Paraphrase Detection in NLP

DataRobot

Yuriy Guts
ML Engineer

Paraphrasing

Any trip to Italy should include a visit to Tuscany to sample their exquisite wines.

Be sure to include a Tuscan wine-tasting experience when visiting Italy.

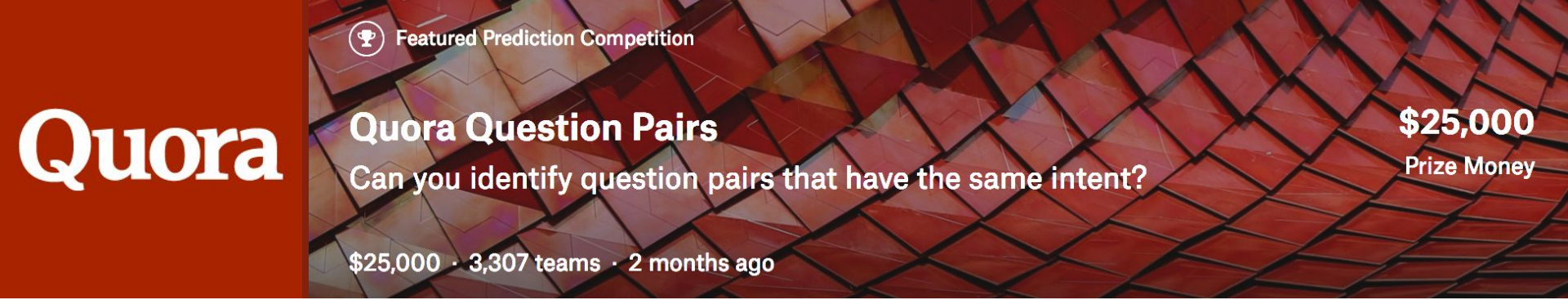
Paraphrase Identification

Where can I get very **professional and reliable** envelope printing service in Sydney?

Where can I get very **affordable** branded envelope printing service in Sydney?

Why are doctors always late?

Why doctors always make you wait for 15-20 minutes before they see you?



Data

Q1 (ID, Title) Q2 (ID, Title)

Target

Binary (metric: log-loss)

404,290 pairs

Training set

2,345,796 pairs

Test set

Challenges

Sometimes a single word matters

What are the best books on **IT** leadership?

What are the best books about leadership?

Will the Miami Heat win the NBA championship in **2011**?

Will the Miami Heat win the NBA championship in **2012**?

How do I lose **20** pounds?

How do I lose **15** pounds?

Sometimes there are almost no shared words

I am unable to talk to girls, leave being friendly with them. Why?

I am shy to talk to any woman because i get nervous and freaked out around them. What is the solution?

Is there a Quora user who have seen a UFO?

Have you seen an alien?

Sometimes ALL the words are shared

Is the Government of **Pakistan** encouraging **India** by not taking any real action against ceasefire violations?

Is the Government of **India** encouraging **Pakistan** by not taking any real action against ceasefire violations?

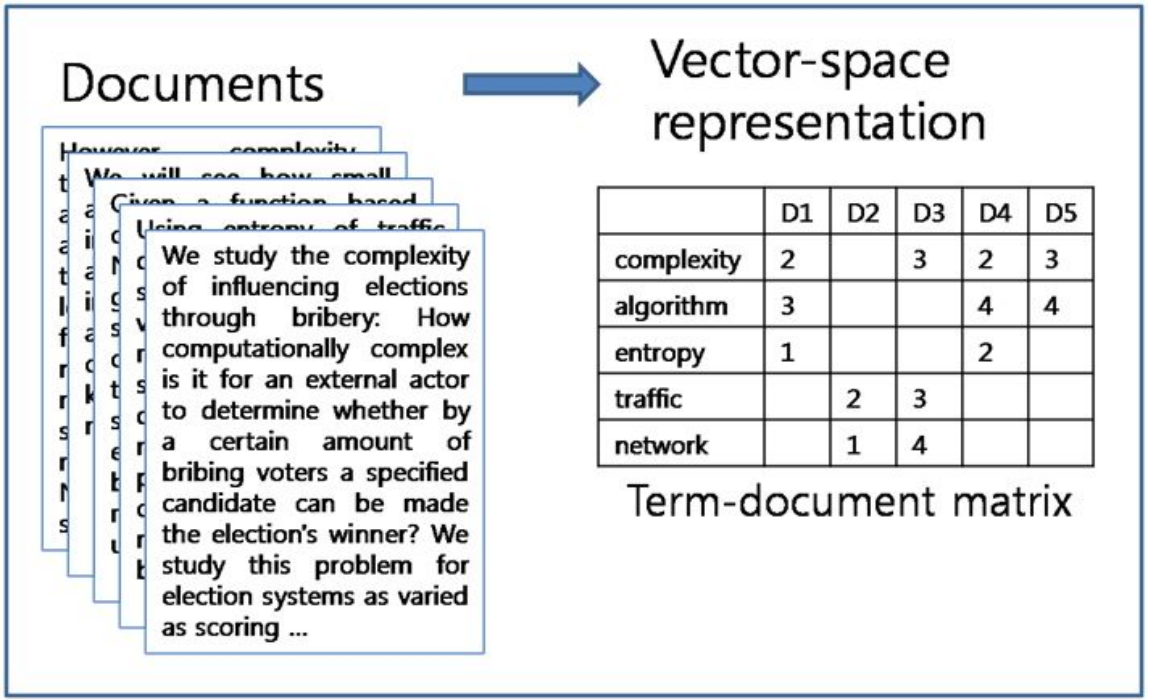
What is the most interesting thing we learned about **Portugal's** World Cup team in their match against **Germany**?

What is the most interesting thing we learned about **Germany's** World Cup team in their match against **Portugal**?

