ECE594: Mathematical Models of Language

Spring 2022

Lecture 7: Transfer Learning and Text Summarization

Logistics

- Project proposal discussions

UNIT 2

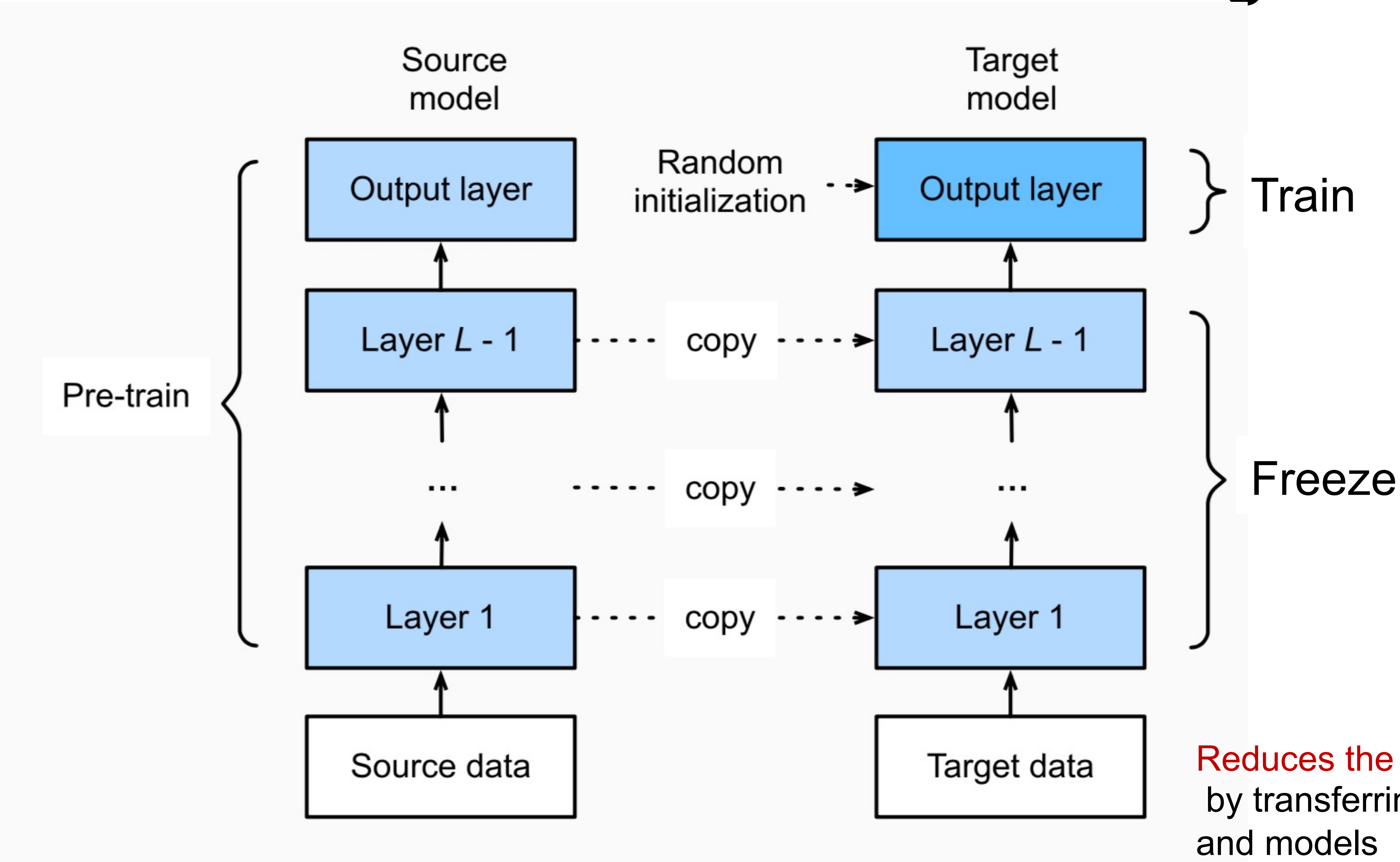
- NLP Applications
  - Summarization

- To appreciate SOTA
  - Transfer learning
  - Contextualized representation
  - Pretrained sentence representation

Supervised, Unsupervised, Semi-supervised

- Most models handled here are **supervised** learning
  - Model P(Y|X), at training time given both
- Sometimes we are interested in **unsupervised** learning
  - Model P(Y|X), at training time given only X
- Or **semi-supervised** learning
  - Model P(Y|X), at training time given both or only X

