University of Colorado Boulder

Removing the Training Wheels:
A Coreference Dataset that Entertains Humans and Challenges Computers

ANUPAM GUHA, MOHIT IYYER, DANNY BOUMAN, JORDAN BOYD-GRABER

Dependence on Newswire

- CL systems for parsing, POS, and just about everything built on newswire
  - news articles, blogs, newsgroups, and transcripts of broadcast news.
- We call such data “newswire”
- **This talk**: Coreference is **deceptively easy** on Newswire

The New York Times

REUTERS

WALL STREET JOURNAL

What this work is and why you should care

- Newswire monoculture offers diminishing returns in coreference resolution
- Newswire lacks complexity of real world coreference.
- New coreference dataset
- more challenging, more fun coreference
- Quiz bowl!

Coreference Resolution is a challenging problem

Monsieur Poirot assured Hastings that he ought to have faith in him

- Introduction
- Dataset
- System
- Results and Conclusion

Coreference Resolution is a challenging problem

Monsieur Poirot assured Hastings that he ought to have faith in him




Limitations of Newswire Data

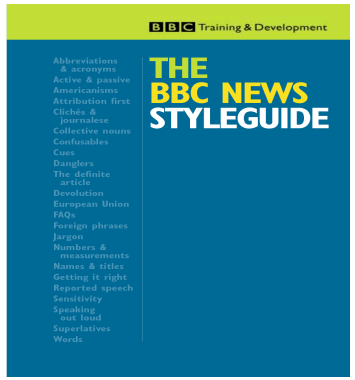
- Sparse
- Singletons
- Domain mismatch

"Systems analysing correctly about 90% of the sequences from a journalistic corpus can sustain a decrease of performance up to 50% on more informal texts. Journalistic redactional constraints often introduce person names with titles (President Chirac) or trigger words (Mister Chirac) . ***This way of writing is not systematic in informal texts****" – (Poibeau and Kosseim, 2001)*

Quotes from journalistic style guides

"**Capitalize *Satan***"

"Always use a person's first and last name the first time they are mentioned in a story."



"The key to good writing is **simple thoughts** simply expressed. Use short sentences and short words."

"Anything **flamboyant**, such as a subordinate clause, is **a potential barrier to understanding**."

Limitations of Newswire Data

- Writers cannot assume that their readers know all participants in the story.
- These constraints make for easy reading and easy coreference resolution.
- Unlike “journalistic” coreference, everyday coreference relies heavily on inferring the identities of entities in language, which requires substantial world knowledge.

Quiz Bowl: A Game of Human Coreference

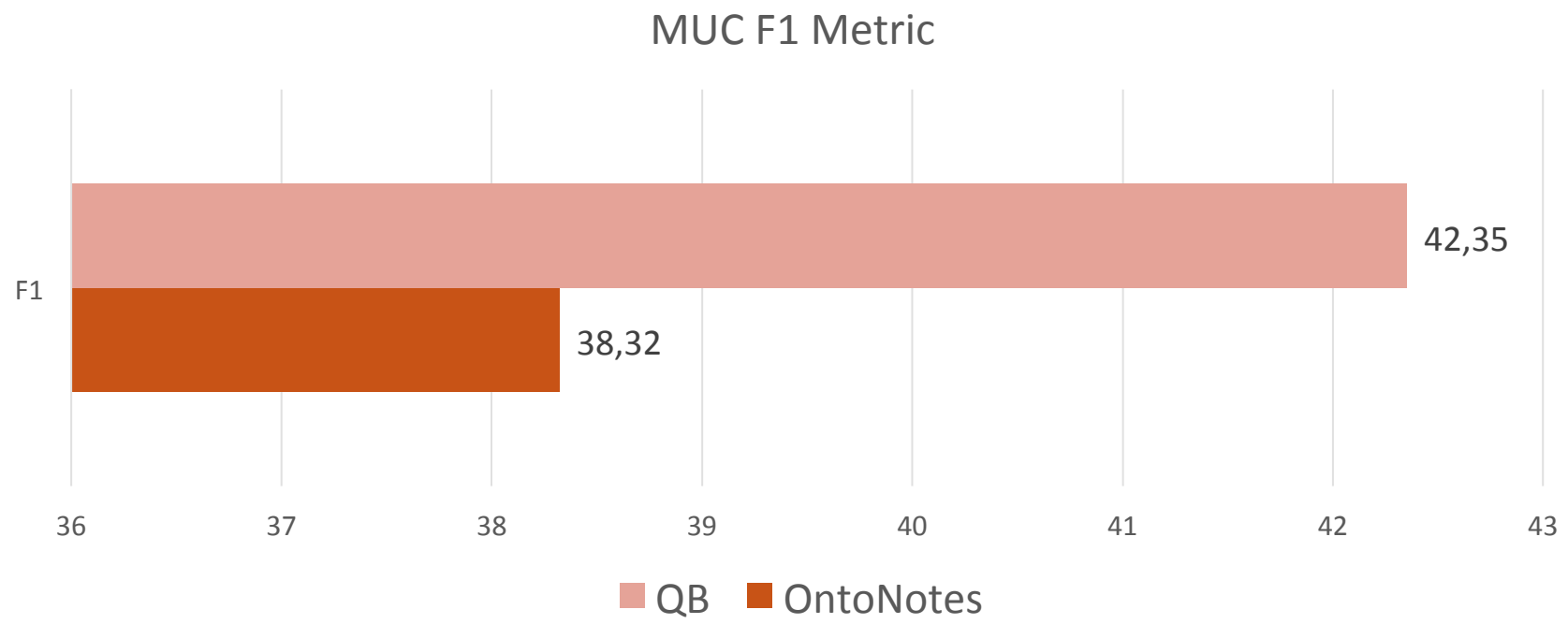
Solve this question. Also, keep track of the coreferent phrases

