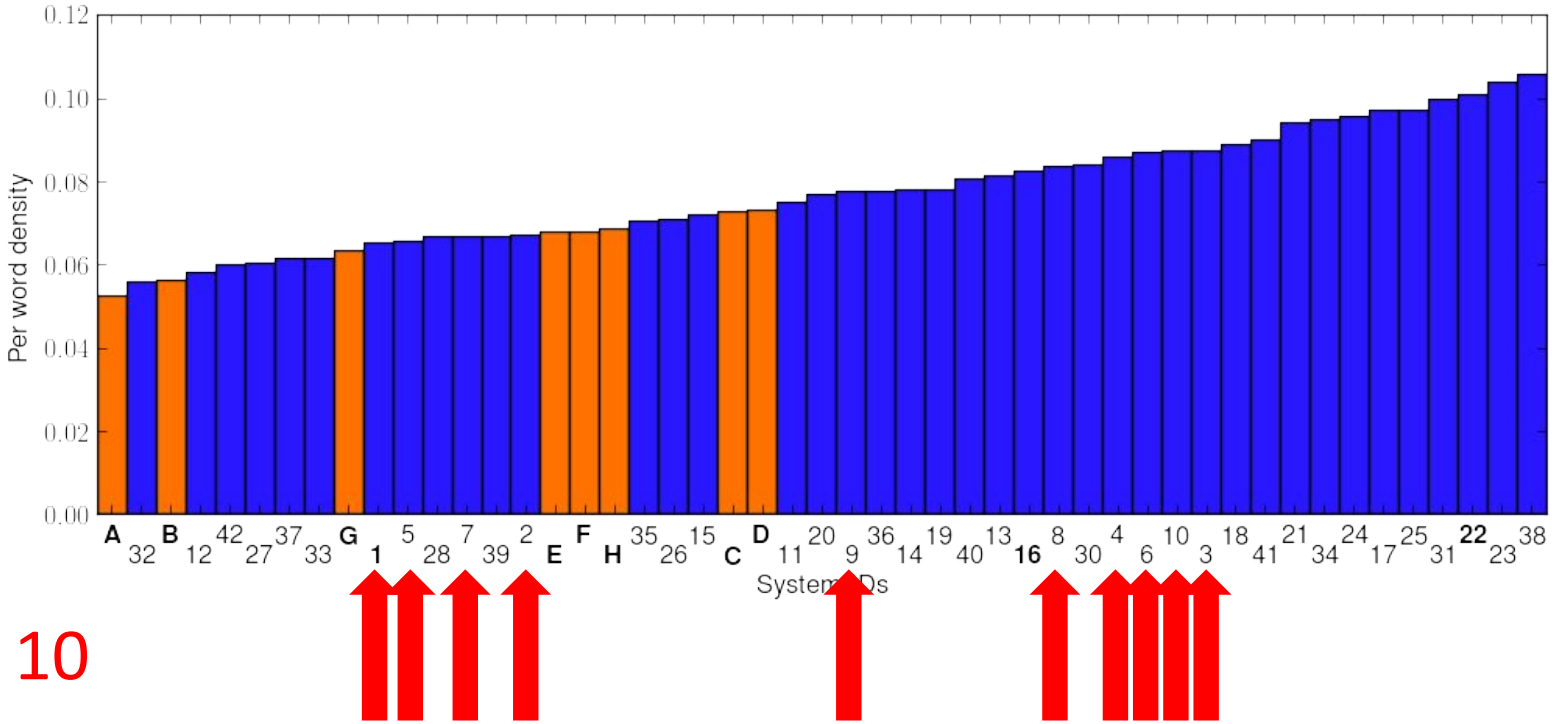
Why Look Beyond Core Sentences

- Automatic summaries *already more central* than human-written gold standard summaries
  - **Signature caseframe density**: a measure of centrality based on method of Lin and Hovy, (2000)



Why Look Beyond Core Sentences

- Automatic summaries *already more central* than human-written gold standard summaries
  - **Signature caseframe density**: a measure of centrality based on method of Lin and Hovy, (2000)



UNSUPERVISED SENTENCE ENHANCEMENT FOR AUTOMATIC SUMMARIZATION

Jackie Chi Kit Cheung and Gerald Penn

University of Toronto

27th October, 2014

Text-to-text Generation

- **Sentence compression** (e.g., Knight and Marcu, 2000)

  Bil Mar Foods Co., a meat processor owned by Sara Lee, announced a recall of certain lots of hot dogs and packaged meat.