Transfer Learning

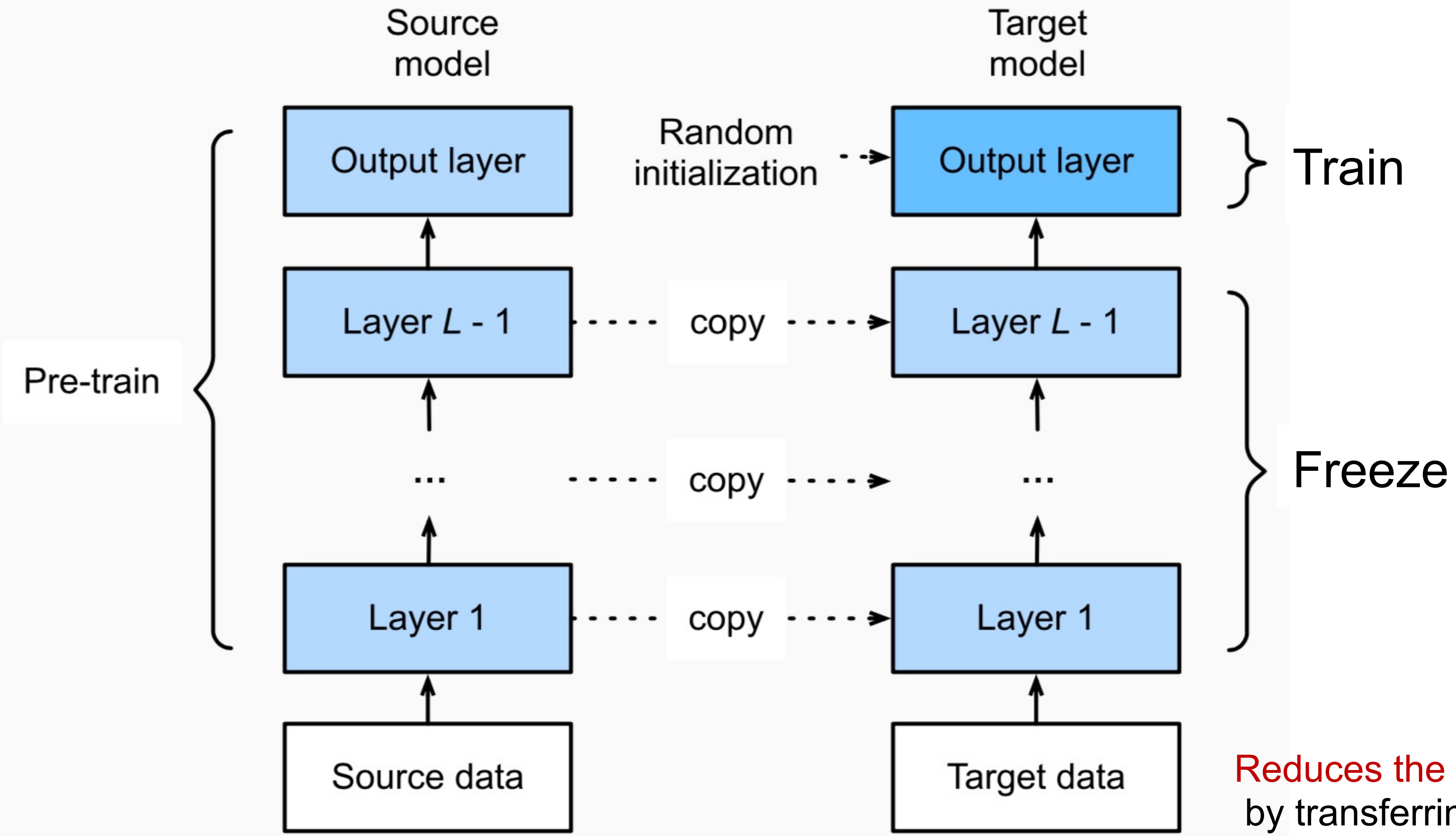


Reduces the need for labeled target data by transferring learned representations and models

So Far

- Word embeddings
  - Distributional hypothesis
  - Acquired knowledge useful in other contexts
  - One representation per word

Transfer Learning



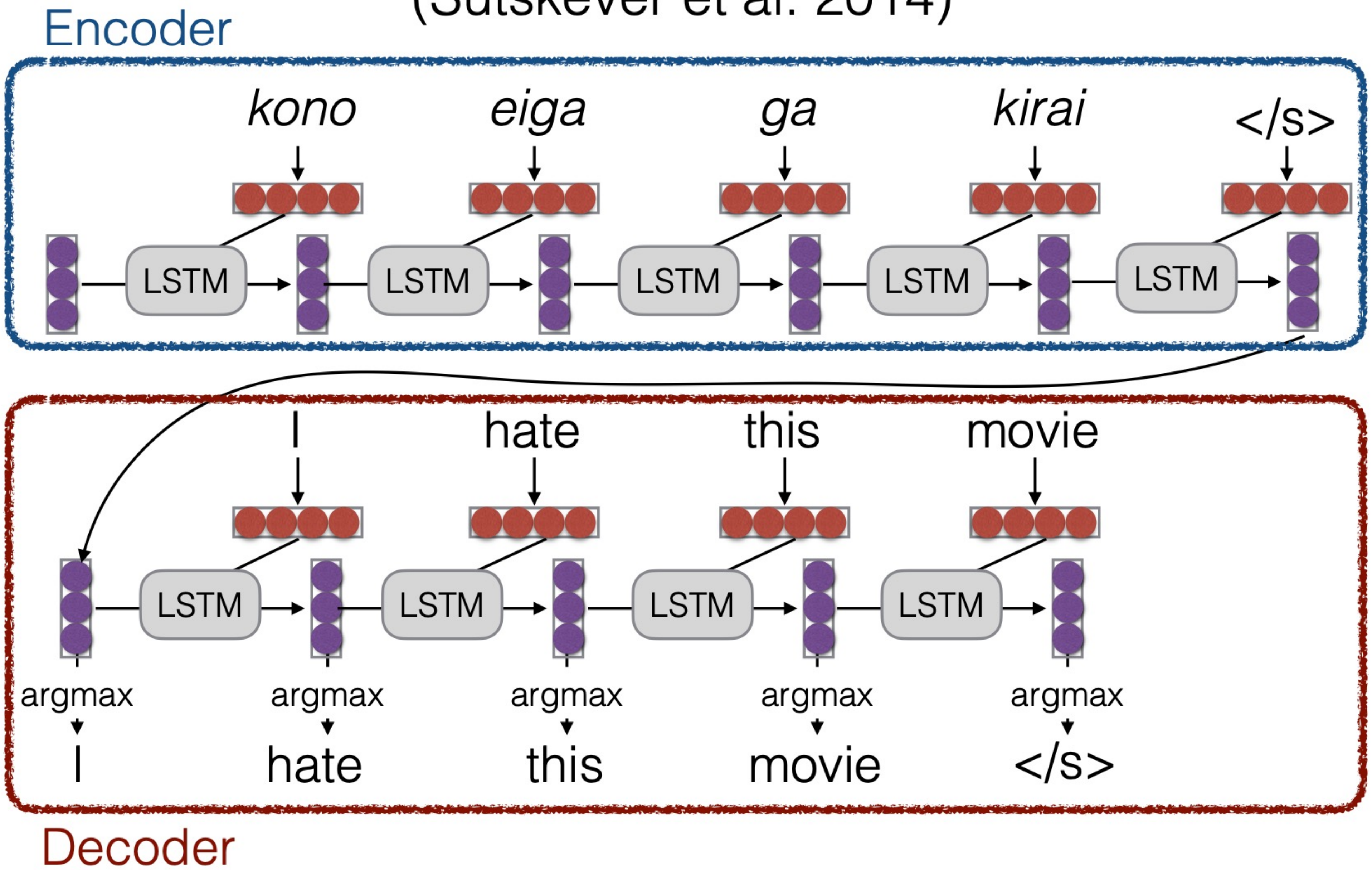
Reduces the need for labeled target data by transferring learned representations and models

Sentence representation

- Aggregated from word representation
- Separate representation

Encoder-decoder Models

(Sutskever et al. 2014)

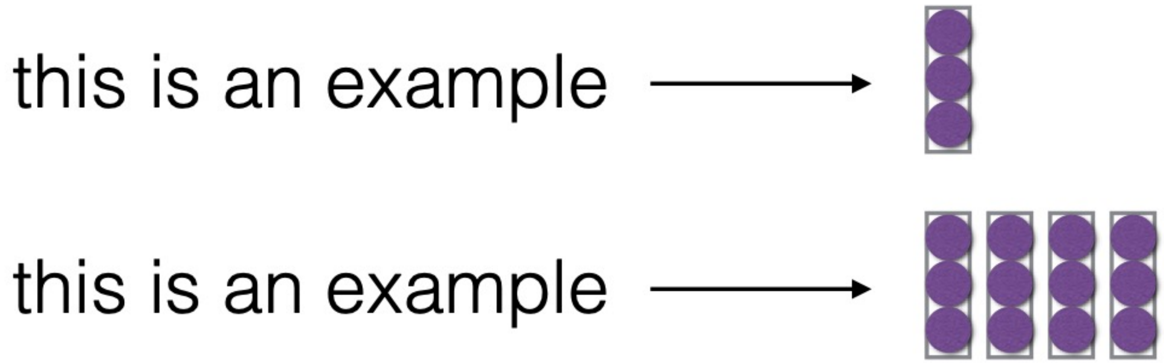


What could be a problem?