Approaches

1. Counting Stuff.
2. Information Theory.
3. Linguistics.
4. Deep Learning.

BoW Representations



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

TFIDF

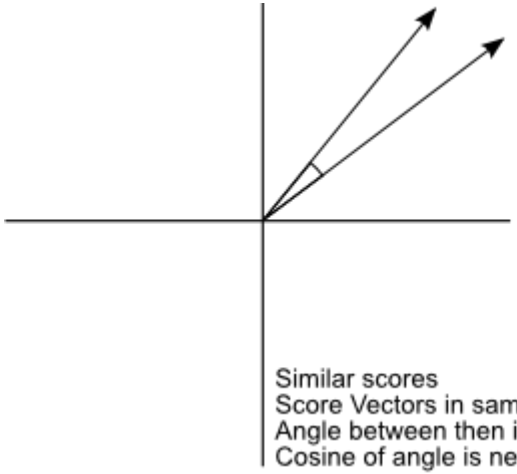
For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

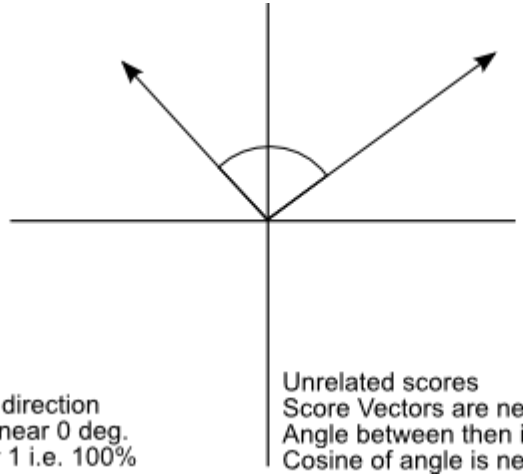
$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Cosine Similarity

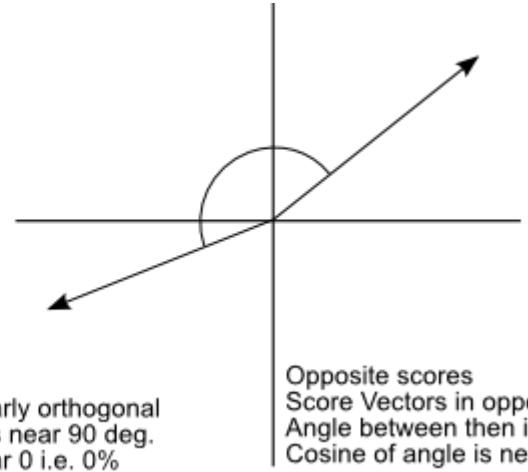
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$



Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%



Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%

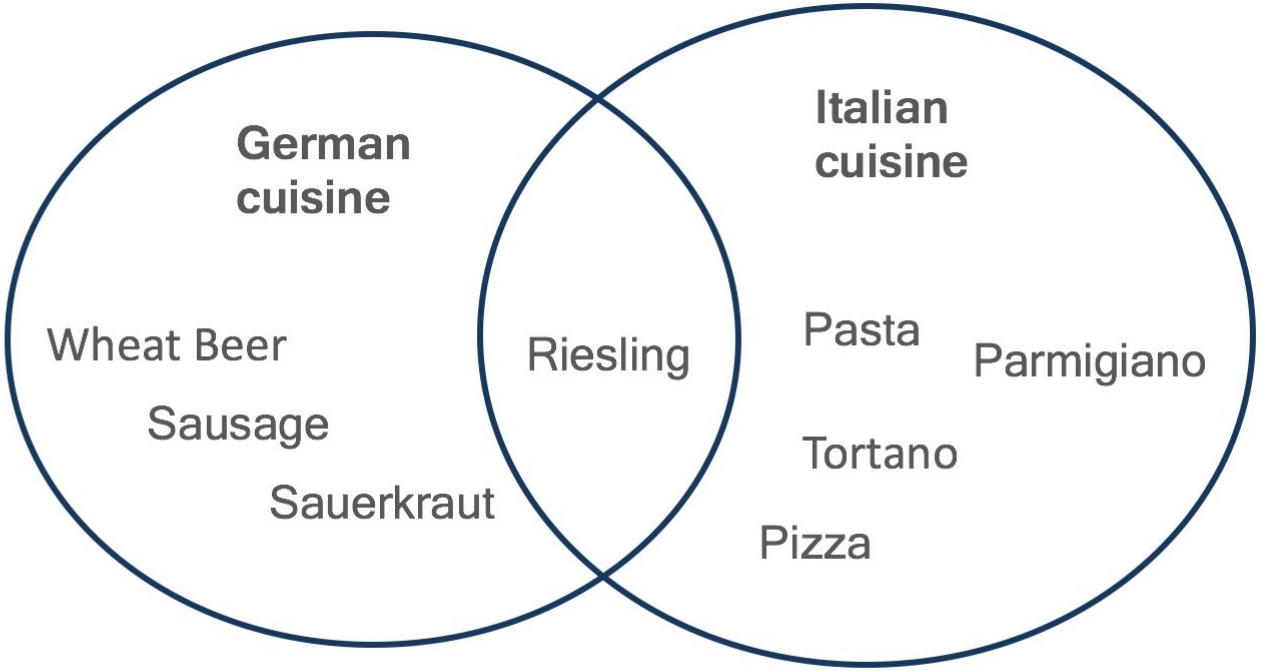


Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

Jaccard Similarity

$$\text{sim}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard similarity



$$\text{sim}(\text{🇩🇪}, \text{🇮🇹}) = \frac{1}{8}$$

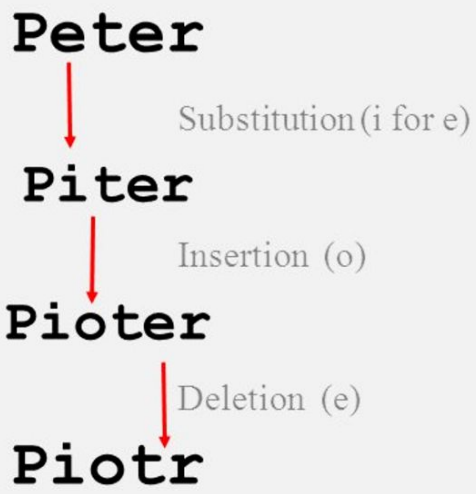
Edit Distances

How similar are the names "Peter" and "Piotr"?

