CS11-747 Neural Networks for NLP

Using/Evaluating Sentence Representations

Graham Neubig

Carnegie Mellon University

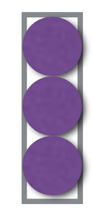
Language Technologies Institute

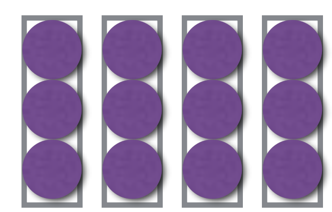
Site

https://phontron.com/class/nn4nlp2017/

Sentence Representations

- We can create a vector or sequence of vectors from a sentence

this is an example → 

this is an example → 

<u>Obligatory Quote!</u>

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

How do We Use/Evaluate Sentence Representations?

- Sentence Classification
- Paraphrase Identification
- Semantic Similarity
- Entailment
- Retrieval

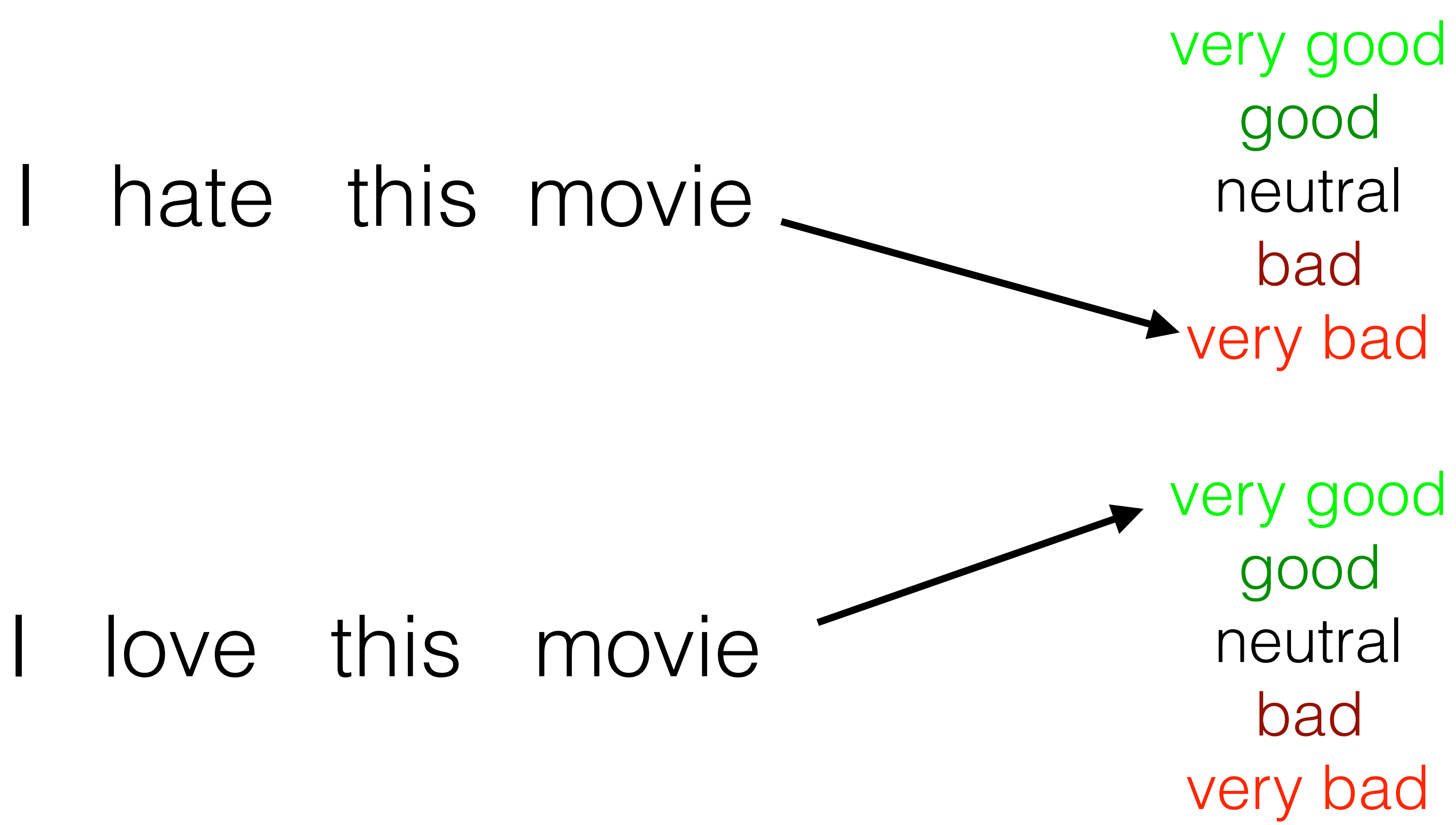
Goal for Today

- Introduce **tasks/evaluation metrics**
- Introduce **common data sets**
- Introduce **methods**, and particularly state of the art results

