# Text Summarization

# Announcements

- Reading: summarization paper for today
  - Monday: language generation paper
- Wed, Nov 20th: Or Biran, Elementary Cognition: Dialog systems.
- Monday, Nov 25th: Bias

# HW4 comments

- Out today to give you a full 3 weeks
- Due Dec. 4th
- Language generation:
  - Restaurant review database
  - Task: generate reviews
  - Includes beam search which we will cover today

# Midterm Curve

- Statistics
  - Median: 71
  - Mean: 69.5
- A+ >-90
- A <90 and >78
- A- <=78 and >76
- B+ <=76 and >74
- B <=74 and >66
- B- <=66 and > 63
- C+ <=63 and >62
- C <=62 and >41
- C- <=41 and >38
- D <=38

# Today

Abstractive summarization: headline generation

Extractive summarization of news articles

A state of the art model

# Moving to neural summarization

- Data!
  - DUC 2000 – 2007
    - Typically 30-50 input document sets with 2-7 summaries per set
- Need on the order of 100,000 – 1,000,000 article/summary pairs
- Constrains the tasks

# A Neural Model

- A Neural Attention Model for Abstraction Summarization Alexander Rush, Sumit Chopra, Jason Weston

# Task: Headline Generation

- Input: first sentence of a news article
- Output:  headline
- Attention Based Summarizer (ABS): Performs abstractive summarization
  - Paraphrasing, compression

# Training Data

- 4 million article/headline pairs from Gigaword
  - A detained iranian-american academic accused of acting out against national security has been released from a tehran prison yesterday after a hefty bail was posted, a top judiciary official said Tuesday.
    Detained iranian-american academic released from prison after hefty bail.
  - Ministers from the european union and its mediterranean neighbors gathered here under heavy security on Monday for an unprecedented conference on economic and political cooperation.
    European mediterranean ministers gather for landmark conference by julie bradford.

# System Output

- russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism.
- russia calls for joint front against terrorism.

# Problem Framework

- Input: $x_1 \ldots x_m$
- Output: $y$ of length $n<m$
  - Output length n fixed and known by system
- Fixed vocabulary $V$ of size $|V|$ and both input and output come from $V$
- Scoring function $S(x,y) = \log P(y|x;\Theta)$ equivalent to:

$$\log p(\mathbf{y}|\mathbf{x};\theta) \approx \sum_{i=0}^{N-1} \log p(\mathbf{y}_{i+1}|\mathbf{x},\mathbf{y}_c;\theta)$$

# Main Focus

- Modeling the local conditional distribution:
  - $P(y_{i+1} | x_i, y_c; \Theta)$
- Neural network is a seq-to-seq model
  - Encoder: a conditional summarization model
  - Decoder: Neural language probabilistic model

# Neural Language Model

- Standard Neural Network Language Model (NNLM) (Banko et al 2003)

$$
\begin{aligned}
p(\mathbf{y}_{i+1}|\mathbf{y}_c, \mathbf{x}; \theta) &\propto \exp(\mathbf{V}\mathbf{h} + \mathbf{W}\mathrm{enc}(\mathbf{x}, \mathbf{y}_c)), \\
\tilde{\mathbf{y}}_c &= [\mathbf{E}\mathbf{y}_{i-C+1}, \ldots, \mathbf{E}\mathbf{y}_i], \\
\mathbf{h} &= \tanh(\mathbf{U}\tilde{\mathbf{y}}_c).
\end{aligned}
$$

- Parameters: $\theta = (\mathbf{E}, \mathbf{U}, \mathbf{V}, \mathbf{W})$
- E is a word embedding matrix
- U,V,W are weight matrices

# Experiments with 3 encoders

1. Bag of word encoder
   - Bag of words of input sentence embedded down to H, size of hidden layer
   - Ignores order and context in input
   - Captures relative importance of words

$$\begin{aligned} \mathrm{enc}_1(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^\top \tilde{\mathbf{x}}, \\ \mathbf{p} &= [1/M, \ldots, 1/M], \\ \tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M]. \end{aligned}$$