Assume the following cost function

Substitution	1 Unit
Insertion	1 Unit
Deletion	1 Unit

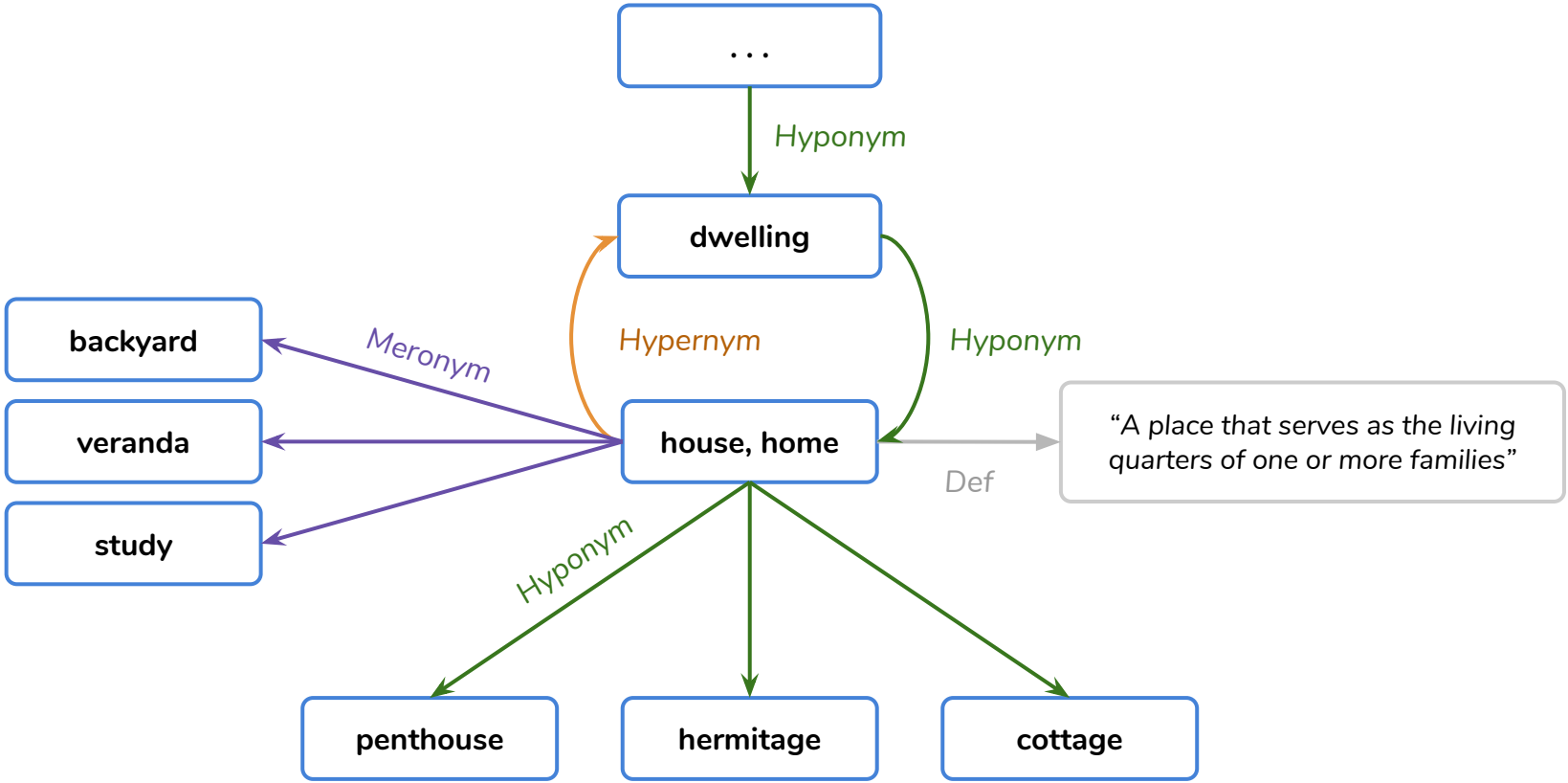
$D(\texttt{Peter}, \texttt{Piotr})$ is 3



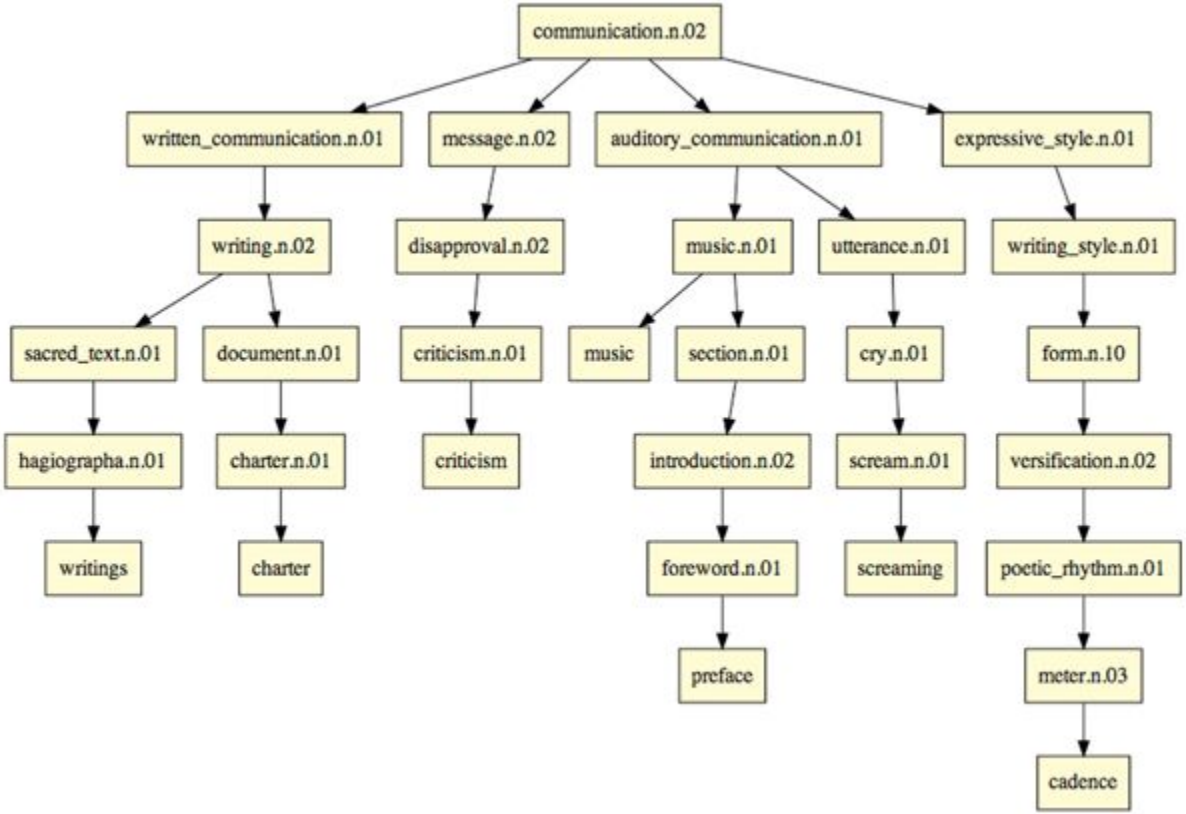
1. Levenshtein
2. Damerau-Levenshtein
3. Jaro
4. Jaro-Winkler

… … …

Lexical Databases and Ontologies



WordNet


communication.n.02
written_communication.n.01
message.n.02
auditory_communication.n.01
expressive_style.n.01
writing.n.02
disapproval.n.02
music.n.01
utterance.n.01
writing_style.n.01
sacred_text.n.01
document.n.01
criticism.n.01
music
section.n.01
cry.n.01
form.n.10
hagiographa.n.01
charter.n.01
criticism
introduction.n.02
scream.n.01
versification.n.02
writings
charter
foreword.n.01
screaming
poetic_rhythm.n.01
preface
meter.n.03
cadence