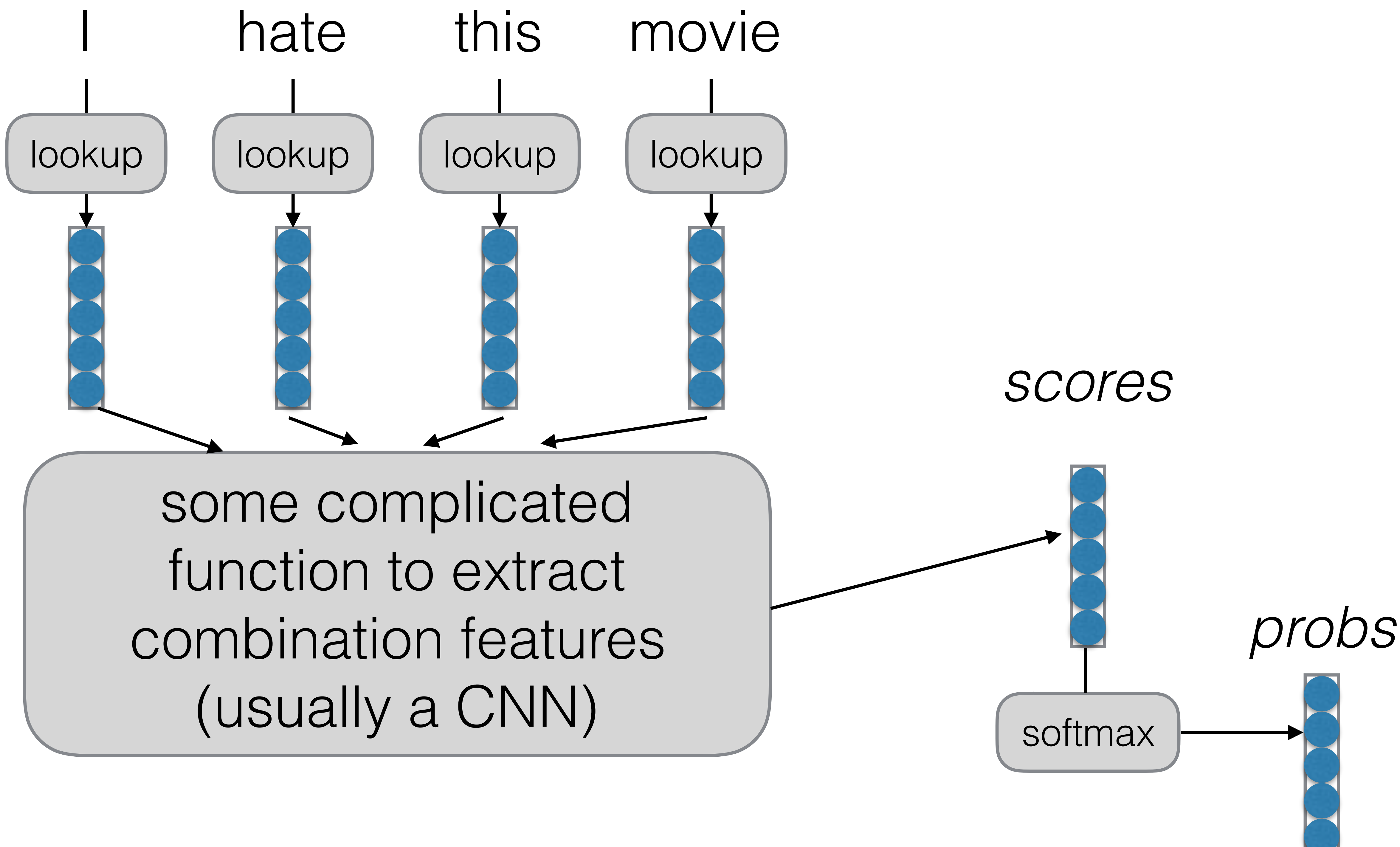
Sentence Classification

Sentence Classification

- Classify sentences according to various traits
- Topic, sentiment, subjectivity/objectivity, etc.



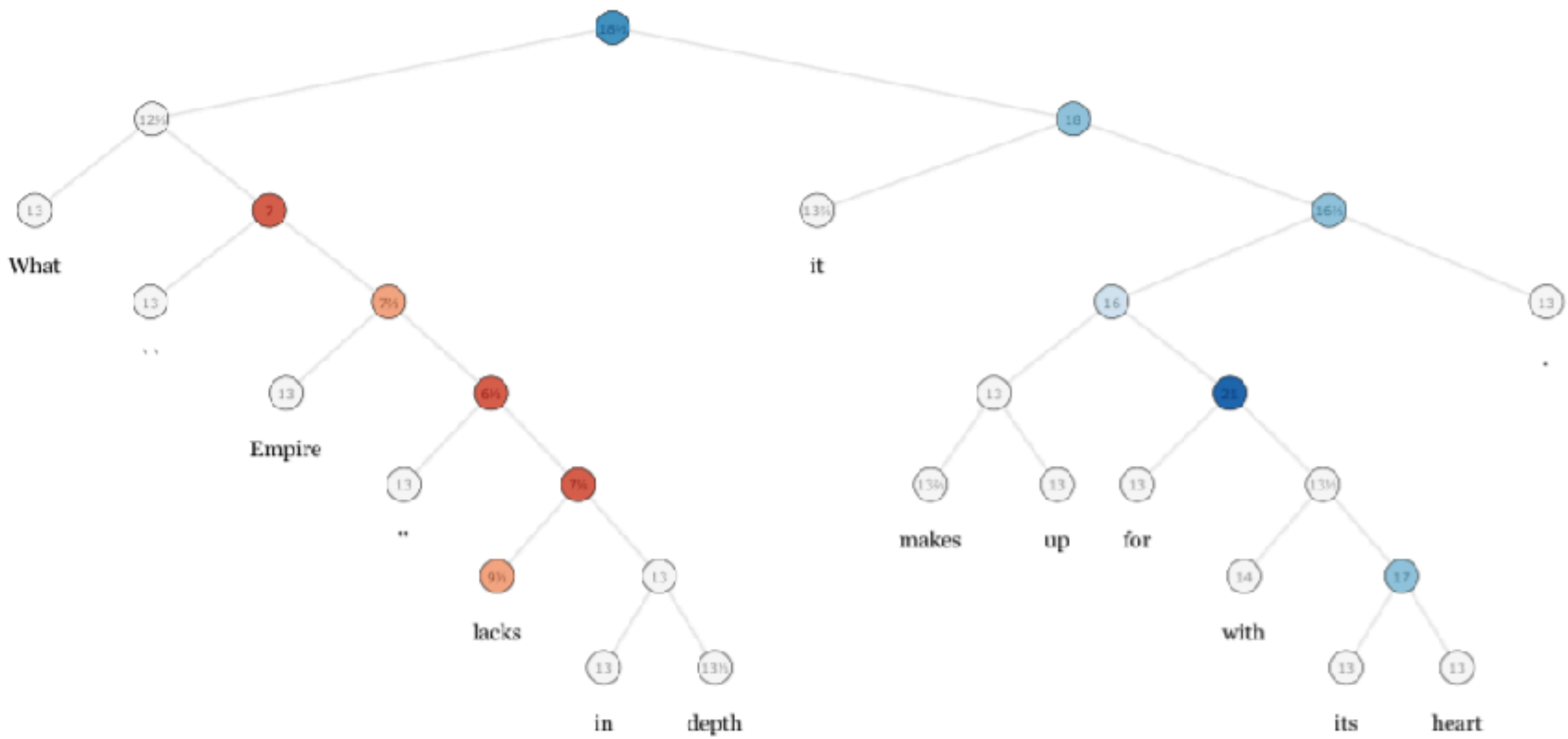
Model Overview (Review)



Data Example: Stanford Sentiment Treebank

(Socher et al. 2013)

- In addition to standard tags, each constituent tagged with a sentiment value



Paraphrase Identification

Paraphrase Identification
(Dolan and Brockett 2005)

- Identify whether A and B mean the same thing

Charles O. Prince, 53, was named as Mr. Weill's successor.