# Experiments with 3 Encoders

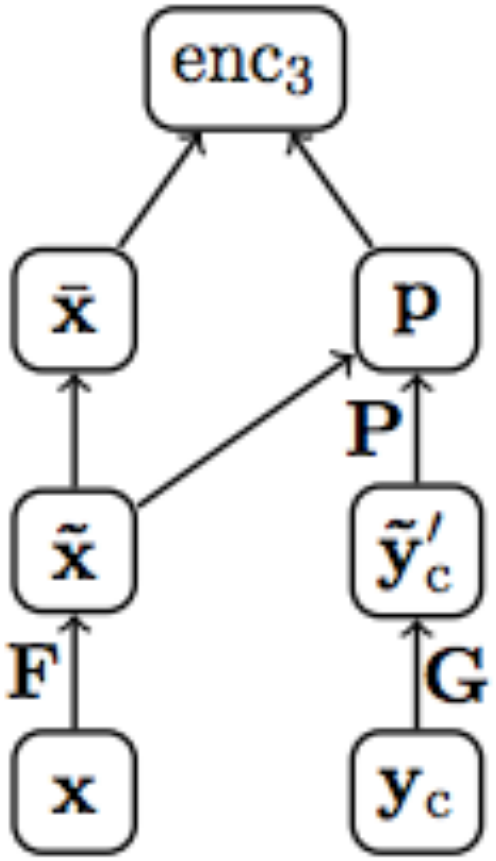
2. Convolutional encoder
   - Allows for local interactions in the input
     - CNNs good at capturing patterns such as n-grams
   - Does not require encoding input context $y_c$
   - But it must produce a single representation for the entire sentence