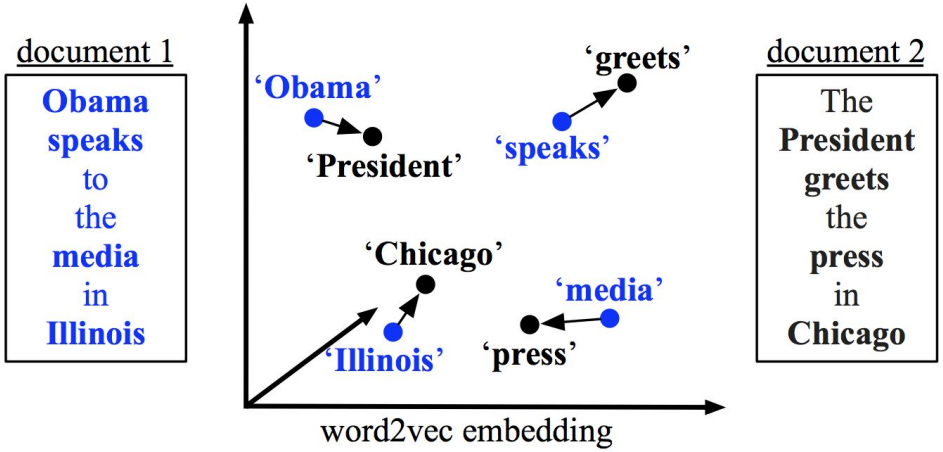
Distributed Hypothesis of Language

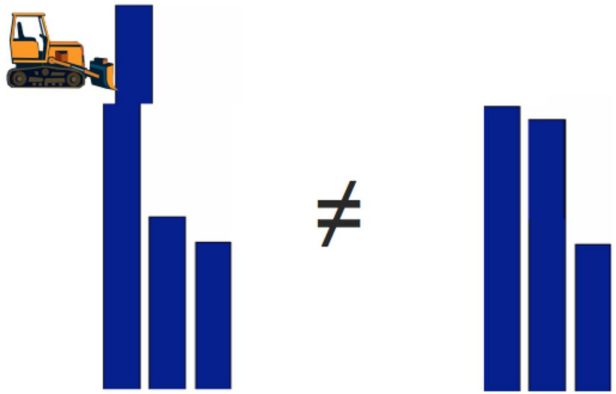
“The complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously”

John R. Firth. The technique of semantics.

Transactions of the Philological Society, 1935.

Word Mover's Distance (WMD)

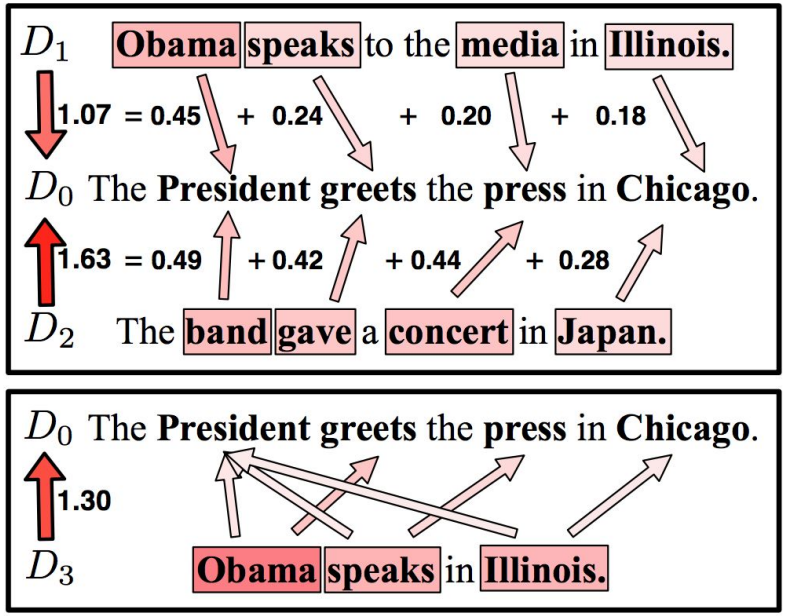




The minimum amount of "work" needed to transform document **1** to document **2**.
Inspired by Earth Mover's Distance, a well-studied transportation problem.

M. Kusner et al. "From Word Embeddings to Document Distances", 2015.
http://proceedings.mlr.press/v37/kusnerb15.pdf

WMD: Linear Optimization Problem



$$d_i = \frac{c_i}{\sum_{j=1}^{n} c_j}$$

nBOW frequency of the i-th word in the document

$$\mathbf{T}_{ij} \geq 0$$

"How much" of word ***i*** travels to word ***j***

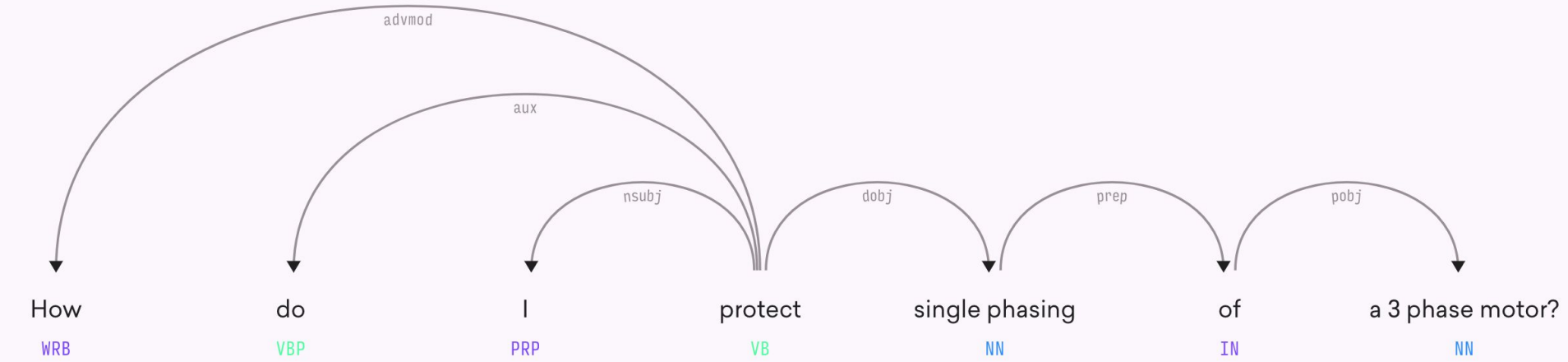
$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}.$$