- **Sentence fusion** – merge parts *of similar sentences*

  (Barzilay and McKeown, 2005; Filippova and Strube, 2008)

Bil Mar Foods Co. announced a recall of certain lots of hot dogs and packaged meat.

The outbreak led to the recall on Tuesday of certain lots of hot dogs and packaged meat produced at the Bil Mar Foods plant.

The outbreak led to the recall on Tuesday of meats produced at the Bil Mar Foods plant.

Sentence Enhancement

- Traditional fusion: align similar *core* sentences

  Bil Mar Foods Co. announced a recall of certain lots of hot dogs and packaged meat.

  The outbreak led to the recall on Tuesday of meats produced at the Bil Mar Foods plant.

- Expand with parts of dissimilar in-domain sentences

  This fact has been underscored in the last few months by two unexpected outbreaks of food-borne illness.

- Output:

  The outbreak of food-borne illness led to the recall on Tuesday of certain lots of hot dogs and packaged meat produced at the Bil Mar Foods plant.

Why Sentence Enhancement?

Still a big gap between text-to-text generation methods and what concept-to-text generation aspires to attain

- Locality of context
- Inference
  - Sentence fusion attractive because it doesn't require deep semantic analysis.
  - *Can we do more with equally little?*

Natural Language Inference

- **Understand paraphrases**

  Gained ten pounds

  A ten-pound weight gain

- **Determine redundant sentences**

  Search crews determined the source of the fire which damaged five homes.

  The fire wrecked five homes.

Why Look Beyond Core Sentences? A Case Study of the TAC 2010 Corpus

- Where do the semantic predicates in human-written summaries come from?
  - **77%** are found in the source text
  - Additional **21%** found in in-domain articles
    - 98% of total found!
  - cf., additional **14%** found if adding irrelevant articles
  - **51%** of source-external sentences use source-internal predicates (often with different arguments)

Why do we look outside the source?

- Source-external predicates are of lower average frequency (95% confidence intervals):

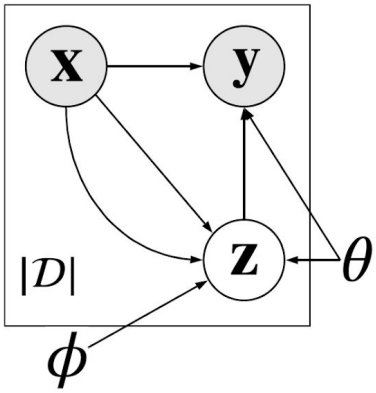
	Average freq (millions)
Source-internal	1.77 (1.57, 2.08)
Source-external	1.15 (0.99, 1.50)

(Wilcoxon ranked-sums $p < 10^{-17}$)

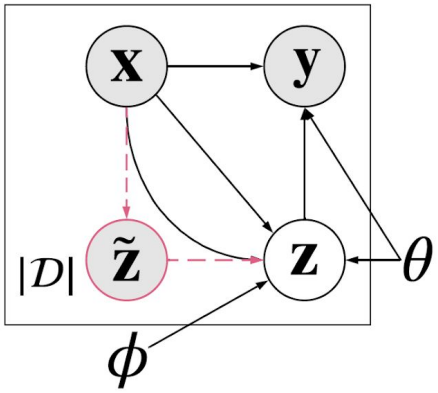
- Source-external predicates have lower average argument entropy (95% confidence intervals):

	Arg entropy
Source-internal	7.94 (7.90, 7.97)
Source-external	7.42 (7.37, 7.48)

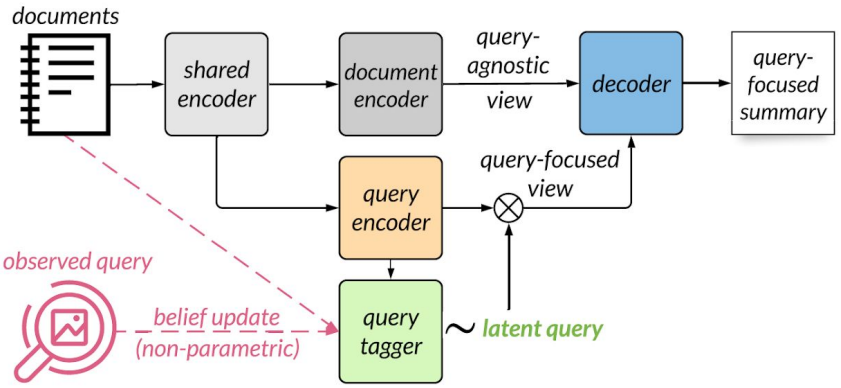
Latent-Query Document Summarization



(a) Generative Process: Training



(b) Generative Process: Testing



(c) Neural Parameterization

(Xu and Lapata, 2022)