The tailor Mustafa in **this story** is led by a slave girl to her owner's house, where he was to sew back together the four pieces of a corpse. A slave conceals the murder and suggests that Cassim was ill. One antagonist finds the house of Cassim's brother and marks it so he could come back later to murder him, but the slave Morgiana foils his plan. For 10 points, name **this story** about a cave full of treasure which is opened by saying Open, sesame, a story told by Scheherezade in One Thousand and One Night.

Quiz bowl is a source of dense coreference

{This entity} made a military pact known as the Tohopesate, and {it} was harassed by forces of Klaus Stortebeker, the leader of the Likedeelers, which were successors of pirates known as the Victual Brothers. After the humiliating Peace of Vordingborg, {this group} forced Valdemar IV of Denmark to sign the Treaty of Stralsund. The Peterhof in Novgorod and Steelyard in London are notable examples of Kontores established by {this polity}, although {its} power faded as the Swedish Empire lessened the importance of cities such as Visby and Lubeck. For 10 points, name {this Northern European trading league} which controlled the Baltic Sea.

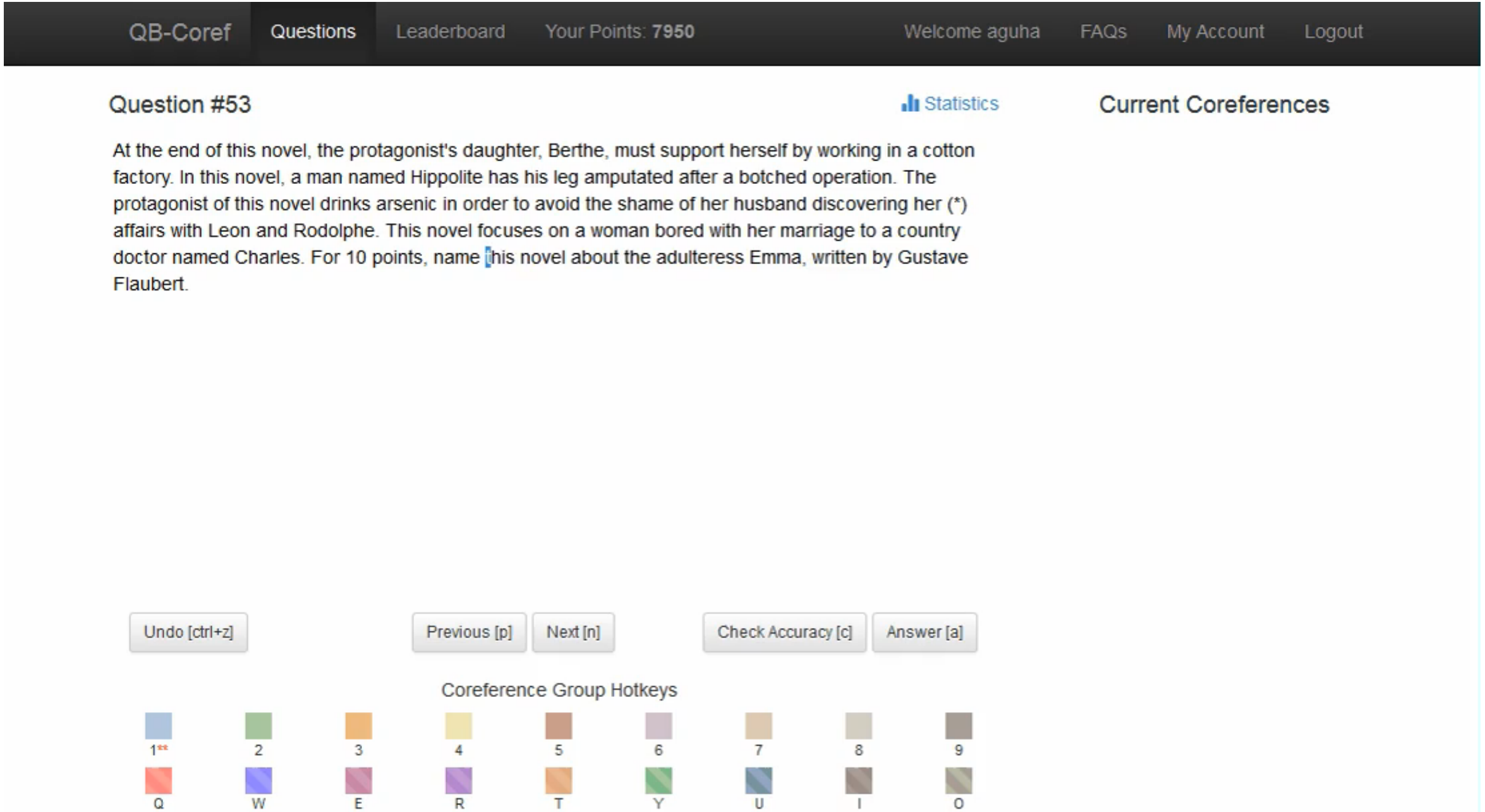
Quiz bowl is hard to solve with models trained on newswire data



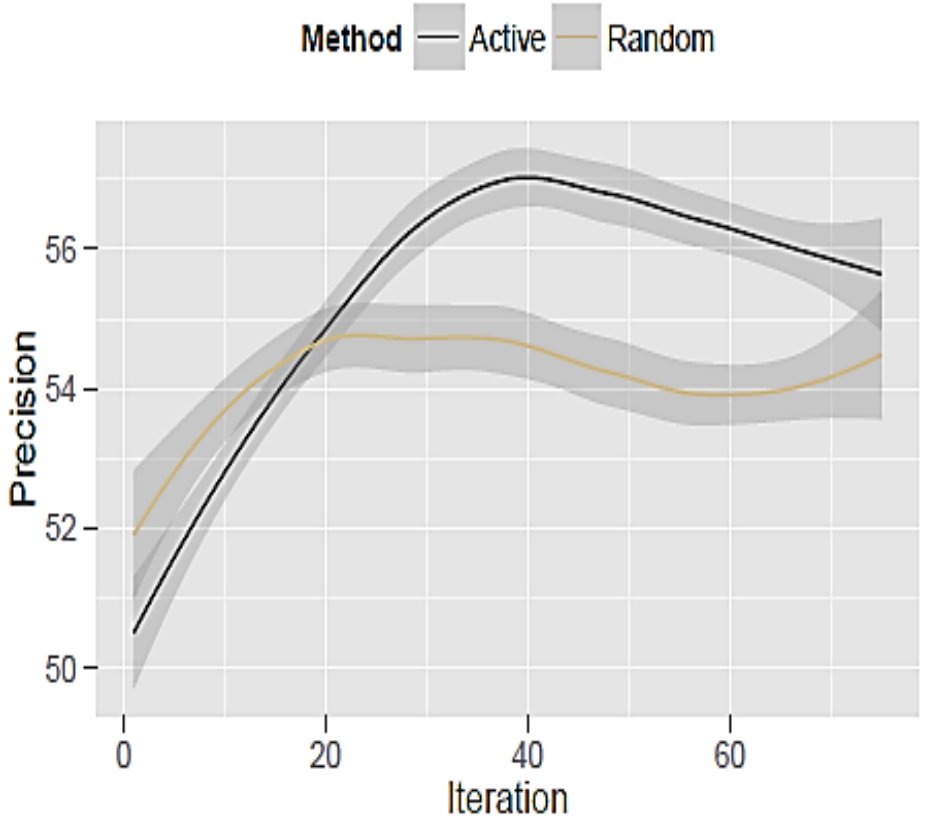
Results of evaluation a quiz bowl dataset with models trained on (Durrett and Klein, 2013) using OntoNotes data vs quiz bowl data