M. Kusner et al. "From Word Embeddings to Document Distances", 2015.
http://proceedings.mlr.press/v37/kusnerb15.pdf

Morphology & Syntax Features

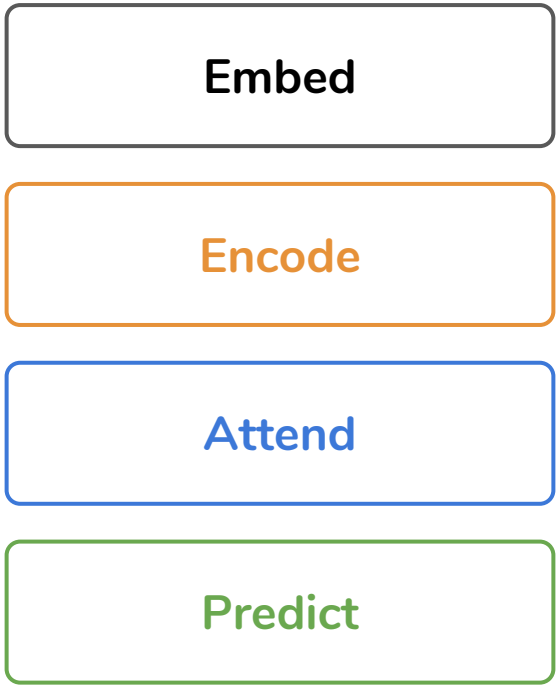


What is the most interesting thing we learned about Portugal GPE 's World Cup EVENT team in their match against Germany GPE ?