$$\mathcal{L} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})\right]}_{\text{conditional language modeling}} + \underbrace{\beta\mathcal{H}\left(q_\phi(\mathbf{z}|\mathbf{x})\right) - \omega\mathcal{H}\left(o(\hat{\mathbf{z}}|\mathbf{x},\mathbf{y}), q_\phi(\mathbf{z}|\mathbf{x})\right)}_{\text{latent query modeling}} \quad (6)$$

where $\mathcal{H}(\cdot)$ denotes posterior entropy and $\mathcal{H}(\cdot,\cdot)$ denotes cross entropy.

Latent-Query Document Summarization

	DUC 2006			DUC 2007			TD-QFS		
Upper Bound & Baselines	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
GOLD	45.4	11.2	16.8	47.5	14.0	18.9	52.2	27.0	30.2
ORACLE	47.5	15.8	20.2	47.6	17.1	20.9	64.9	48.3	49.4
LEAD	32.1	5.3	10.4	33.4	6.5	11.3	33.5	5.2	10.4
LEXRANK$_Q$	34.2	6.4	11.4	35.8	7.7	12.7	35.3	7.6	12.2
Distantly Supervised	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
QUERYSUM* (Xu and Lapata, 2020)	41.6	9.5	15.3	43.3	11.6	16.8	44.3	16.1	20.7
BART-CAQ (Su et al., 2020)	38.3	7.7	12.9	40.5	9.2	14.4	—	—	—
PQSUM (Laskar et al., 2020b)	40.9	9.4	14.8	42.2	10.8	16.0	—	—	—
Few- or Zero-shot Abstractive	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
MARGESUM† (Xu and Lapata, 2021)	40.2	9.7	15.1	42.5	12.0	16.9	45.5	16.6	20.9
BART (Lewis et al., 2020)	38.3	7.8	13.1	40.2	9.9	14.6	45.1	16.9	21.4
GSUM+LEXRANK$_Q$	38.1	7.9	13.1	39.5	9.5	14.3	45.5	18.0	**22.4**
LQSUM	**39.1**	**8.5**	**13.7**	**40.4**	**10.2**	**15.0**	**45.7**	**18.1**	22.1

Table 7: Multi-document QFS, zero-shot setting, **DUC** (queries are narratives) and **TD-QFS** (queries are keywords) test sets. $*/\dagger$ denotes extractive/few-shot systems.

	CNN/DM			WikiRef			Debatepedia			DUC 2006			DUC 2007			TD-QFS		
Model	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
LQSUM	45.1	22.0	41.9	31.1	12.6	27.1	23.5	7.2	20.6	39.1	8.5	13.7	40.4	10.2	15.0	45.7	18.1	22.1
$-\Delta(\hat{\mathbf{z}}\|\mathbf{x},\mathbf{z})$	—	—	—	↓0.1	↓0.2	↓0.2	↓0.5	↓0.3	↓0.6	↓0.6	↓0.2	↓0.6	↑0.1	↓0.1	↓1.3	↑0.1	↓0.6	↓0.4
–Joint training	↓0.4	↓0.3	↓0.4	↓2.9	↓0.9	↓2.8	↓2.8	↓1.1	↓2.8	↓2.9	↓1.7	↓1.6	↓2.4	↓2.0	↓1.7	↓0.7	↓0.6	↓0.4
–Weak supervision	↓0.6	↓0.7	↓0.7	↓0.7	↓0.2	↓0.5	↓1.0	↓0.5	↓1.3	↓0.2	↓0.2	↓0.2	↓0.2	↓0.3	↓0.3	↓0.1	↓0.3	↓0.0
–Dual view	↓2.7	↓3.5	↓2.5	↓12.2	↓9.3	↓10.5	↓7.9	↓3.3	↓6.6	↓6.3	↓1.8	↓1.8	↓6.5	↓3.0	↓2.5	↓2.5	↓3.3	↓2.8
–Posterior dropout	↓0.7	↓0.6	↓0.8	↓0.8	↓0.3	↓0.7	↓1.1	↓0.3	↓1.2	↓0.2	↓0.2	↓0.2	↓0.4	↓0.4	↓0.5	↑0.2	↓0.0	↑0.1