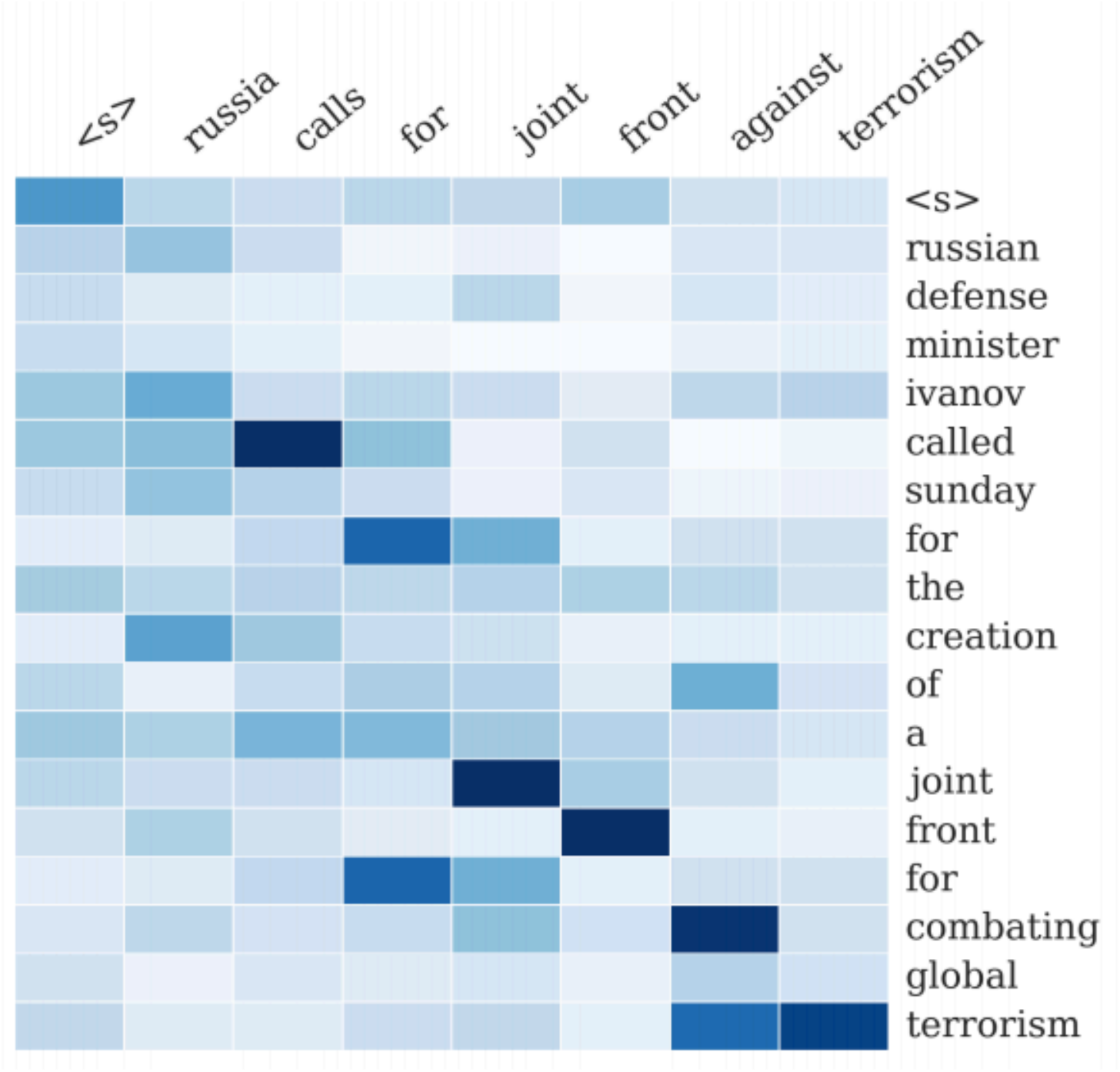
# Experiments with 3 Encoders

- Attention-based encoder
  - Constructs a representation using generation context ($y_c$)
  - Replaces uniform distribution of bag of words with a learned soft alignment between input and summary
    - Used to weight the smoothed input when construction the representation

# Attention based encoder

$$
\begin{aligned}
\mathrm{enc}_3(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^\top \bar{\mathbf{x}}, \\
\mathbf{p} &\propto \exp(\tilde{\mathbf{x}} \mathbf{P} \tilde{\mathbf{y}}'_c), \\
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\tilde{\mathbf{y}}'_c &= [\mathbf{G}\mathbf{y}_{i-C+1}, \ldots, \mathbf{G}\mathbf{y}_i], \\
\forall i \quad \bar{\mathbf{x}}_i &= \sum_{q=i-Q}^{i+Q} \tilde{\mathbf{x}}_i / Q.
\end{aligned}
$$




<s>
russia
calls
for
joint
front
against
terrorism
<s>
russian
defense
minister
ivanov
called
sunday
for
the
creation
of
a
joint
front
for
combating
global
terrorism

# Interesting extensions

- Uses a beam search decoder
  - Maintains full vocabulary
  - Limits to K potential hypotheses at each position of the summary
- Add features to promote using words of input (extractive features)
  - Combine the local conditional probability with indicator features for n-grams from input