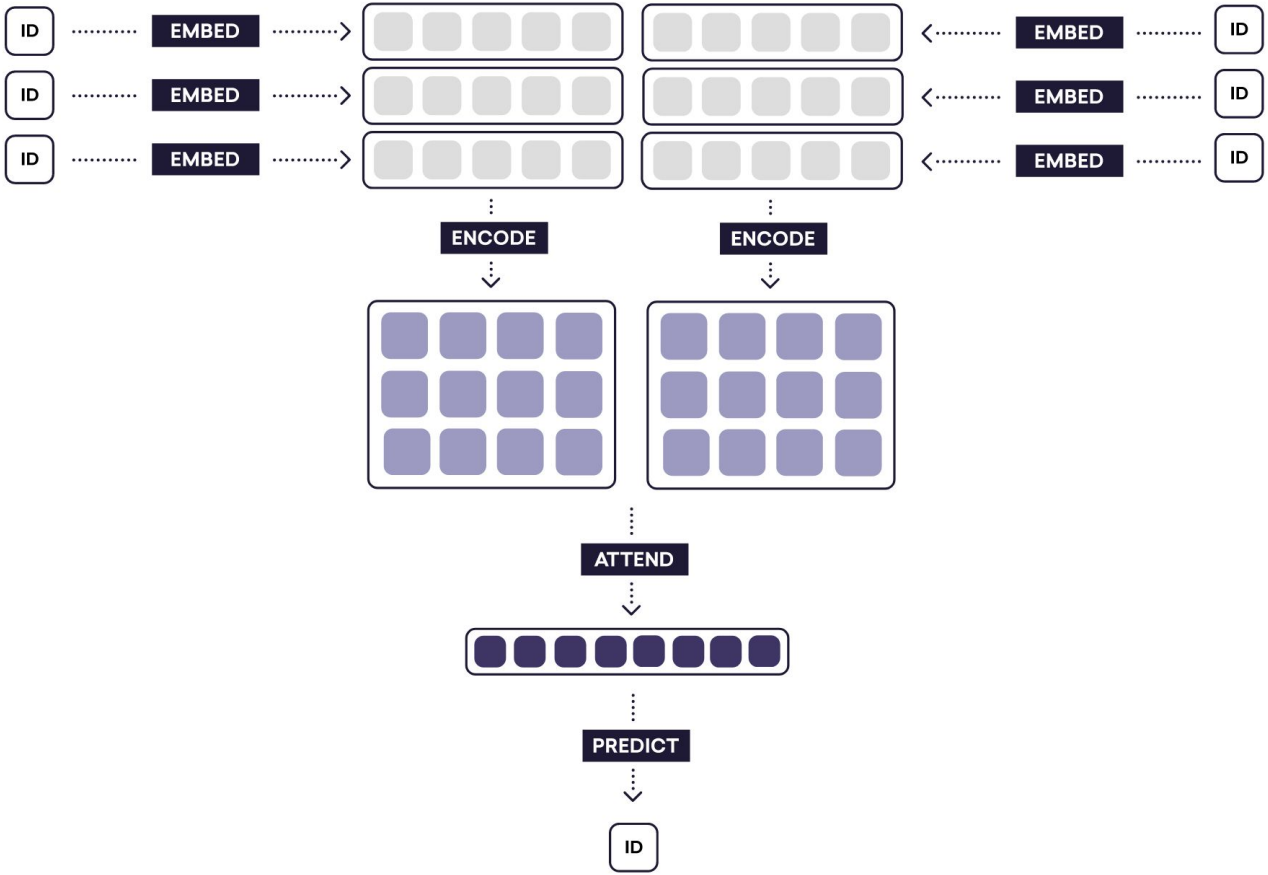
https://demos.explosion.ai/displacy/

Architectural Principles

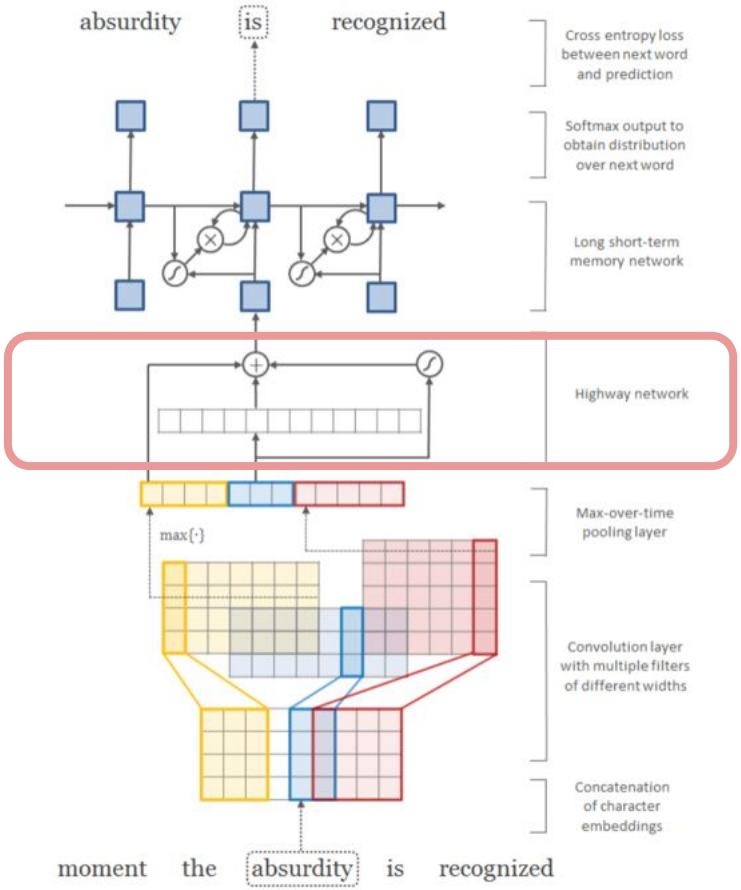
for State-of-the-Art Neural NLP


Embed
Encode
Attend
Predict

State-of-the-Art NLP Pipeline



Recurrent Highway Networks

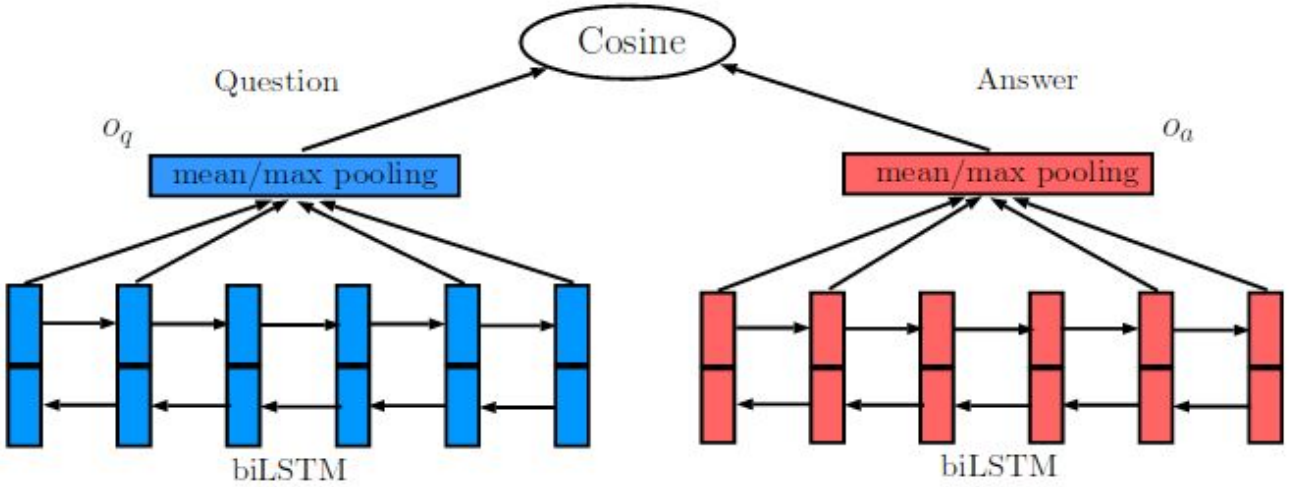


Highway layer:

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot (1 - T(\mathbf{x}, \mathbf{W_T}))$$

Srivastava et al. “*Recurrent Highway Networks*”, *2015*
https://arxiv.org/pdf/1607.03474.pdf

Two Input Documents? Siamese Network



Run the same encoding step for every input.

Share encoder weights, don't learn W_{E1} and W_{E2} separately.

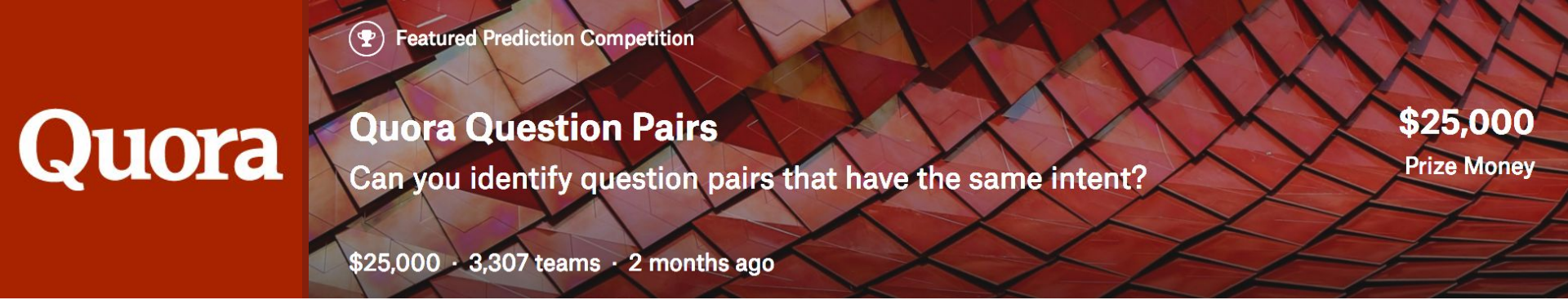
Distance Learning, Contrastive Loss

$$L(W, Y, \vec{X}_1, \vec{X}_2) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

Penalize similar pairs by a monotonically **increasing** function of their learned distance.

Penalize different pairs by a monotonically **decreasing** function of their learned distance.

Hadsell, Chopra, LeCun “Dimensionality Reduction by Learning an Invariant Mapping”, 2006.
http://yann.lecun.com/exdb/publis/pdf/hadsell-chopra-lecun-06.pdf



Data

Q1 (ID, Title) Q2 (ID, Title)

Target

Binary (metric: log-loss)

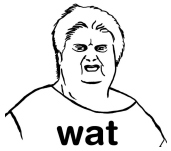
404,290 pairs

Training set

2,345,796 pairs

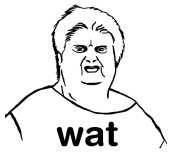
Test set

Sometimes the labeling is just plain wrong



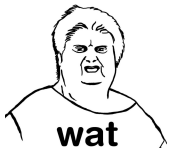
How can I prevent pneumonia?

How do you prevent pneumonia?



What is the car in this picture?

What car is in this picture?



How do I protect single phasing of a 3 phase motor?