↕

Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

- **Note:** *exactly* the same thing is too restrictive, so use a loose sense of similarity

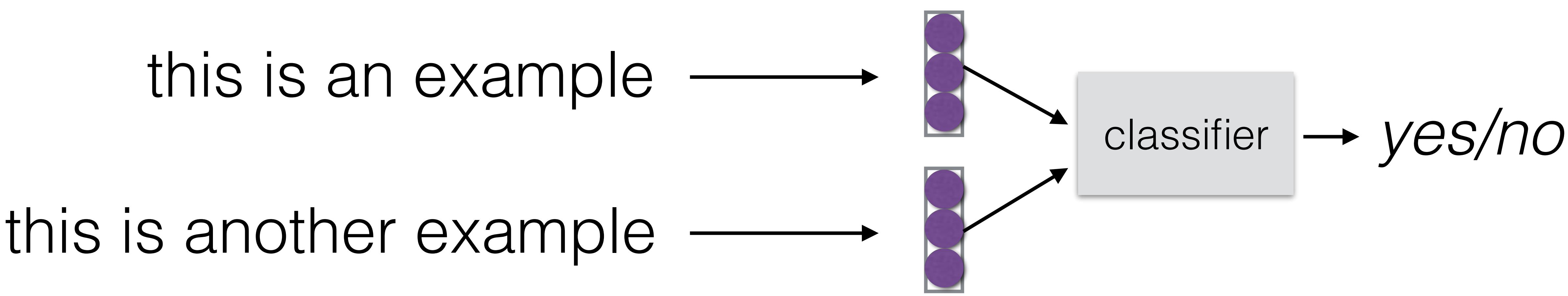
Data Example: Microsoft Research Paraphrase Corpus

(Dolan and Brockett 2005)

- **Construction procedure**
  - Crawl large news corpus
  - Identify sentences that are similar automatically using heuristics or classifier
  - Have raters determine whether they are in fact similar (67% were)
- Corpus is **high quality but small**, 5,800 sentences
- **c.f.** Other corpora based on translation, image captioning

Models for Paraphrase Detection (1)

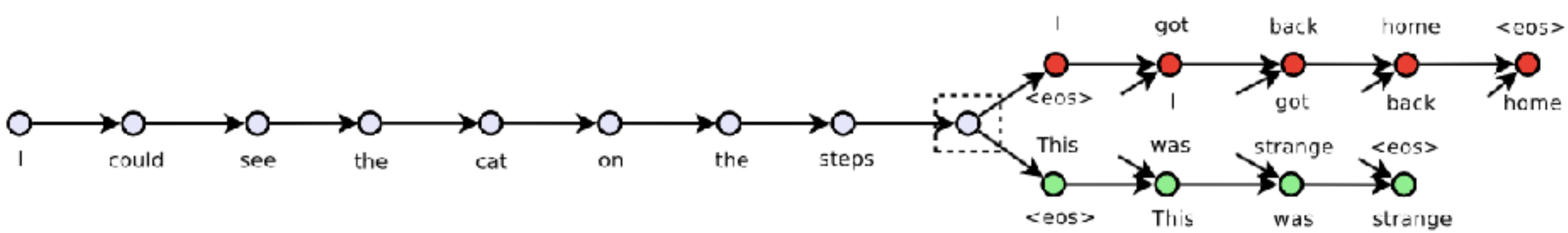
- Calculate vector representation
- Feed vector representation into classifier



Model Example: Skip-thought Vectors

(Kiros et al. 2015)

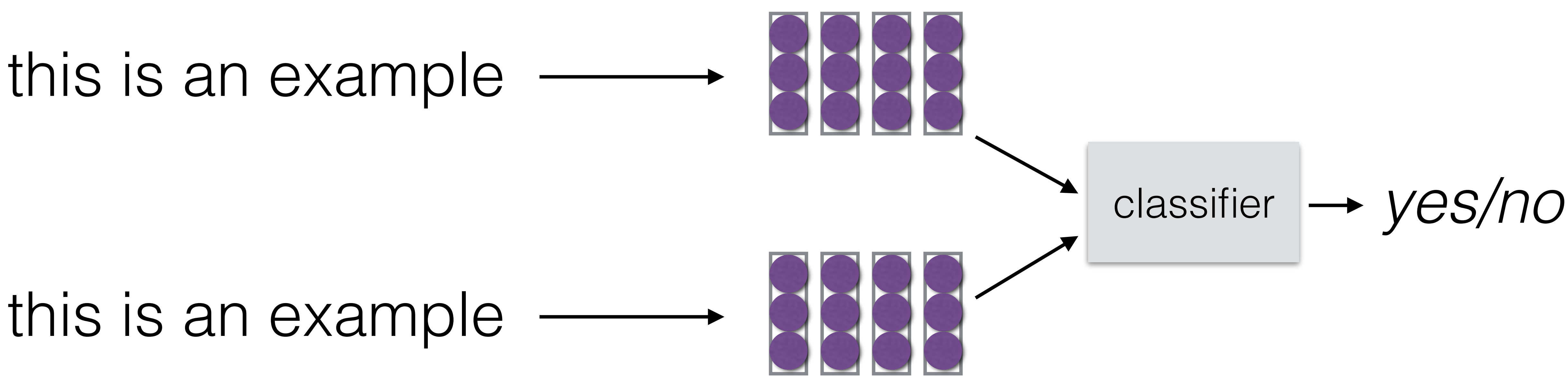
- General method for sentence representation
  - Unsupervised training: predict surrounding sentences on large-scale data (using encoder-decoder)
  - Use resulting representation as sentence representation