<u>Problem!</u>

> "You can't cram the meaning of a whole %&!$ing sentence into a single $&!*ing vector!"
> — Ray Mooney

- But what if we could use multiple vectors, based on the length of the sentence.

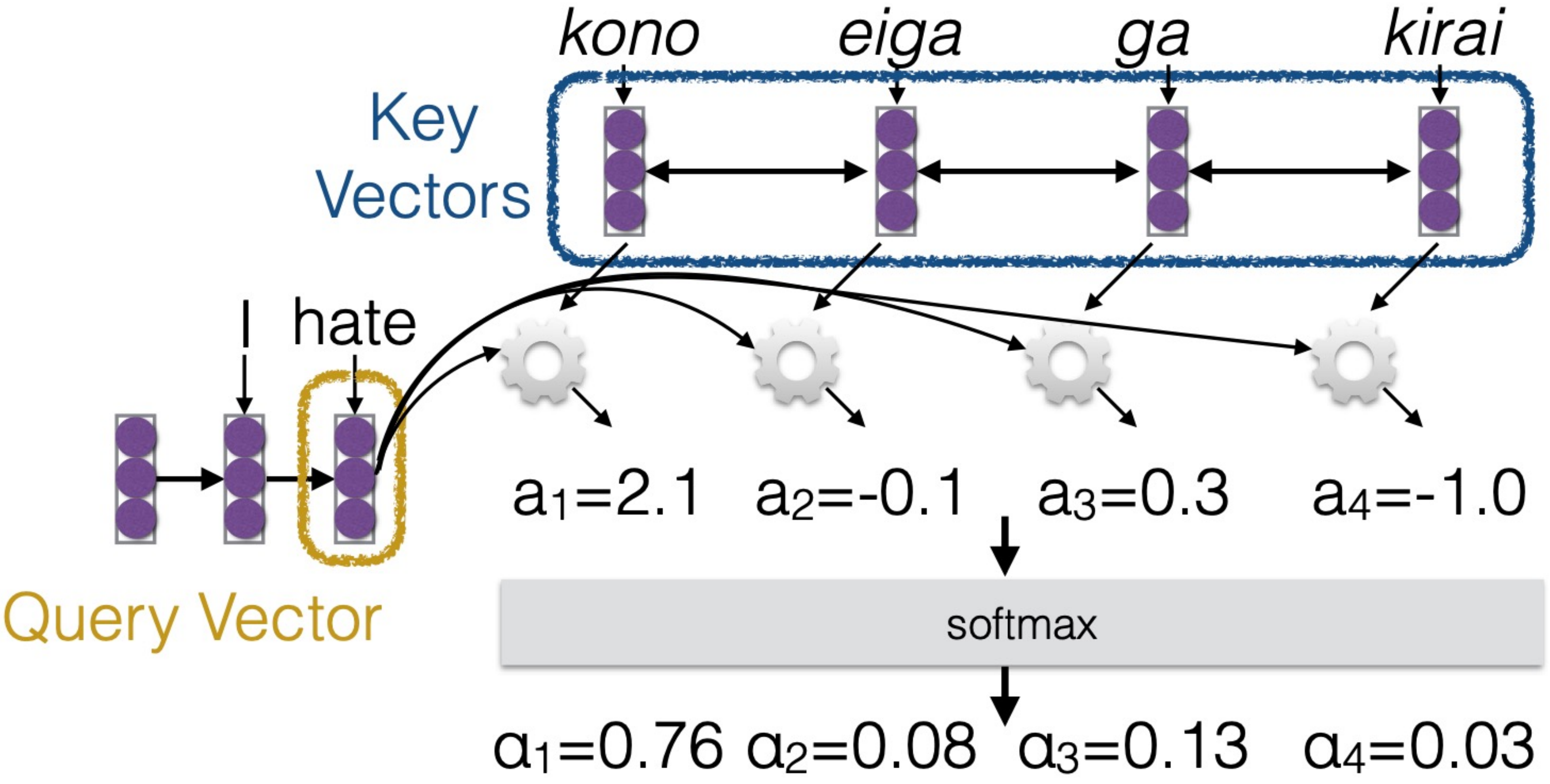


Attention (Bahdanau et al. 2015)

- Encode each word in the sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by “attention weights”
- Use this combination in picking the next word