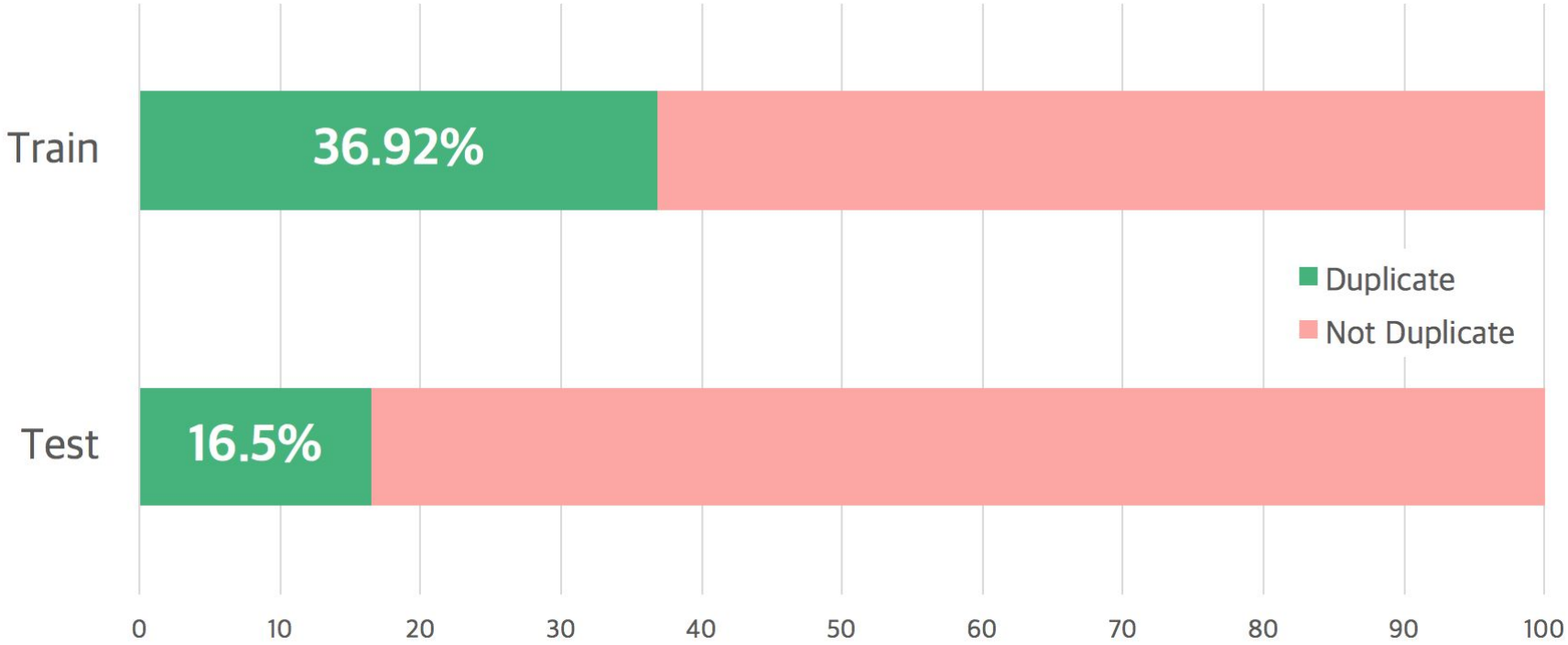
What is single phase and 3 phase?

People misspell. A lot.

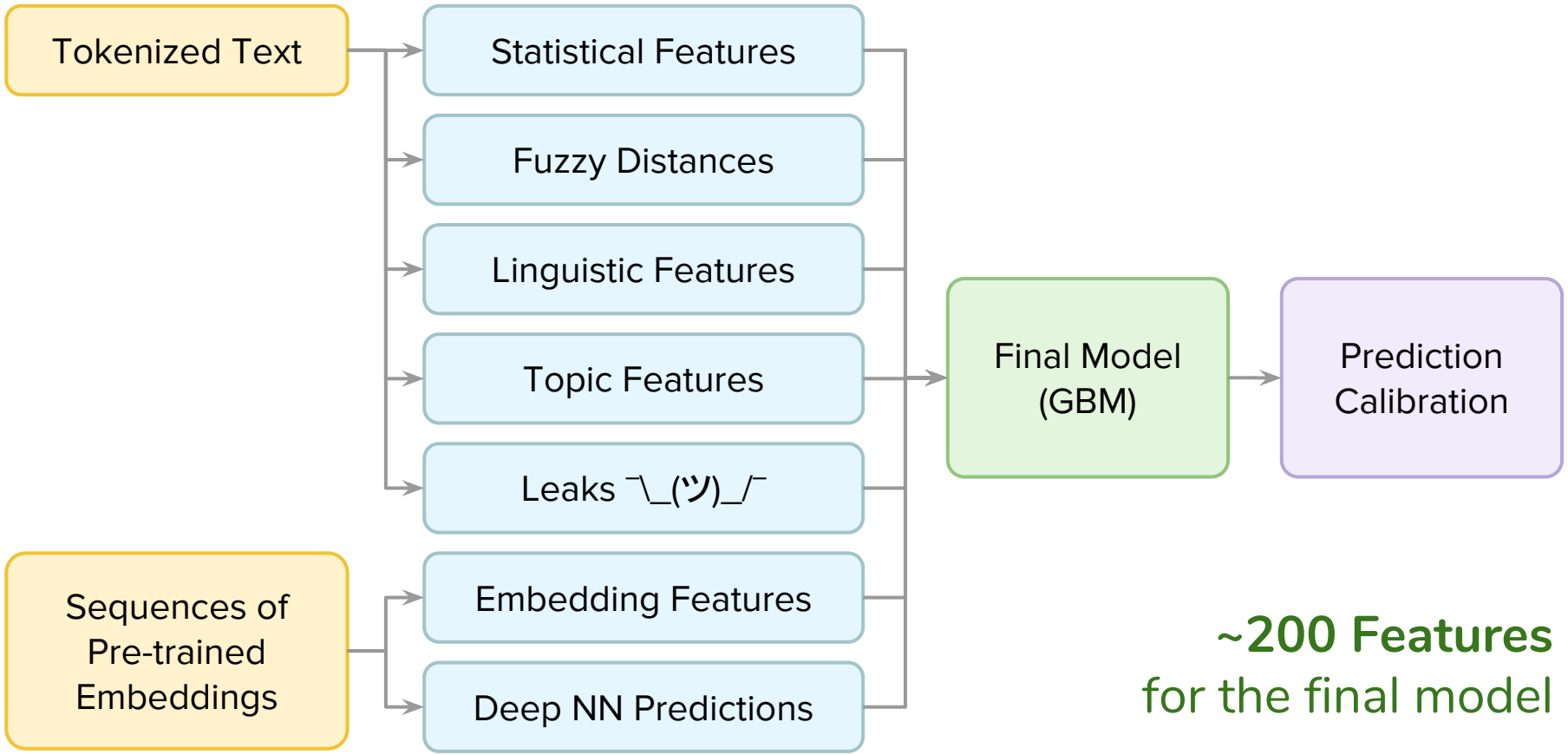
demonitization	demonitozation	masturbution	masturbtion
demonitizing	demonetiztion	masturation	mastburation
demonetising	demoneitisation	mastrubating	masturabate
demonitisation	demoneticzation	masturbuation	mastrubration
demonitize	demonitizition	masterbuting	mastubrate
demonatisation	demonetisations	mastubate	masturbute
demonitazation	demonitasation	mastribution	mastribute
demonitesation	demonestisation	masturburation	masterbuate
demoneytisation	demotenization	mastrabutation	mastubating
demonestation	demonitised	mastubation	
demonitising	demonsation		