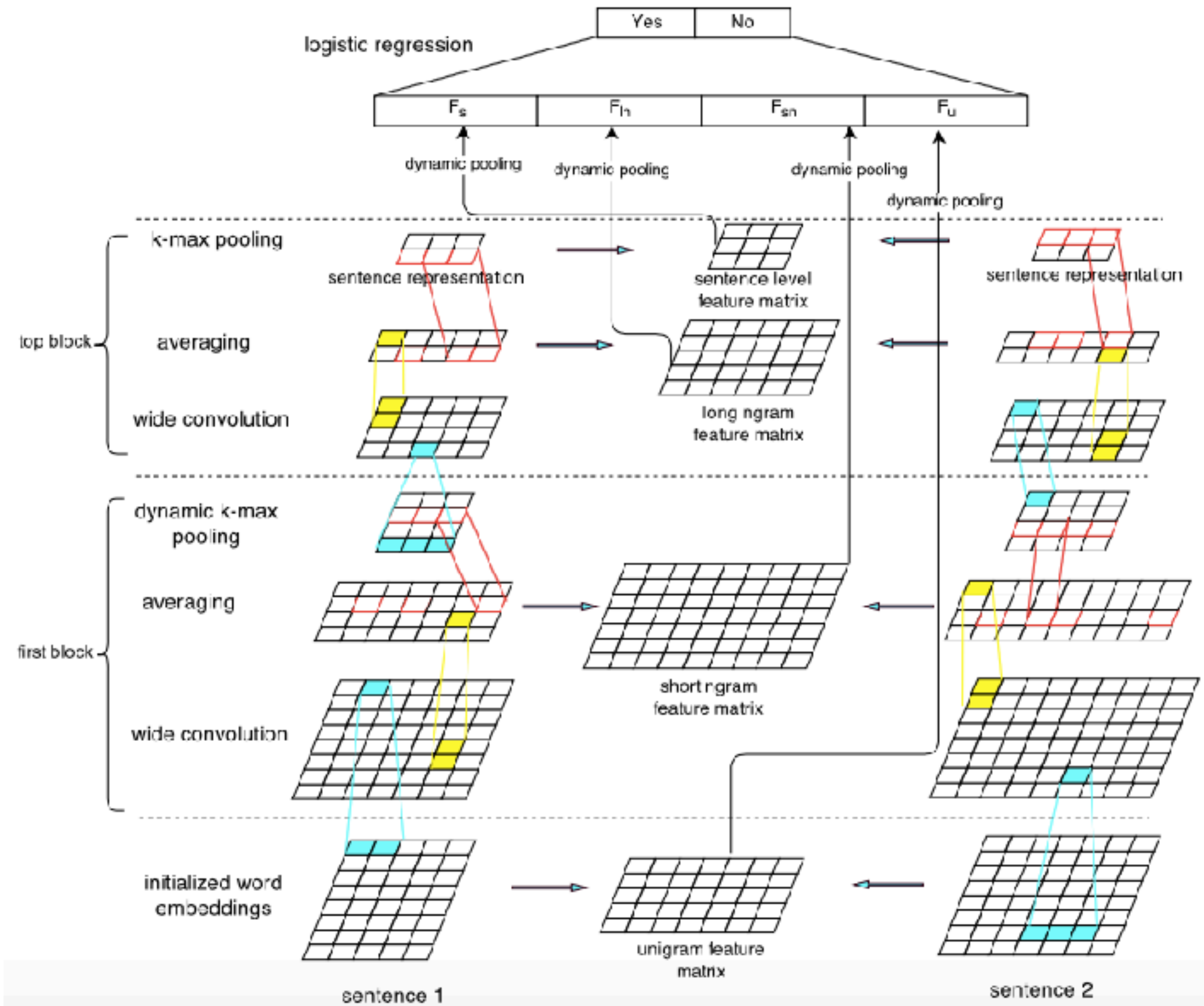
- Train logistic regression on [|u-v|; u*v] (component-wise)

Models for Paraphrase Detection (2)

- Calculate multiple-vector representation, and combine to make a decision



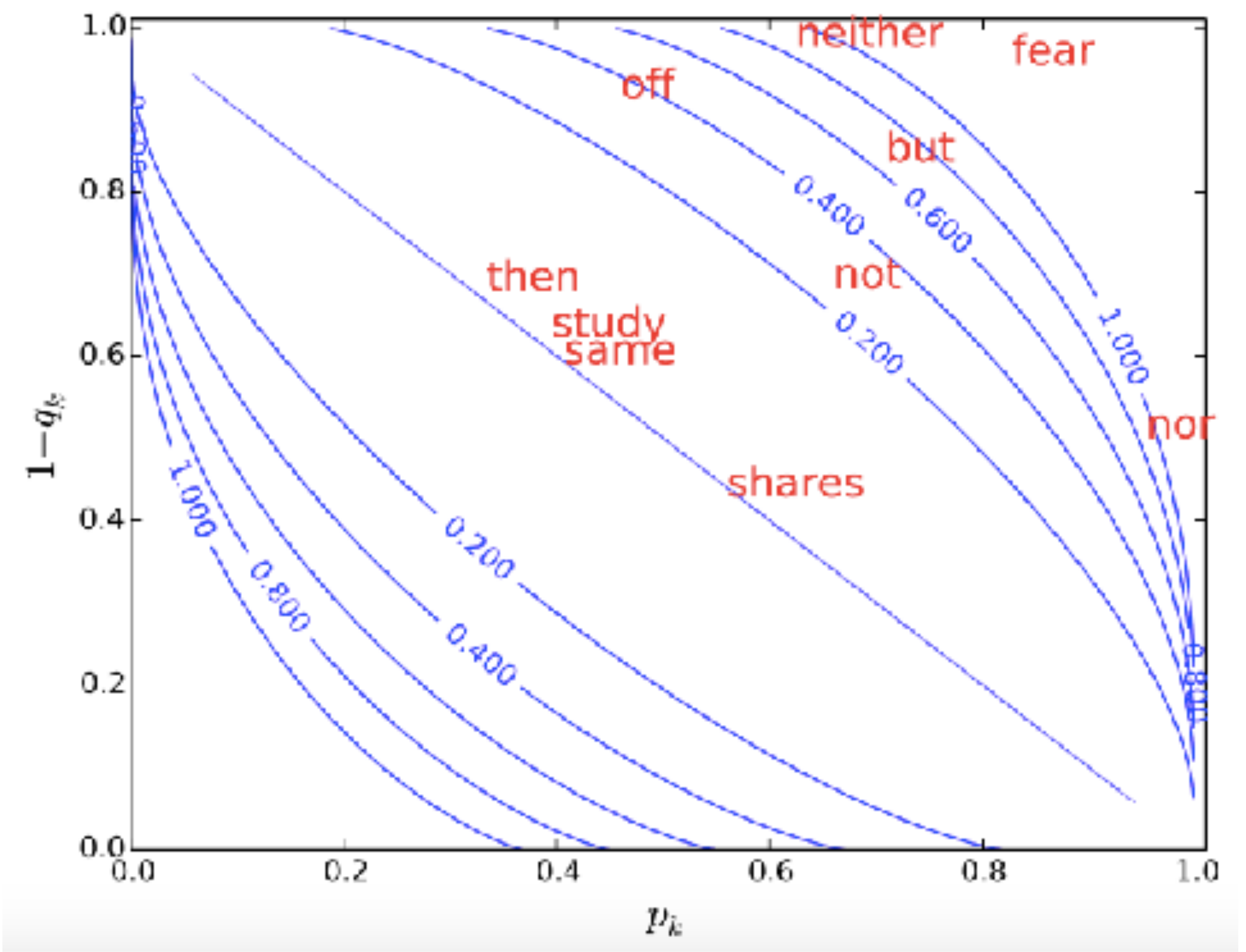
Model Example: Convolutional Features + Matrix-based Pooling (Yin and Schutze 2015)



Model Example: Paraphrase Detection w/ Discriminative Embeddings

(Ji and Eisenstein 2013)

- Perform matrix factorization of word/ context vectors
    - Weight word/context vectors based on discriminativeness



- *Also add features regarding surface match*
- Current state-of-the-art on MSRPC

Semantic Similarity

Semantic Similarity/Relatedness

(Marelli et al. 2014)

- Do two sentences mean something similar?

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

- Like paraphrase identification, but with shades of gray.