# Adding in features to encourage using input words

$$s(\mathbf{y}, \mathbf{x}) = \sum_{i=0}^{N-1} \alpha^\top f(\mathbf{y}_{i+1}, \mathbf{x}, \mathbf{y}_c).$$

$$f(\mathbf{y}_{i+1}, \mathbf{x}, \mathbf{y}_c) = [\, \log p(\mathbf{y}_{i+1} | \mathbf{x}, \mathbf{y}_c; \theta),$$
$$\mathbb{1}\{\exists j.\, \mathbf{y}_{i+1} = \mathbf{x}_j\},$$
$$\mathbb{1}\{\exists j.\, \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0, 1\}\},$$
$$\mathbb{1}\{\exists j.\, \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0, 1, 2\}\},$$
$$\mathbb{1}\{\exists k > j.\, \mathbf{y}_i = \mathbf{x}_k, \mathbf{y}_{i+1} = \mathbf{x}_j\}\, ].$$

# Beam search decoder

**Algorithm 1** Beam Search

**Input:** Parameters $\theta$, beam size $K$, input $\mathbf{x}$
**Output:** Approx. $K$-best summaries

$\pi[0] \leftarrow \{\epsilon\}$
$\mathcal{S} = \mathcal{V}$ if abstractive else $\{\mathbf{x}_i \mid \forall i\}$
**for** $i = 0$ to $N - 1$ **do**

▷ Generate Hypotheses

$$\mathcal{N} \leftarrow \{[\mathbf{y}, \mathbf{y}_{i+1}] \quad | \quad \mathbf{y} \in \pi[i], \mathbf{y}_{i+1} \in \mathcal{S}\}$$

▷ Hypothesis Recombination

$$\mathcal{H} \leftarrow \left\{ \mathbf{y} \in \mathcal{N} \quad | \quad \begin{array}{c} s(\mathbf{y}, \mathbf{x}) > s(\mathbf{y}', \mathbf{x}) \\ \forall \mathbf{y}' \in \mathcal{N} \text{ s.t. } \mathbf{y}_c = \mathbf{y}'_c \end{array} \right\}$$