100,000+
spelling corrections

Different target balance for Train and Test



Final Solution (Simplified)



~200 Features
for the final model

Original Datasets
train.csv test.csv

Auxiliary Data

- Spelling corrections
- Curated stopword list

- Spellcheck
- Expand shortened words
- Spell digits
- Tokenize
- (sometimes) Lowercase
- (sometimes) Remove stopwords

Encode questions as padded sequences

Dataset augmentation:
Flip the pairs and append the mirrored input to the original one.

When predicting, make two mirrored predictions and average them.

Simple summary statistics (word/char lengths, diffs, ratios)

TF-IDF distances

Fuzzy string matching scores

Jaccard distances over character n-grams

Bag-of-POS-tags distances
Bag-of-NER-tags distances

LDA topic distances

WordNet path similarity

Phrase embedding distances

Word Mover's Distance

Magic features
PageRank on question graph

Magic features
Question occurrence counts

Magic features
Co-occurrence matrix distances

A subset of the magic features is concatenated with the intermediate feature layer

FastText embeddings
pre-trained on Wikipedia

Neural Network
Multi-Layer Perceptron

Neural Network
Word-CNN

Neural Network
Bidirectional word-LSTM

Neural Network
Siamese LSTM with attention
Contrastive loss

Out-of-fold predictions
MLP

Out-of-fold predictions
Word-CNN

Out-of-fold predictions
Word-Bi-LSTM

Out-of-fold predictions
Siamese word-LSTM w/attention

Word2Vec embeddings
pre-trained on Google News

3rd-party features
@abhishek

3rd-party features
@dasolmar

3rd-party features
@mephistopheles

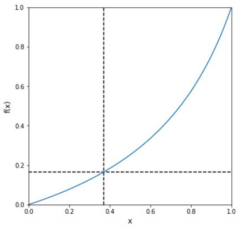
3rd-party features
@puneetsl

LightGBM
Averaged predictions
from K-fold CV

Recalibrate for a different prior on the test set
(36.9% → 16.5%)

$$\alpha = \frac{p_{test}}{p_{train}} \qquad \beta = \frac{1 - p_{test}}{1 - p_{train}}$$

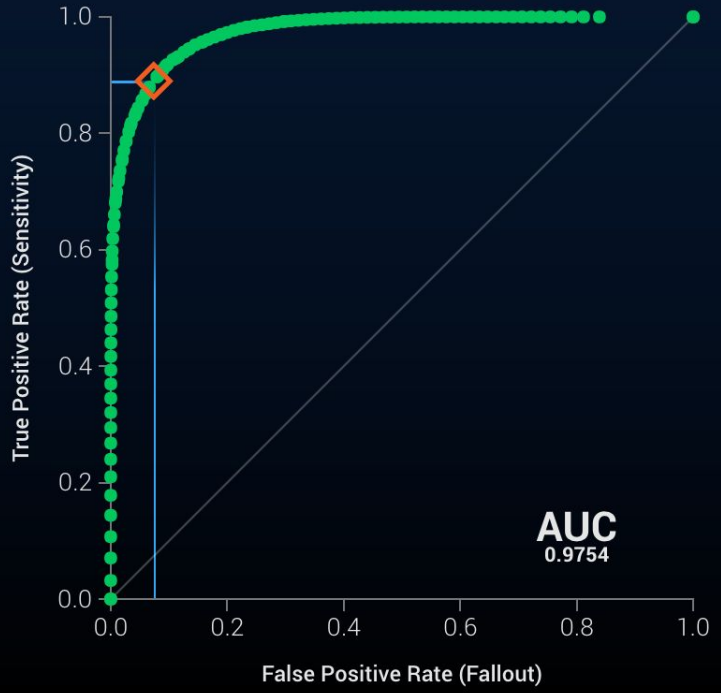
$$\hat{y}'_{test} = \frac{\alpha \cdot \hat{y}_{test}}{\alpha \cdot \hat{y}_{test} + \beta \cdot (1 - \hat{y}_{test})}$$



F1 Score	True Positive Rate (Sensitivity)	False Positive Rate (Fallout)	True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy
0.8829	0.89	0.0738	0.9262	0.876	0.935	0.9129

ROC Curve

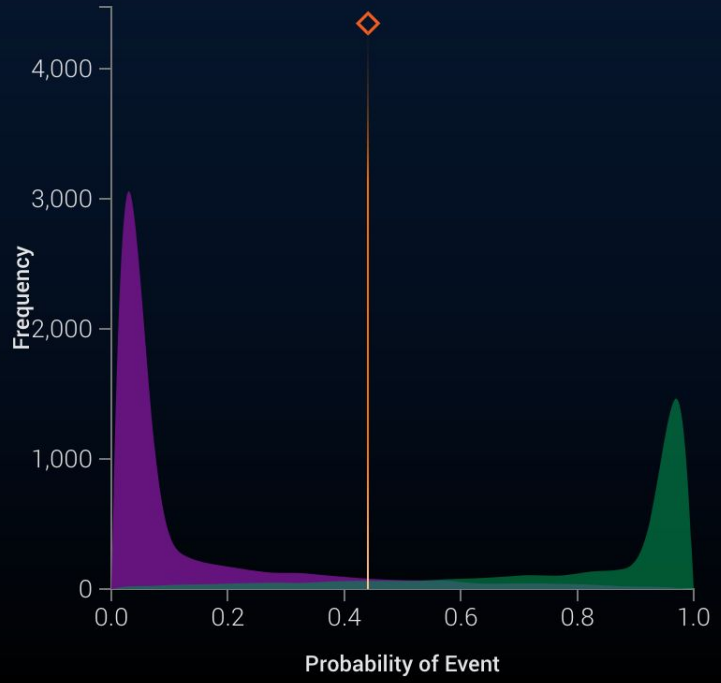
Data Source: Holdout



Prediction Distribution

Threshold (0-1): 0.4409

Frequency



facebook.com/yuriy.guts

github.com/YuriyGuts/kaggle-quora-question-pairs