Data Example: SICK Dataset
(Marelli et al. 2014)

- Procedure to create sentences
  - Start with **short flickr/video description sentences**
  - **Normalize** sentences (11 transformations such as active↔passive, replacing w/ synonyms, etc.)
  - **Create opposites** (insert negation, invert determiners, replace words w/ antonyms)
  - **Scramble words**
- Finally **ask humans to measure semantic relatedness** on 1-5 Likert scale of “completely unrelated - very related”

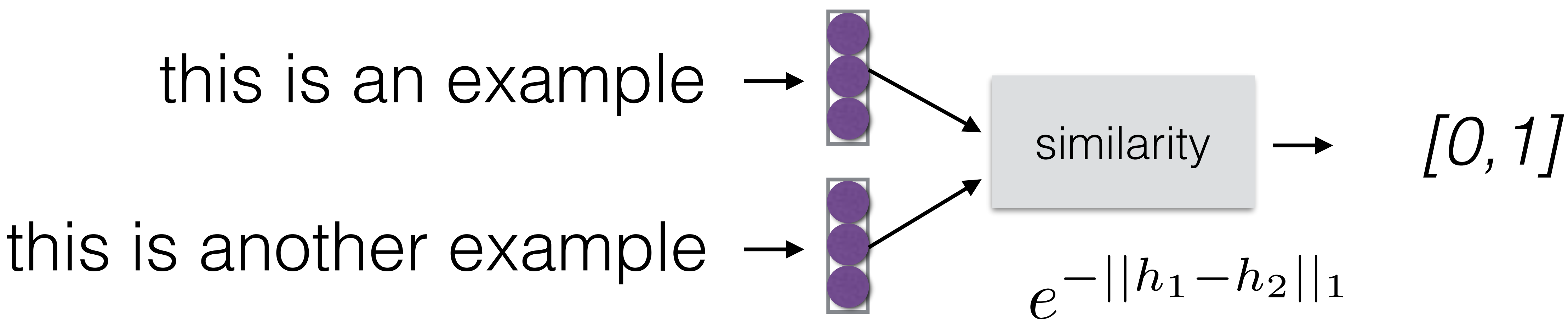
Evaluation Procedure

- Input two sentences into model, calculate score
- Measure correlation of the machine score with human score (e.g. Pearson's correlation)

Model Example:
Siamese LSTM Architecture

(Mueller and Thyagarajan 2016)

- Use **siamese LSTM architecture** with e^-L1 as a similarity metric



- **Simple model!** Good results due to engineering? Including pre-training, using pre-trained word embeddings, etc.
- Results in best reported accuracies for SICK task

Textual Entailment

Textual Entailment
(Dagan et al. 2006, Marelli et al. 2014)

- **Entailment:** if A is true, then B is true (c.f. paraphrase, where opposite is also true)
  - The woman bought a sandwich for lunch
    → The woman bought lunch
- **Contradiction:** if A is true, then B is not true
  - The woman bought a sandwich for lunch
    → The woman did not buy a sandwich
- **Neutral:** cannot say either of the above
  - The woman bought a sandwich for lunch
    → The woman bought a sandwich for dinner

Data Example: Stanford Natural Language Inference Dataset

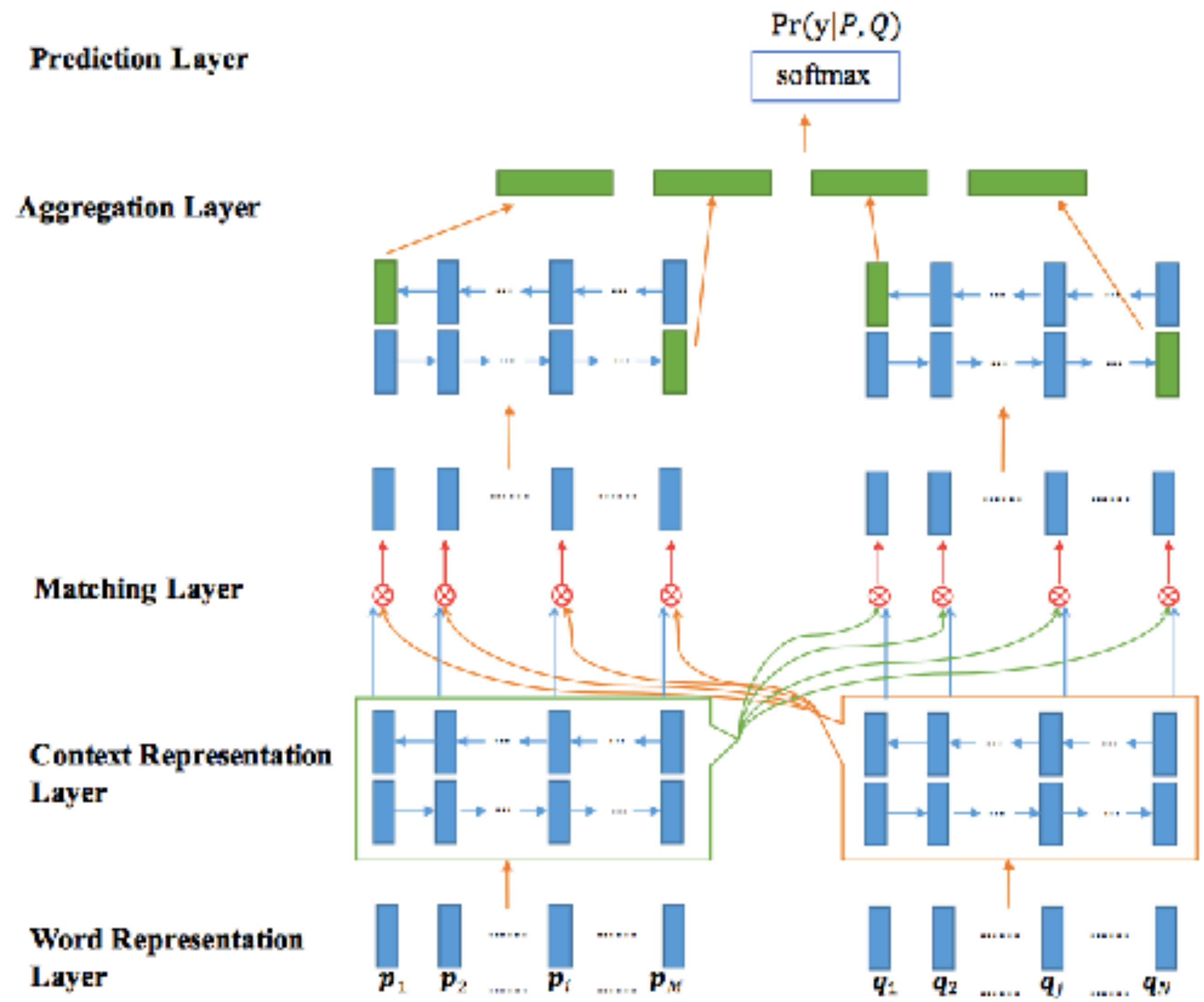
(Bowman et al. 2015)