▷ Filter K-Max

$$\pi[i+1] \leftarrow \text{K-arg}\max_{\mathbf{y} \in \mathcal{H}} g(\mathbf{y}_{i+1}, \mathbf{y}_c, \mathbf{x}) + s(\mathbf{y}, \mathbf{x})$$

**end for**
**return** $\pi[N]$

# Example

- Input: Kathy has an adorable but sometimes feisty bulldog named Roscoe
- Summary: Kathy's feisty bulldog Roscoe.

# Experiments

- Rouge-1, -2 and –L as metrics
- On headlines:
  - 17 point jump in Rouge-1 from worst baseline (IR)
  - 11 point jump in Rouge 1 from compressive system
- On DUC-2004, 17 and 10 point jumps

| Model | DUC-2004 ROUGE-1 | DUC-2004 ROUGE-2 | DUC-2004 ROUGE-L | Gigaword ROUGE-1 | Gigaword ROUGE-2 | Gigaword ROUGE-L | Ext. % |
|---|---|---|---|---|---|---|---|
| IR | 11.06 | 1.67 | 9.67 | 16.91 | 5.55 | 15.58 | 29.2 |
| PREFIX | 22.43 | 6.49 | 19.65 | 23.14 | 8.25 | 21.73 | 100 |
| COMPRESS | 19.77 | 4.02 | 17.30 | 19.63 | 5.13 | 18.28 | 100 |
| W&L | 22 | 6 | 17 | - | - | - | - |
| TOPIARY | 25.12 | 6.46 | 20.12 | - | - | - | - |
| MOSES+ | 26.50 | 8.13 | 22.85 | 28.77 | 12.10 | 26.44 | 70.5 |
| ABS | 26.55 | 7.06 | 22.05 | 30.88 | 12.22 | 27.77 | 85.4 |
| ABS+ | 28.18 | 8.49 | 23.81 | 31.00 | 12.65 | 28.34 | 91.5 |
| REFERENCE | 29.21 | 8.38 | 24.46 | - | - | - | 45.6 |

Table 1: Experimental results on the main summary tasks on various ROUGE metrics . Baseline models are described in detail in Section 7.2. We report the percentage of tokens in the summary that also appear in the input for Gigaword as `Ext %`.

BASELINES:

IR: index training set and retrieve the title with best match to input

COMPRESS: Clarke&Lapata (learning using syntactic structures and language model)

TOPIARY: Best system on DUC 2004 (linguistic rules)

PREFIX: first 75 characters

MOSES: MT approach

# Example output

- Input: a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted, a top judiciary offical said Tuesday.
- G: iranian-american academic held in tehran released on bail
- A: detained iranian-american academic released from jail after posting bail.
- A+ detained iranian-american academic released from prison after hefty bail

# Syntactically incorrect

- I: the white house on Thursday warned iran of possible new sanctions after the un nuclear watchdog reported that tehran had begun sensitive nuclear work at a key site in defiance of un resolutions
- G: us warns iran of step backward on nuclear issue
- A: iran warns of possible new sanctions on nuclear work
- A+: un nuclear watchdog warns iran of possible new sanctions

# ABS is more creative

- Input: australian foreign minister stephen smit Sunday congratulated new zealand's new prime minister-elect john key as he praised ousted leader helen clark as a "gutsy" and respected politician.
- G: time caught up with nz's gutsy clark says australian fm
- A: australian foreign minister congratulates new nz pm after election
- A+ australian foreign minister congratulates new zealand as leader

# What other similar task?

- Could we use Roscoe? (in image and in print)

# Another neural summarization approach

- Extractive summarization of news
  - Single document summarization
- Data source: Daily News
  - Bulleted highlights of each article
- Neural Summarization by Extracting Sentences and Words
  - Cheng and Lapata, Edinburgh

# Example from Daily News

**AFL star blames vomiting cat for speeding**

Adelaide Crows defender Daniel Talia has kept his driving license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat.

The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car.

In the Adelaide magistrates court on Wednesday, Magistrate Bob Harrap fined Talia $824 for exceeding the speed limit by more than 30km/h.