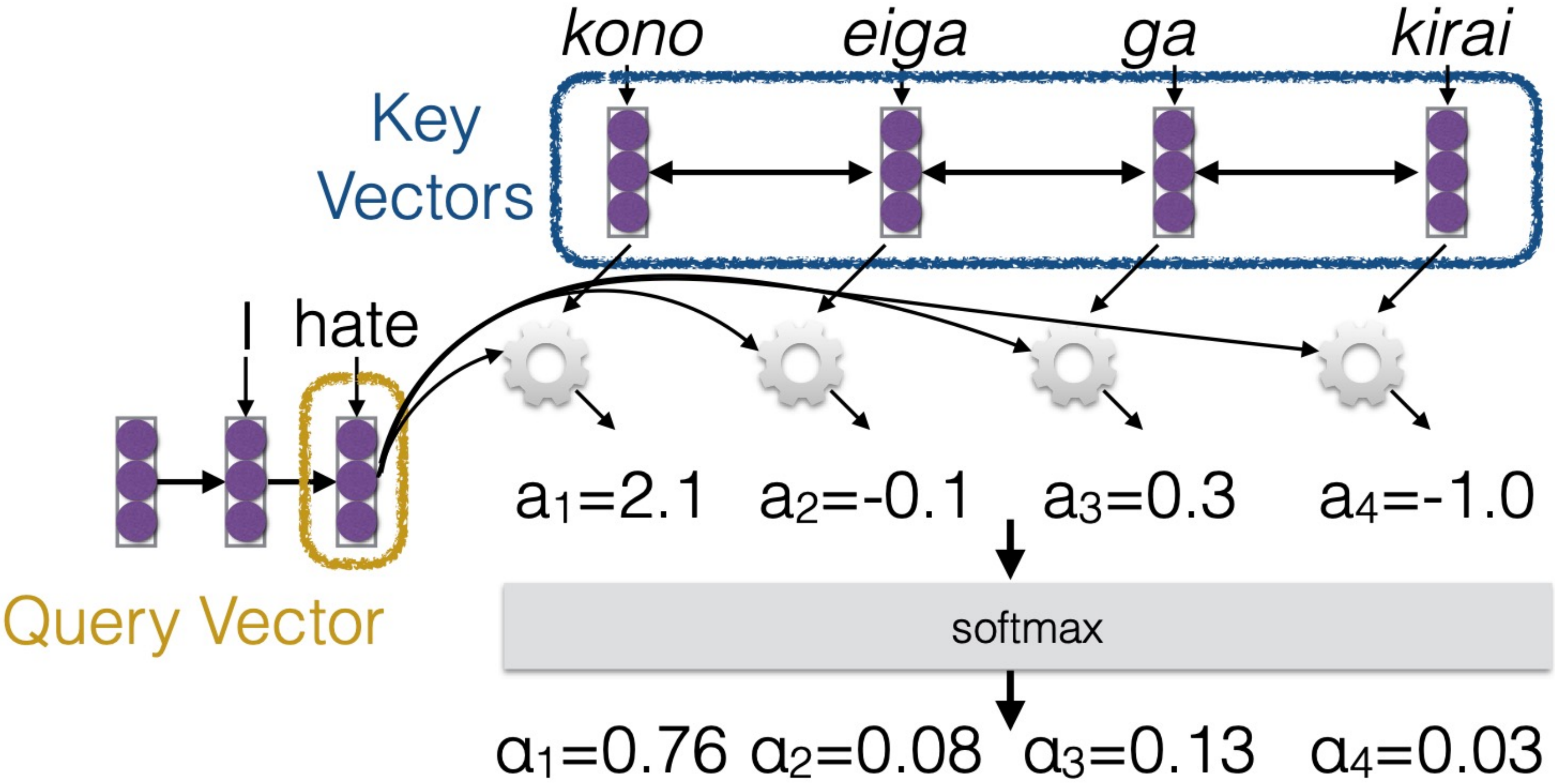
Calculating Attention

- Use "query" vector (decoder state) and "key" vectors (all encoder states)
- For each query-key pair, calculate weight
- Normalize to add to one using softmax



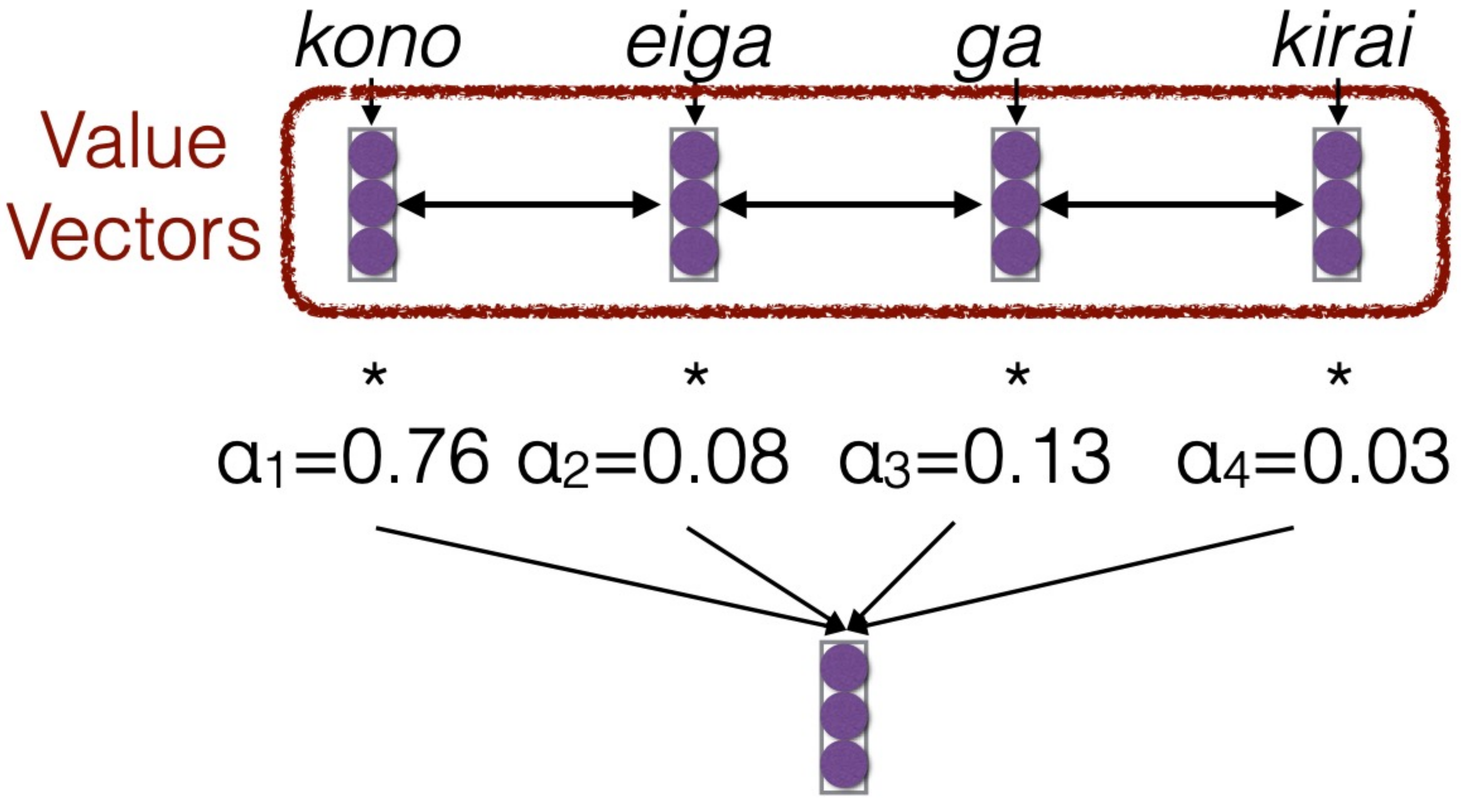
Calculating Attention (1)

- Use "query" vector (decoder state) and "key" vectors (all encoder states)
- For each query-key pair, calculate weight
- Normalize to add to one using softmax



Calculating Attention (2)

- Combine together value vectors (usually encoder states, like key vectors) by taking the weighted sum



- Use this in any part of the model you like

Scoring Functions

- $\boldsymbol{q}$ is the query and $\boldsymbol{k}$ is the key
- **Multi-layer Perceptron** (Bahdanau et al. 2015

$$a(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{w}_2^\intercal \tanh(W_1[\boldsymbol{q}; \boldsymbol{k}])$$

  - Flexible, often very good with large data
- **Bilinear** (Luong et al. 2015)

$$a(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{q}^\intercal W \boldsymbol{k}$$

- **Dot Product** (Luong et al. 2015)