- Data created from **Flickr captions**
- **Crowdsource** creation of one entailed, neutral, and contradicted caption for each caption
- **Verify** the captions with 5 judgements, 89% agreement between annotator and “gold” label
- Also, **expansion to multiple genres: MultiNLI**

Model Example: Multi-perspective Matching for NLI (Wang et al. 2017)

- Encode, aggregate information in both directions, encode one more time, predict

- Strong results on SNLI



- Lots of other examples on SNLI web site: https://nlp.stanford.edu/projects/snli/

Interesting Result: Entailment → Generalize

(Conneau et al. 2017)

- Skip-thought vectors are **unsupervised training**
- Simply: can **supervised training** for a task such as inference learn generalizable embeddings?
  - Task is more difficult and requires capturing nuance → yes?
  - Data is much smaller → no?
- Answer: **yes**, generally better

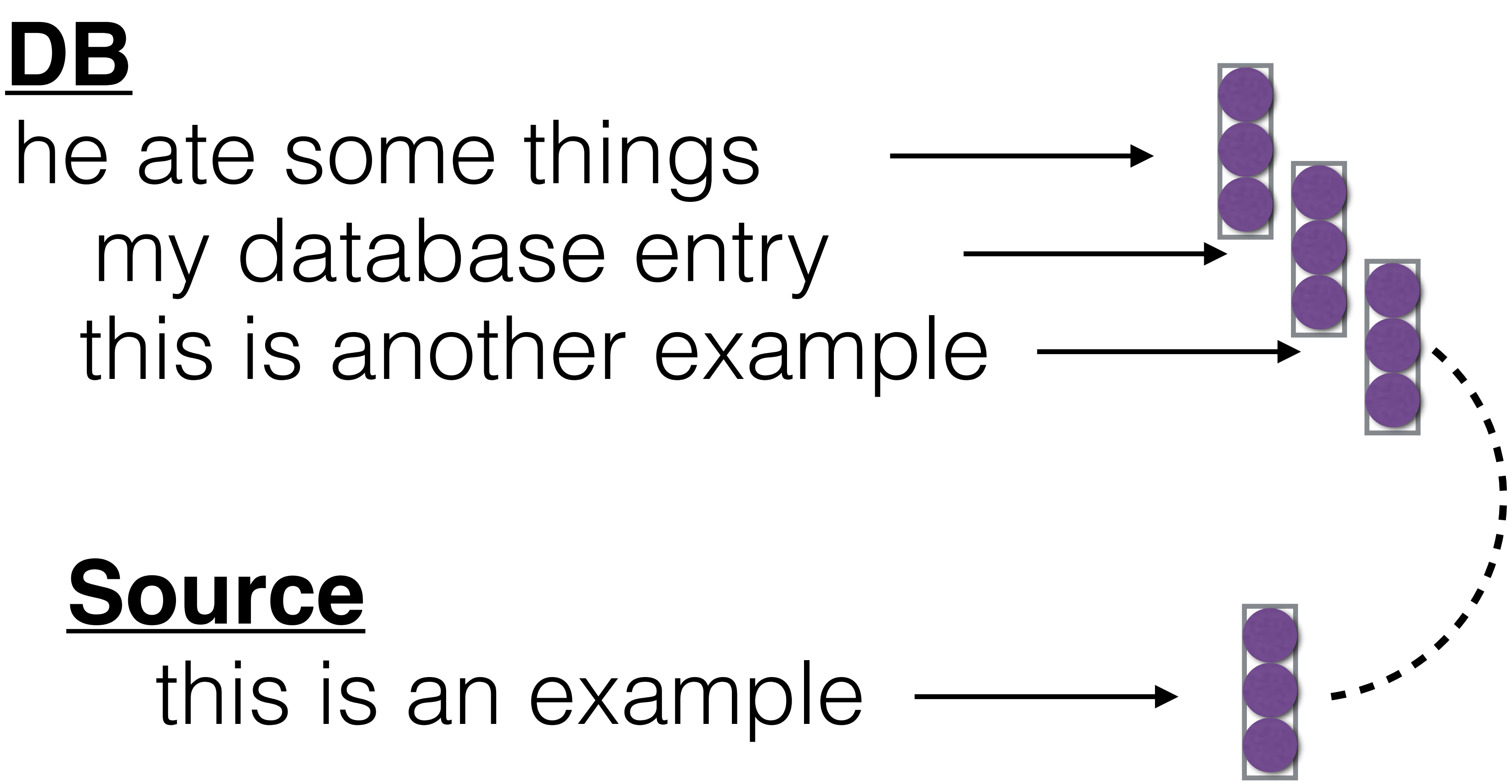
Retrieval

Retrieval Idea

- Given an input sentence, find something that matches
  - Text → text (Huang et al. 2013)
  - Text → image (Socher et al. 2014)
  - Anything to anything really!