He lost four demerit points, instead of seven, because of his significant training commitments.

- *Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car.*
- *22-year-old Talia was fined $824 and four demerit points, instead of seven, because of his 'significant' training commitments.*

Figure 1: DailyMail news article with highlights. Underlined sentences bear label 1, and 0 otherwise.

Cheng and Lapata 2016

# Example from Daily News

**AFL star blames vomiting cat for speeding**

Adelaide Crows defender Daniel Talia has kept his driving license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat.

The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car.

In the Adelaide magistrates court on Wednesday, Magistrate Bob Harrap fined Talia $824 for exceeding the speed limit by more than 30km/h.

He lost four demerit points, instead of seven, because of his significant training commitments.

- *Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car.*
- *22-year-old Talia was fined $824 and four demerit points, instead of seven, because of his 'significant' training commitments.*

Figure 1: DailyMail news article with highlights. Underlined sentences bear label 1, and 0 otherwise.

**Paraphrasing**
**Compression**
**Fusion**

Cheng and Lapata 2016

# Two Tasks

- Input: Document D: $\{s_1,\ldots s_m\}$ consisting of words $w_1,\ldots.w_n$
- Sentence extraction
  - Select a subset of j sentences, $j<m$
  - Score each sentence and predict label $y_L \in \{0,1\}$
  - Objective: Maximize all sentence labels given D and weights $\theta$
- Word extraction
  - Find a subset of words in D and their optimal ordering
  - Language generation task with output vocabulary restricted to input D vocabulary
  - Objective: Maximize the likelihood of generated sentences, further decomposed by considering conditional dependencies among their words

# Training Data

- Sentence extraction
  - Highlights are abstracts
  - Find the s in D that most closely matches a highlight sentence
    - Positive, unigram and bigram matches, #entities
  - 200K document/summary pairs, summary size = 30% document
- Word extraction
  - Retain highlights with all words from D
  - Find neighbors of words not in D and substitute
  - 170K document/summary pairss

# Neural Summarization Architecture

- Hierarchical document reader
  - Derive meaning representation of document from its constituent sentences
- Attention based hierarchical content extractor
- Encoder-decoder architecture

# Document Reader

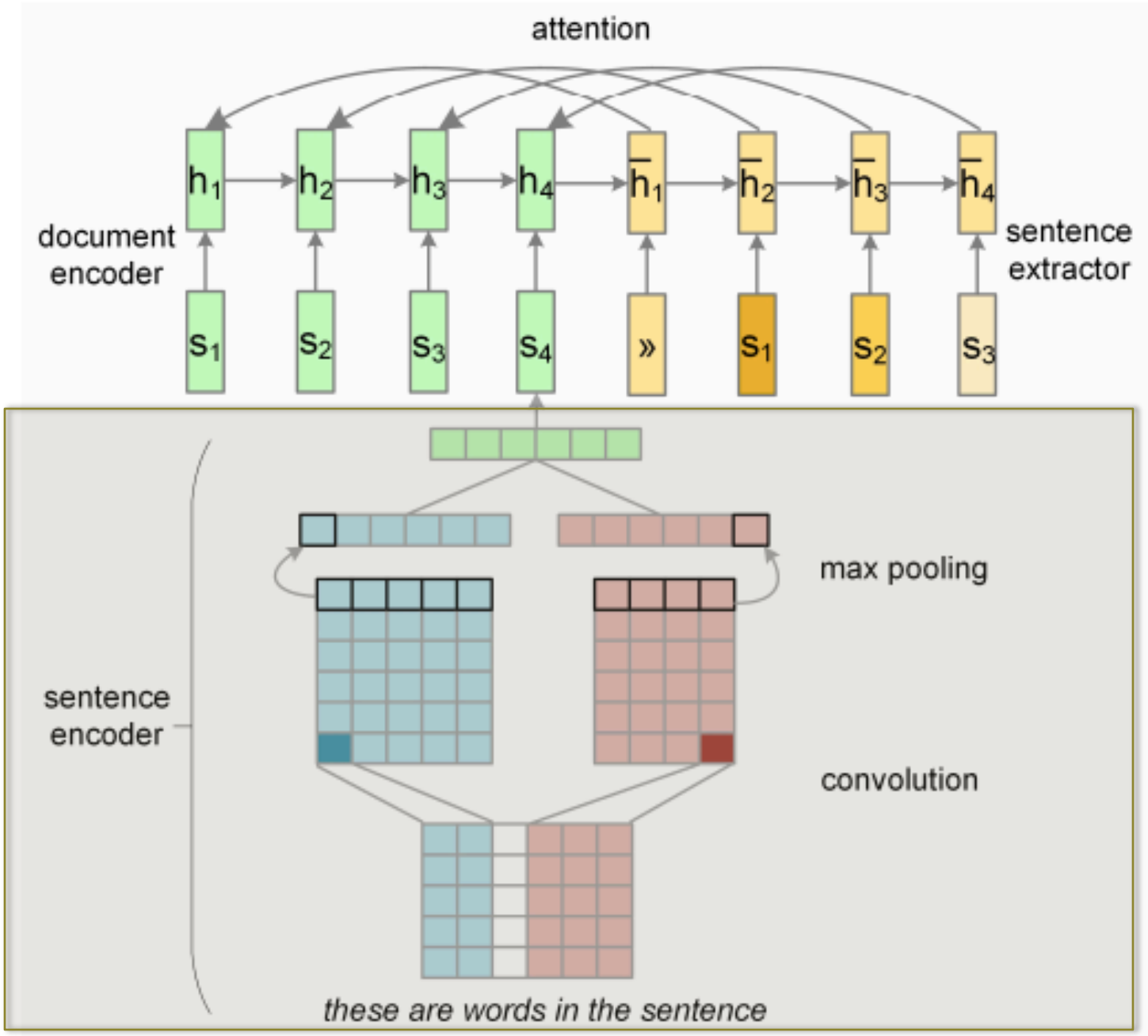
- CNN sentence encoder
  - Useful for sentence classification
  - Easy to train
- LSTM document encoder
  - Avoids vanishing gradients