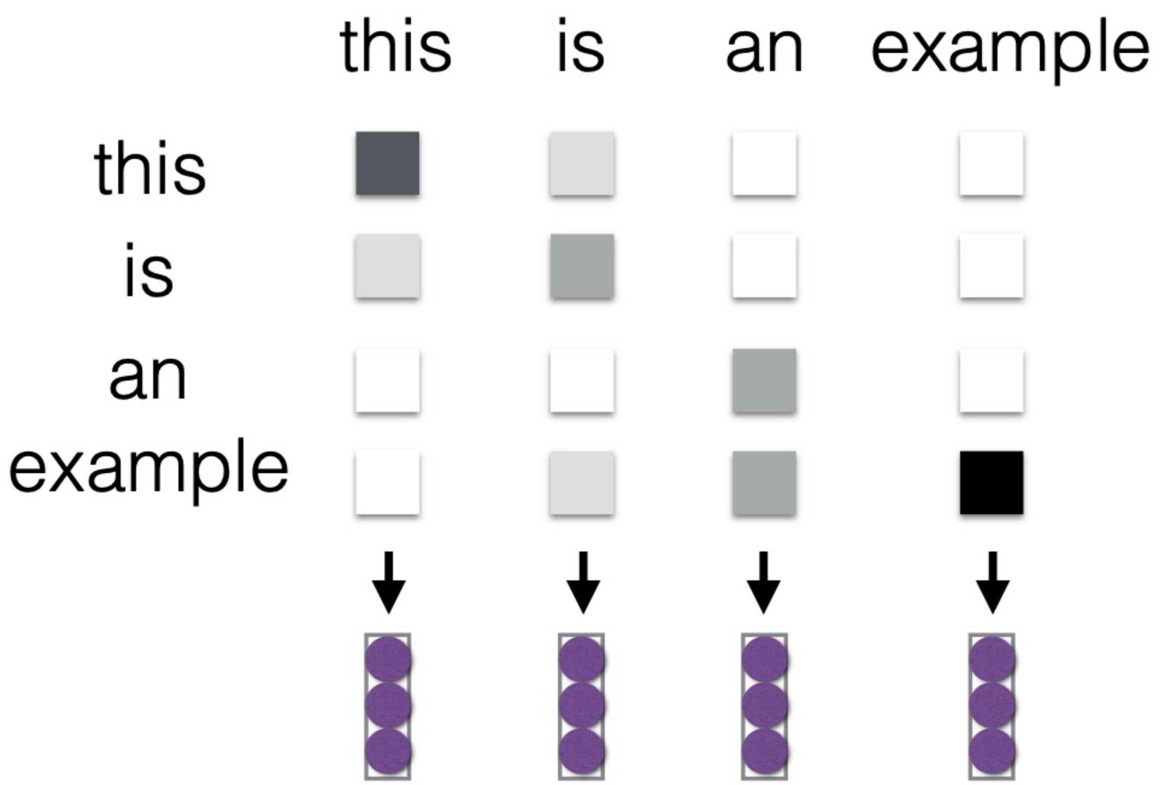
$$a(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{q}^\intercal \boldsymbol{k}$$

  - No parameters! But requires sizes to be the same.
- **Scaled Dot Product** (Vaswani et al. 2017)
  - *Problem:* scale of dot product increases as dimensions get larger
  - *Fix:* scale by size of the vector

$$a(\boldsymbol{q}, \boldsymbol{k}) = \frac{\boldsymbol{q}^\intercal \boldsymbol{k}}{\sqrt{|\boldsymbol{k}|}}$$

Self-Attention
(Cheng et al. 2016, Vaswani et al. 2017)

- Each element in the sentence attends to other elements → context sensitive encodings!

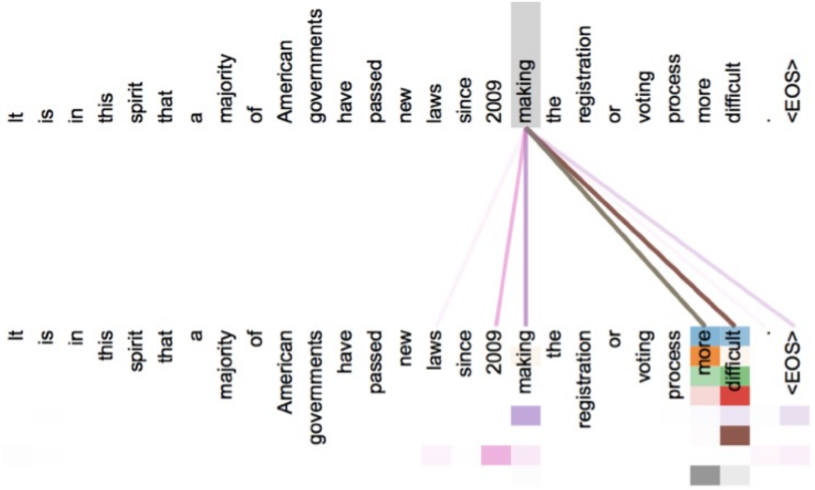


Multi-Headed Attention

- Different parts of the sentence attended to by different 'heads'
  - e.g. Different heads for "copy" vs regular (Allamanis et al. 2016)

	Target		Attention Vectors	λ
m_1	set	$\boldsymbol{\alpha} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.012
		$\boldsymbol{\kappa} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_2	use	$\boldsymbol{\alpha} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.974
		$\boldsymbol{\kappa} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_3	browser	$\boldsymbol{\alpha} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.969
		$\boldsymbol{\kappa} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_4	cache	$\boldsymbol{\alpha} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.583
		$\boldsymbol{\kappa} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	
m_5	END	$\boldsymbol{\alpha} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	0.066
		$\boldsymbol{\kappa} =$	<s> { this . use Browser Cache = use Browser Cache ; } </s>	

  - Or multiple independently learned heads (Vaswani et al. 2017)