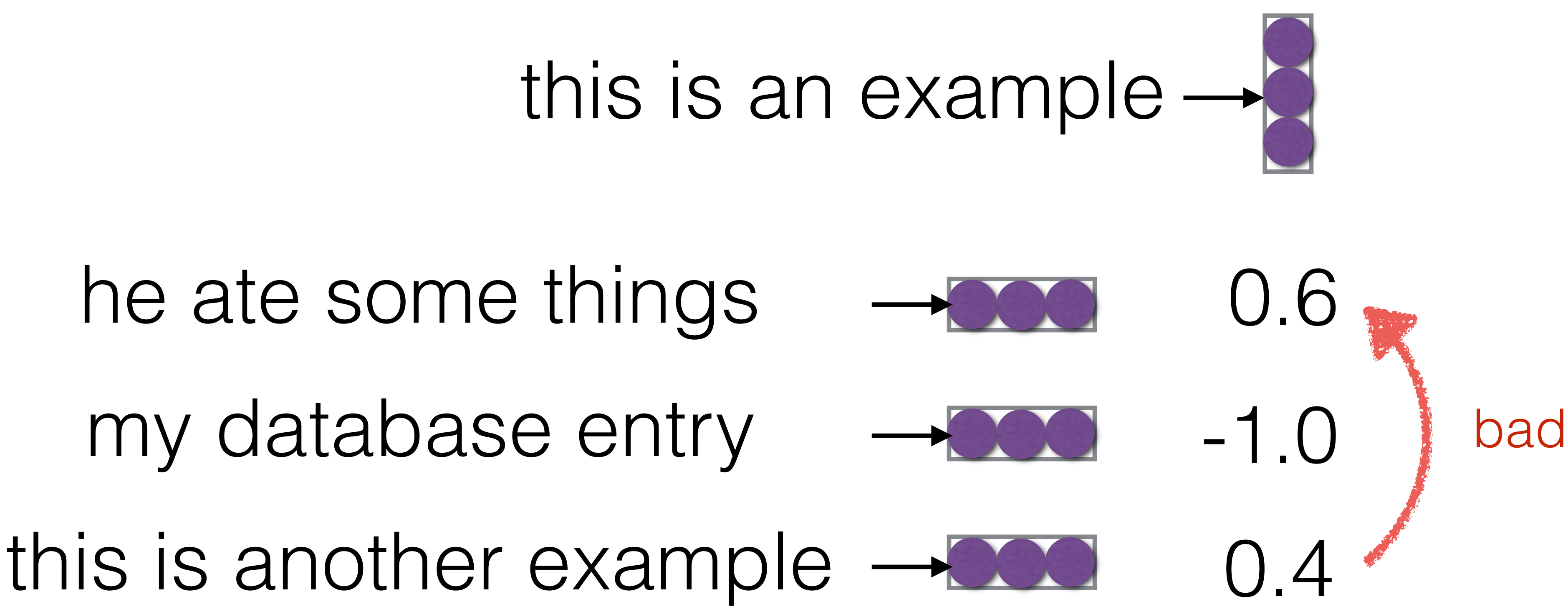
Basic Idea

- First, encode entire target database into vectors
- Encode source query into vector
- Find vector with minimal distance