# CNN

$$\mathbf{f}_j^i = \tanh(\mathbf{W}_{j:j+c-1} \otimes \mathbf{K} + b)$$

- Where W $\varepsilon \mathbb{R}^{nXd}$ and d = word embedding dimension, n = #words in sentence
- K a kernel of width c, b the bias
- $f^i_j$ = the jth item in the ith feature map $f^i$
- Perform max pooling over time to obtain a single feature to represent the sentence

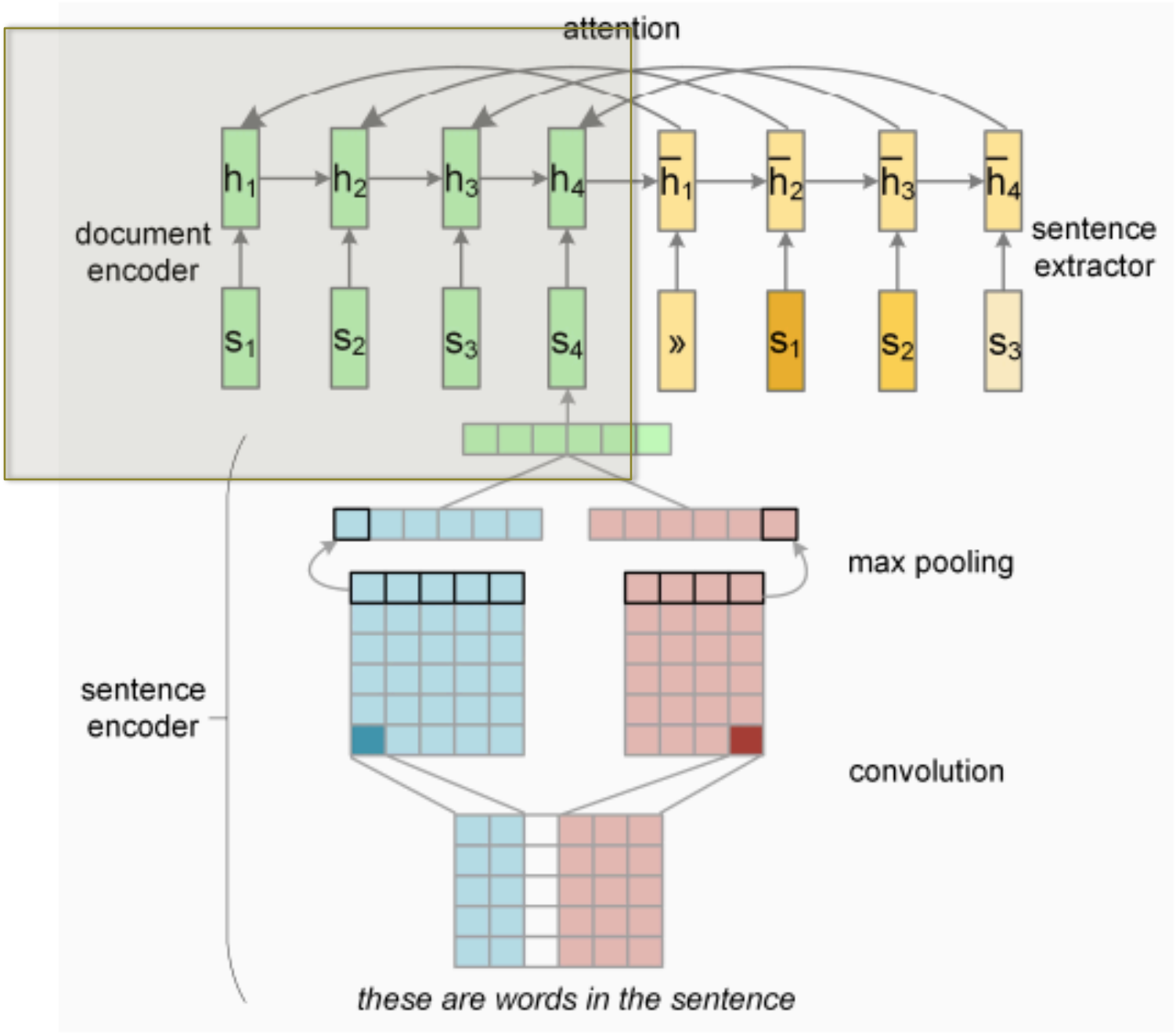
$$\mathbf{s}_{i,\mathbf{K}} = \max_j \mathbf{f}_j^i$$