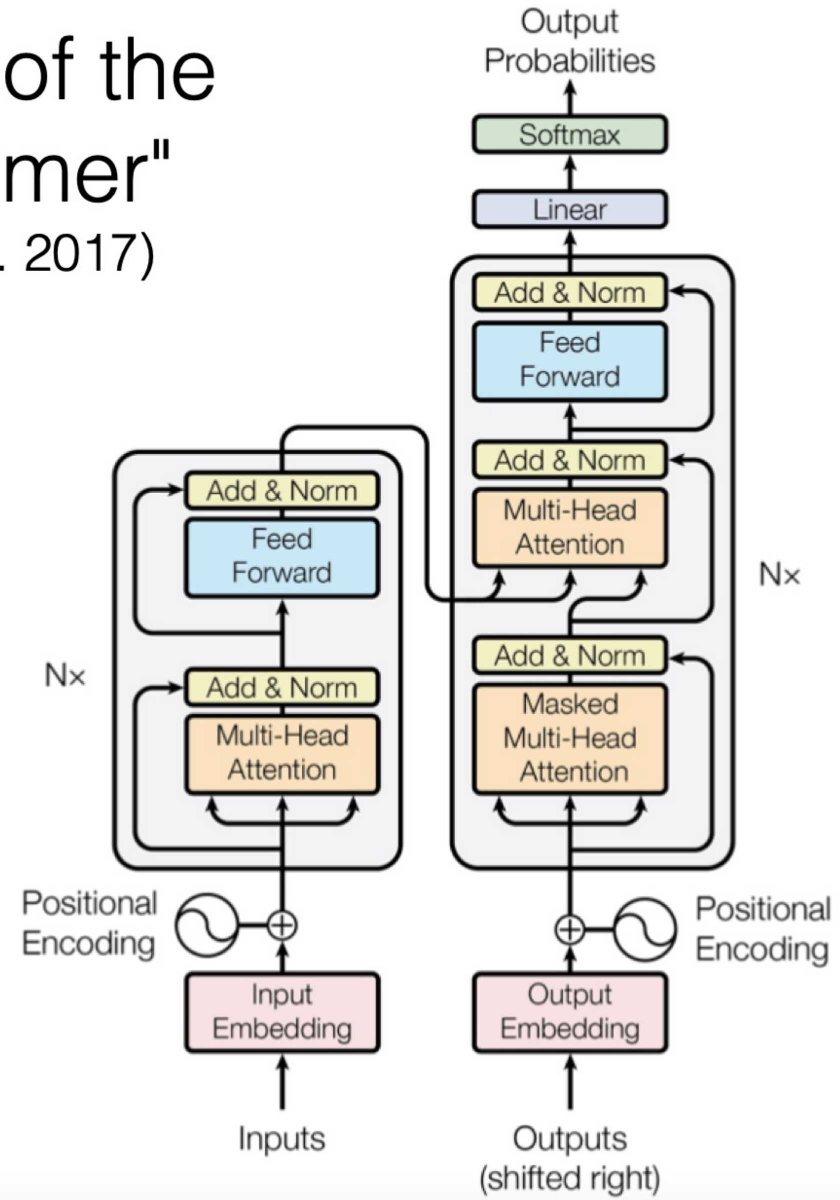
  - Or one head for every hidden node! (Choi et al. 2018)

Transformer

Summary of the “Transformer"

(Vaswani et al. 2017)

- A sequence-to-sequence model based entirely on attention
- Strong results on translation, a wide variety of other tasks
- Fast: only matrix multiplications

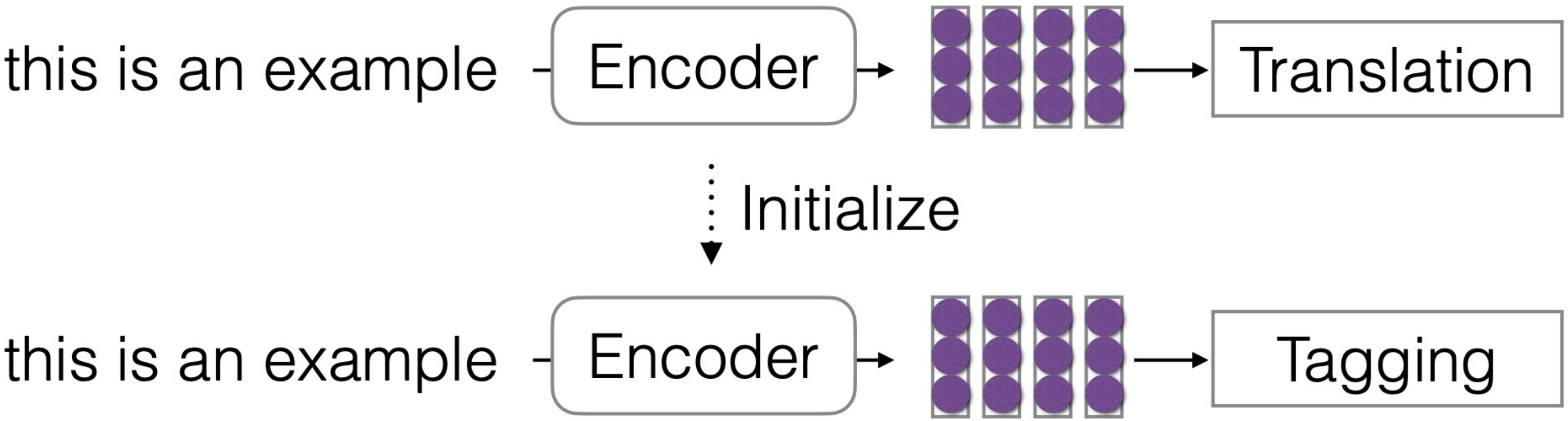


Pretrained Sentence Representation

- Needed for
  - Paraphrase ID, retrieval
  - Sentence classification

Pretrained Sentence Representation

- First train on one task, then train on another



- Widely used in word embeddings (Turian et al. 2010)

- Also pre-training sentence encoders or contextualized word representations (Dai et al. 2015, Melamud et al. 2016)

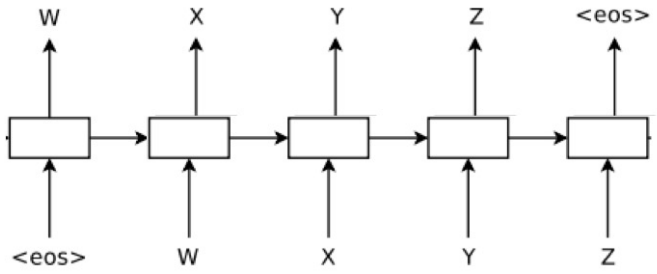
Pretrained Representations

- Many methods have names like SkipThought, ParaNMT, CoVe, ELMo, BERT along with pre-trained models
- These often refer to a combination of
  - **Model:** The underlying neural network architecture
  - **Training Objective:** What objective is used to pre-train
  - **Data:** What data the authors chose to use to train the model

Language Model + Transfer

(Dai and Le 2015)

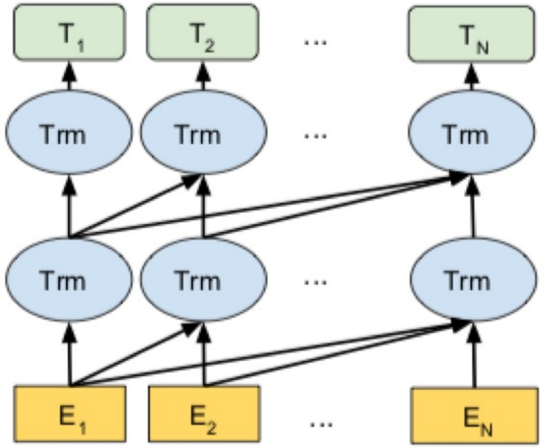
- **Model:** LSTM
- **Objective:** LM objective
- **Data:** Classification data itself, or Amazon reviews



- **Downstream:** On text classification, initialize weights and continue training