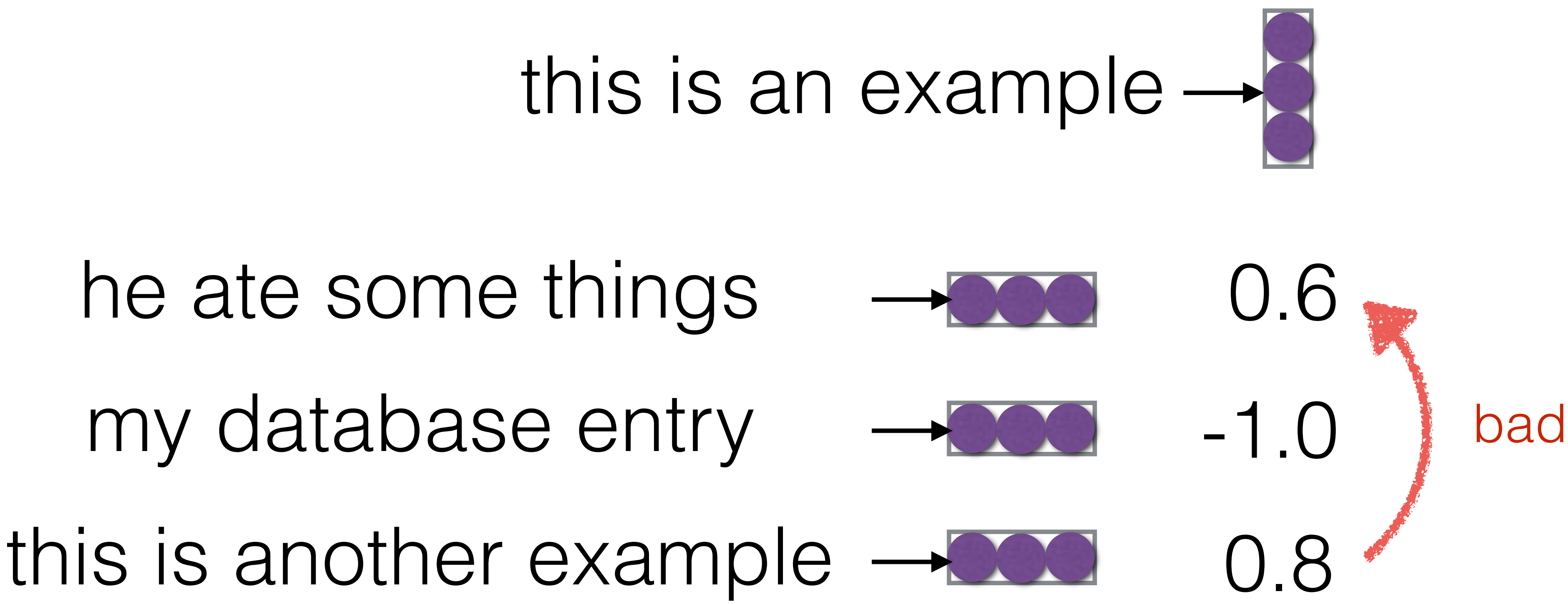
A First Attempt at Training

- Try to get the score of the **correct answer higher than the other answers**



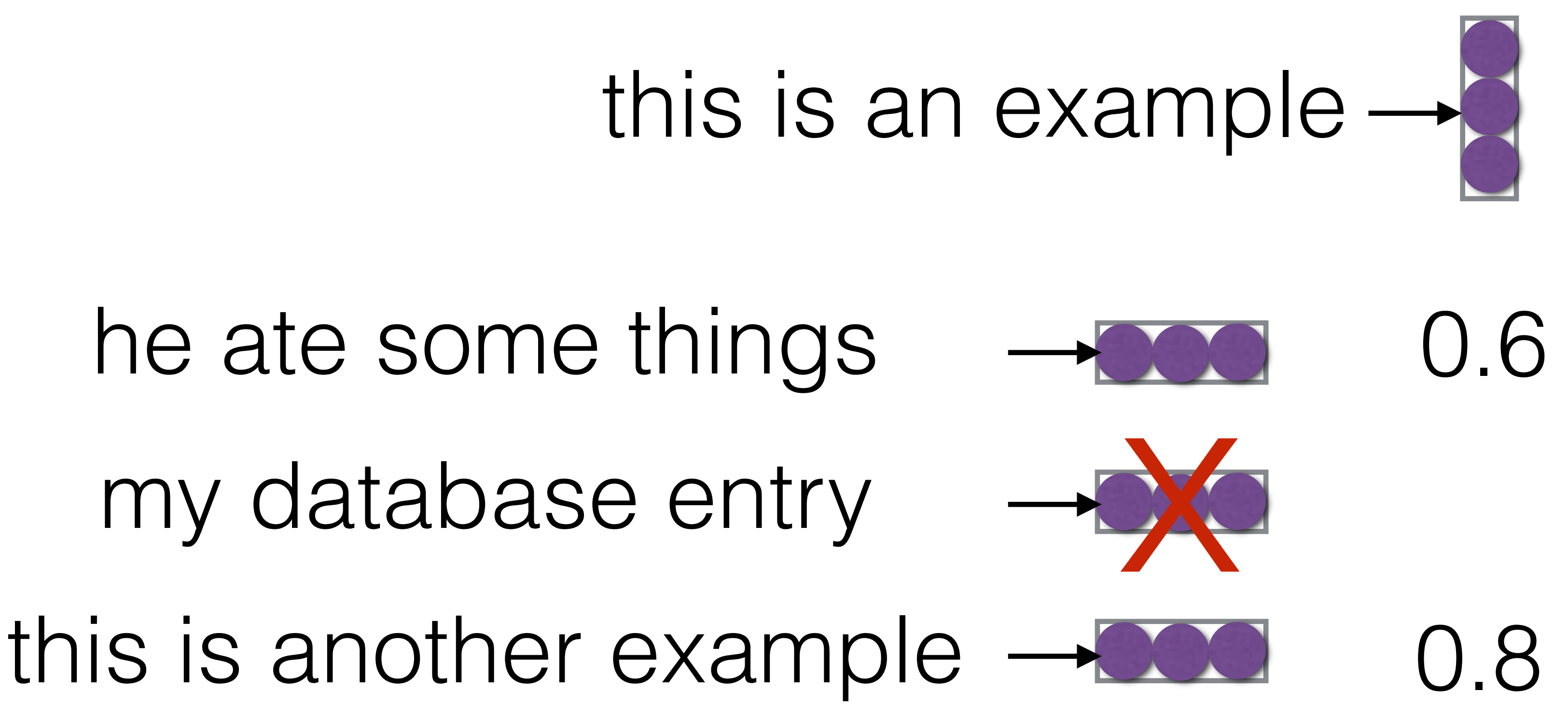
Margin-based Training

- Just “better” is not good enough, want to **exceed by a margin (e.g. 1)**



Negative Sampling

- The database is too big, so only use a **small portion of the database as negative samples**



Loss Function In Equations

$$\underbrace{L(x^*, y^*, S)}_{} = \sum_{x \in S} \max(0, \underbrace{1 + s(x, y^*)}_{\text{incorrect score plus one}} - \underbrace{s(x^*, y^*)}_{\text{correct score}})$$

correct input

correct output

negative samples

incorrect score plus one

correct score

Evaluating Retrieval Accuracy

- **recall@X:** “is the correct answer in the top X choices?”
- **mean average precision:** area under the precision recall curve for all queries

Let’s Try it Out
(on text-to-text)

`lstm-retrieval.py`

Efficient Training

- Efficiency improved when using **mini-batch training**
- **Sample a mini-batch**, calculate representations for all inputs and outputs
- Use **other elements of the minibatch as negative samples**

Bidirectional Loss

- Calculate the hinge loss in both directions
- Gives **a bit of extra training signal**
- **Free computationally** (when combined with mini-batch training)

Efficient Retrieval

- Again, the database may be too big to retrieve, use approximate nearest neighbor search
- Example: locality sensitive hashing

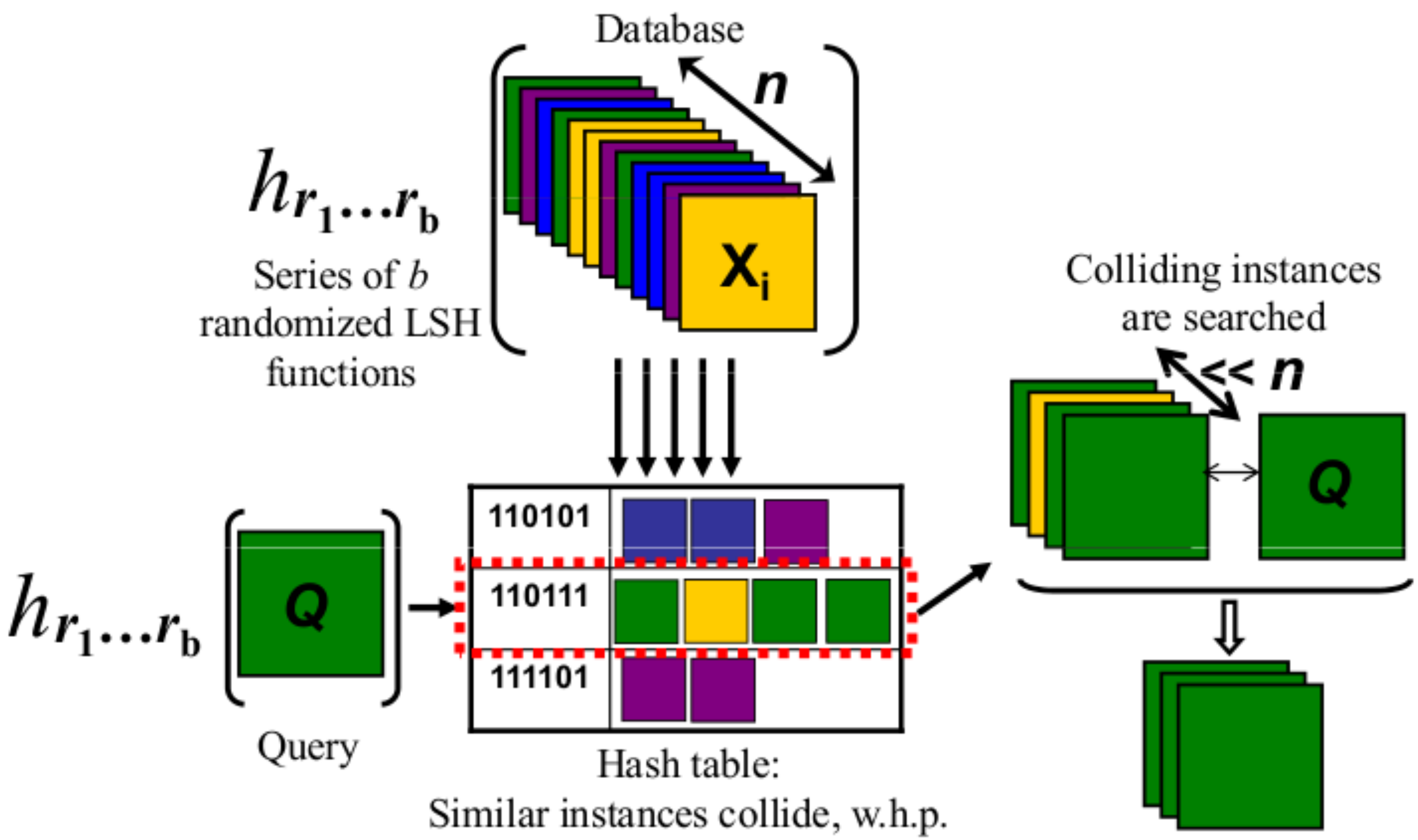


Image Credit: https://micvog.com/2013/09/08/storm-first-story-detection/

Data Example:
Flickr8k Image Retrieval
(Hodosh et al. 2013)

- Input text, output image
- 8000 images x 5 captions each
- Gathered by asking Amazon mechanical turkers to generate captions

Questions?