Cheng and Lapata 2016

# Recurrent document encoder

- LSTM to compose a sequence of sentence vectors into a document vector
- The hidden states of the LSTM = a list of partial representations
  - Each focuses on the corresponding input sentence given previous content
- Altogether constitute document representation




Cheng and Lapata 2016

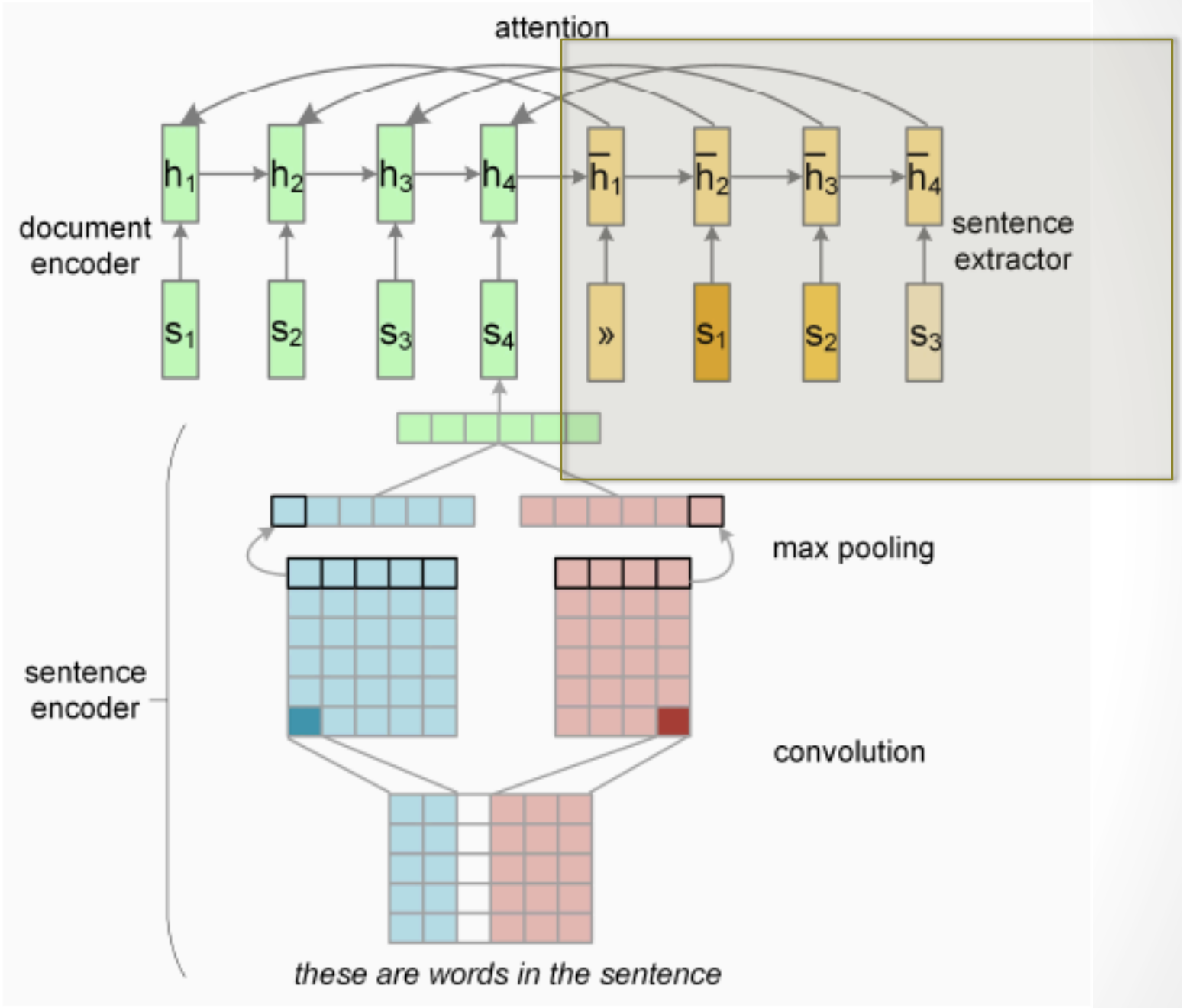
# Sentence Extractor

- Applies attention to directly extract sentences after reading them

$$\bar{\mathbf{h}}_t = \mathrm{LSTM}(p_{t-1}\mathbf{s}_{t-1}, \bar{\mathbf{h}}_{t-1})$$

$$p(y_L(t) = 1|D) = \sigma(\mathrm{MLP}(\bar{\mathbf{h}}_t : \mathbf{h}_t))$$

- ħ extractor hidden state, h encoder hidden state
  - Attends to relation between extractor and encoder hidden state
- MLP takes as input concatenated ħ and h
- $P_{t-1}$ degree to which extractor believes previous sentence should be extracted




Cheng and Lapata 2016

# Word Extractor

- Instead of extracting sentence, extracts next word
- Uses hierarchical attention to attend to sentence and word within sentence
- Output vocabulary restricted to input sentence

- -> conditional language model with vocabulary constraint

# Datasets

- Daily Mail
  - 200K training
  - 500 test
- DUC 2002
  - 567 documents with 2 summaries each

# Results

| DUC 2002 | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| LEAD | 43.6 | 21.0 | 40.2 |
| LREG | 43.8 | 20.7 | 40.3 |
| ILP | 45.4 | 21.3 | 42.8 |
| NN-ABS | 15.8 | 5.2 | 13.8 |
| TGRAPH | 48.1 | **24.3** | — |
| URANK | **48.5** | 21.5 | — |
| NN-SE | 47.4 | 23.0 | **43.5** |
| NN-WE | 27.0 | 7.9 | 22.8 |

| DailyMail | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| LEAD | 20.4 | 7.7 | 11.4 |
| LREG | 18.5 | 6.9 | 10.2 |
| NN-ABS | 7.8 | 1.7 | 7.1 |
| NN-SE | **21.2** | **8.3** | **12.0** |
| NN-WE | 15.7 | 6.4 | 9.8 |

Table 1: ROUGE evaluation (%) on the DUC-2002 and 500 DailyMail samples.

# Later results

(Nallapati et al 2017):

- RNN over sentence embeddings and output concatenated, representation of entire document by averaging, representation of summary so far by summing outputs

- (Kedzie, McKeown and Daume 2018): simpler is better
  - Averaging word embeddings, pre-trained fine, no need for summary so far

# State of the Art

- http://nlpprogress.com/english/summarization.html

- Is this a good task?

- Could you imagine other summarization tasks for which there might be data?