"GPT" (Radford et al. 2018)

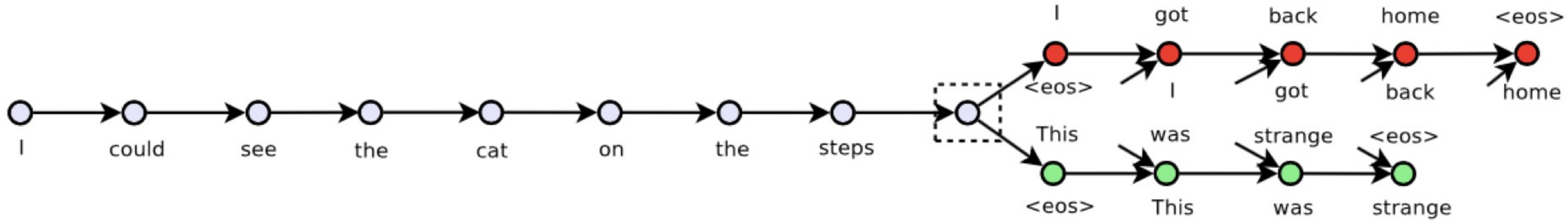
- **Model:** Masked self-attention
- **Objective:** LM objective
- **Data:** BooksCorpus



Downstream: Some task fine-tuning, other tasks additional multi-sentence training

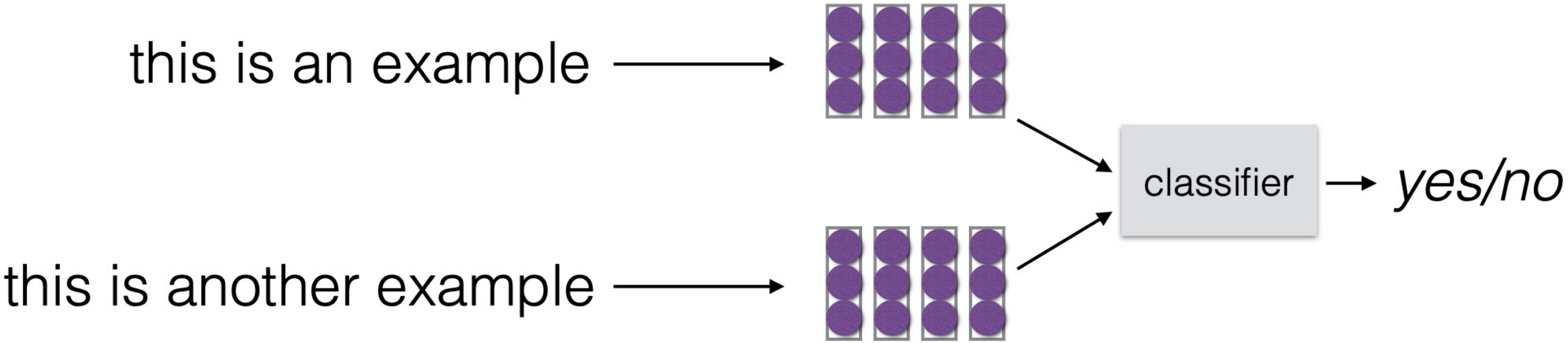
Sentence-level Context Prediction+Transfer: "Skip-thought Vectors" (Kiros et al. 2015)

- **Model:** LSTM
- **Objective:** Predict the surrounding sentences
- **Data:** Books, important because of context



Contextualized Word Representations

- Instead of one vector per sentence, one vector per word!



How to train this representation?

Text Summarization

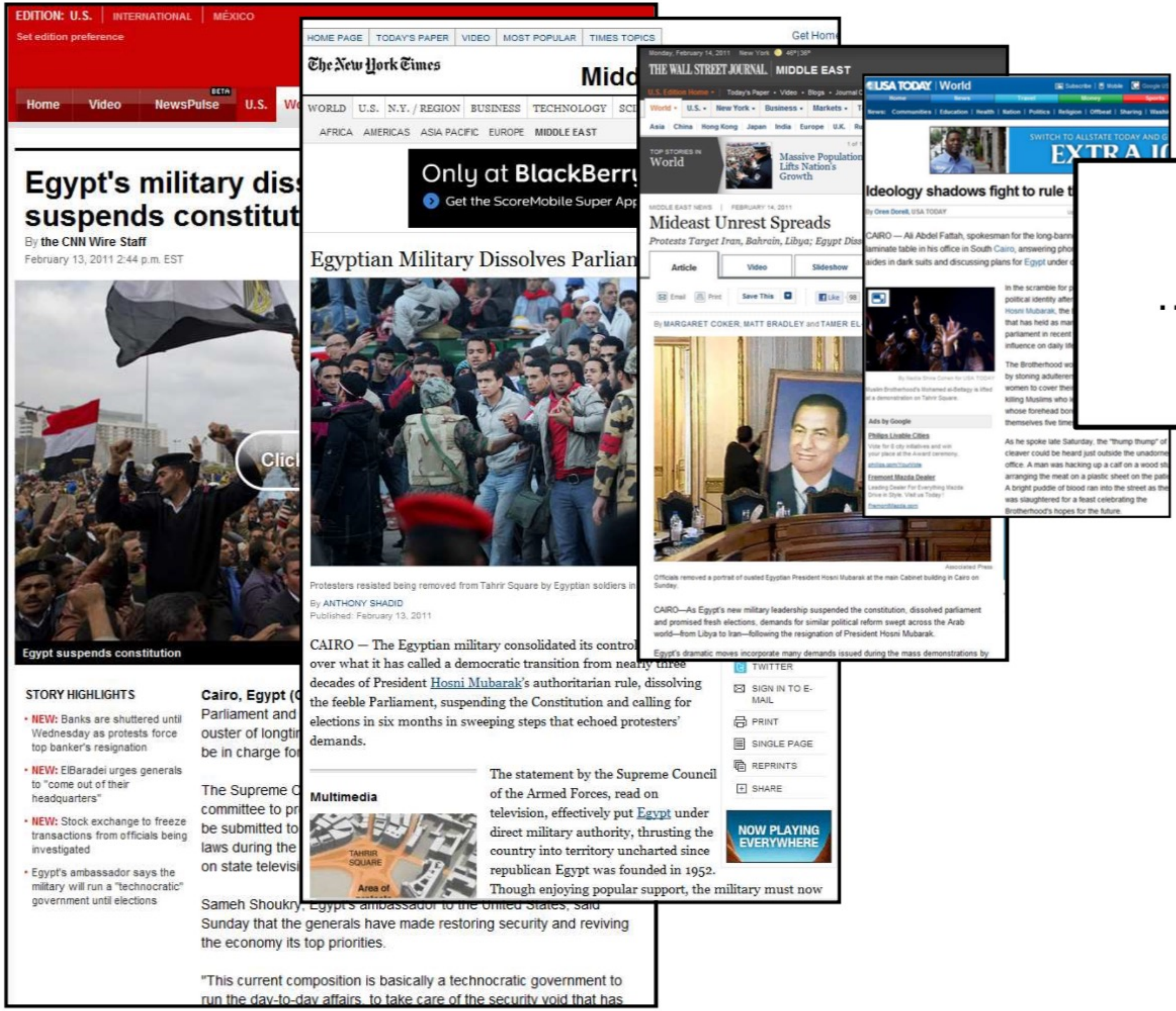
- Extractive summarization
  - summary is a subset of original text
- Abstractive summarization
  - summary is paraphrase of original text

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle- field of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

Figure 23.12 The Gettysburg Address. Abraham Lincoln, 1863.

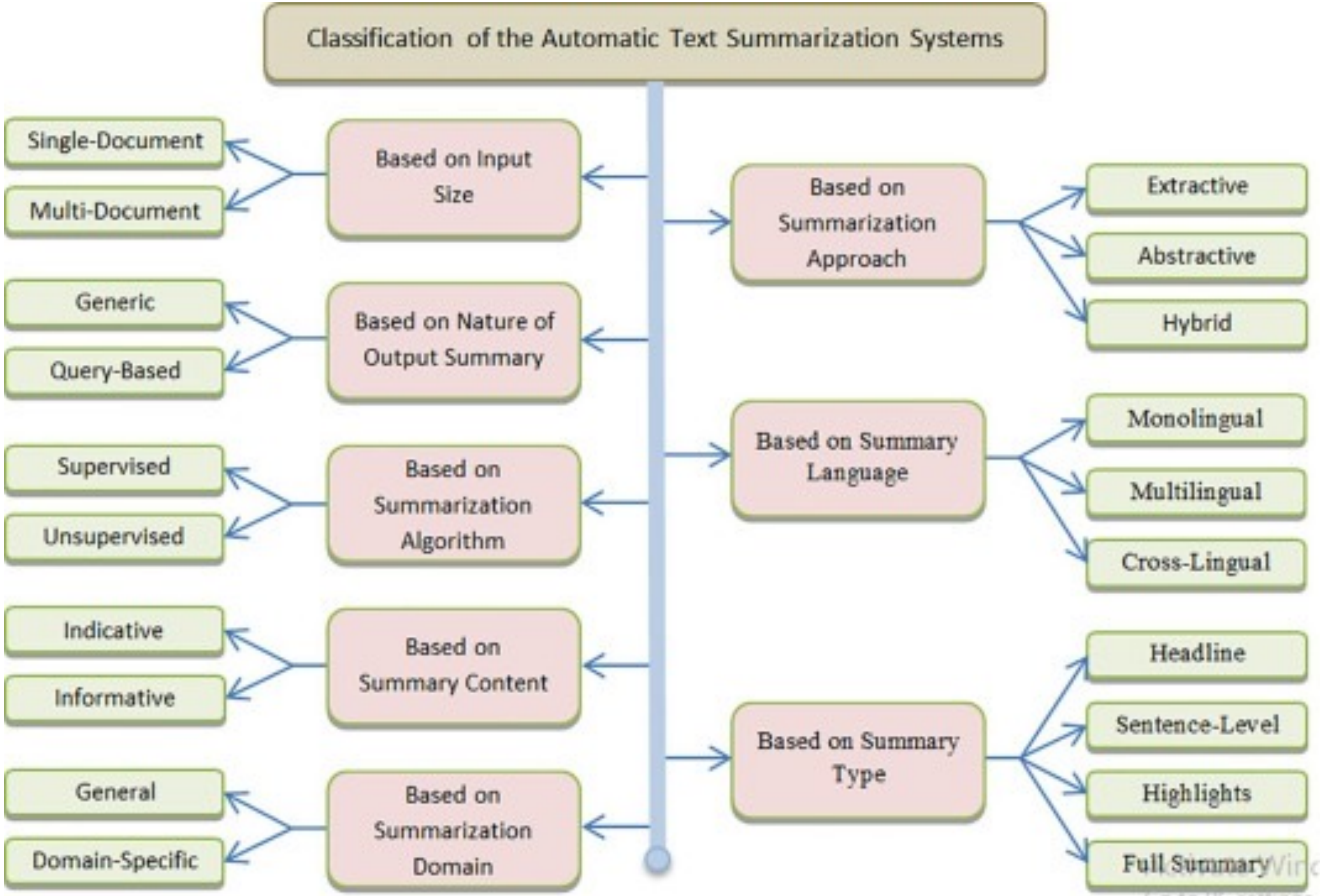
Extract from the Gettysburg Address:
Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field. But the brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that government of the people, by the people for the people shall not perish from the earth.
Abstract of the Gettysburg Address:
This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

Figure 23.13 An extract versus an abstract from the Gettysburg Address (abstract from Mani (2001)).


Egypt's military dis
suspends constitut
By the CNN Wire Staff
February 13, 2011 2:44 p.m. EST
Egypt suspends constitution
STORY HIGHLIGHTS
The New York Times
Egyptian Military Dissolves Parlian
By ANTHONY SHADID
Published: February 13, 2011
CAIRO — The Egyptian military consolidated its control over what it has called a democratic transition from nearly three decades of President Hosni Mubarak's authoritarian rule, dissolving the feeble Parliament, suspending the Constitution and calling for elections in six months in sweeping steps that echoed protesters' demands.
THE WALL STREET JOURNAL. MIDDLE EAST
Mideast Unrest Spreads
USA TODAY World
Ideology shadows fight to rule t

… 27,000+ more

Text Summarization



- Generic summarization:
  - Summarize the content of a document
- Query-focused summarization:
  - summarize a document with respect to an information need expressed in a user query.
  - a kind of complex question answering:
    - Answer a question by summarizing a document that has the information to construct the answer

Extractive Summarization

- Select units from the original
  - Typically sentences
  - No simplification/rewriting
- Baseline
  - Extract the first few sentences (news genre)

Extractive Summarization

- Long history
  - Baxendale (1958)
  - Luhn (1958; technical documents)

- Heuristics
  - Position of sentences
    - Analyzed 200 paragraphs; first and last are topic sentences
  - Sentences with content terms (frequency/uniqueness)
  - Cue words (*hardly, significant, impossible*)

Extractive Summarization

- Problems
  - Paice (1990)
    - Lack of balance (e.g., single views)
    - Lack of cohesion (antecedent not mentioned/incorrectly cited)
- Solutions
  - Rhetorical structure theory
  - Anaphors
    - *That:* nonanaphoric if preceded by a research verb (demonstrated)

      : nonanaphoric if followed by pronoun, article, quantifier

Summarization Tasks

- Content selection
  - Choose sentences to extract
- Information ordering
  - Order sentences
- Realization
  - Cleanup and present

Text Summarization