Creating our dataset
A webapp to annotate quiz bowl questions with coreference tags

- Introduction
- Dataset
- System
- Results and Conclusion


QB-Coref
Questions
Leaderboard
Your Points: 7950
Welcome aguha
FAQs
My Account
Logout
Question #53
Statistics
Current Coreferences
At the end of this novel, the protagonist's daughter, Berthe, must support herself by working in a cotton factory. In this novel, a man named Hippolite has his leg amputated after a botched operation. The protagonist of this novel drinks arsenic in order to avoid the shame of her husband discovering her (*) affairs with Leon and Rodolphe. This novel focuses on a woman bored with her marriage to a country doctor named Charles. For 10 points, name this novel about the adulteress Emma, written by Gustave Flaubert.
Undo [ctrl+z]
Previous [p]
Next [n]
Check Accuracy [c]
Answer [a]
Coreference Group Hotkeys
1**
2
3
4
5
6
7
8
9
Q
W
E
R
T
Y
U
I
O

Active Learning



- The original dataset has 7,000 literature questions, and we want to tag only the useful ones.
- We use active learning for document-level coreference rather than at the mention level.
- We use voting sampling.

Our dataset

- The webapp was advertised to quiz bowl players before a national tournament and attracted competent annotators.
- Annotation rubric was provided.
- Documents were either tagged by four users with a degree of similarity and verified by authors (150 docs), or tagged by the authors in committee (250 docs).

Statistics of our dataset

Number of …	Quiz bowl	OntoNotes
Documents	400	1667
Tokens	50,347	955,317
Mentions	9,471	94,155
Nested Mentions	2,194	11,454
Singletons	2,461	0

Interesting examples from our dataset

Example which needs **understanding of thematic content**

[One author from [this country]$_1$]$_8$ describes [a man]$_9$ who sells [his]$_9$ family's land to [The Emperor]$_{10}$ in [The Silent Cry]$_{11}$. In addition to [Oe]$_8$, [a poet from [this country]$_1$]$_2$ wrote about the sea at Sado. In addition to [the author of [Narrow Road to the Deep North]$_7$]$_2$, [another author from [this country]$_1$]$_3$ described [Lady Aoi]$_6$'s relationship with [the title prince]$_5$ in [The Tale of [Genji]$_5$]$_4$. For 10 points, name [this homeland of [Basho]$_2$ and [Lady Murasaki]$_3$]

Interesting examples from our dataset

Example which needs **world knowledge** to solve

[The protagonist of [one of [this man]$_1$'s works]$_2$]$_4$ erects a sign claiming that [[that story]$_2$'s title figure]$_3$ will fly to heaven from a pond. Identify [this author of [[Dragon]$_3$: the Old Potter's Tale]$_2$]$_1$

Interesting examples from our dataset

Example with mentions having long text spans

[Stephen Hero]$_4$ was [a publication that contained large portions of [another of [his]$_1$ works]$_3$]$_4$, while [Humphrey Chimpdon]$_6$ appears in [another work]$_5$, [Finnegan's Wake]$_5$. "Eveline", "Araby", and "The Dead" appear in [one of his works depicting the life of [residents of the title city]$_2$]$_2$, while [another work about [Leopold Bloom]$_8$]$_7$ contained allusions to [a work by [Homer]$_{10}$]$_9$. For 10 points, identify [this Irish author of [A Portrait of the Artist as a Young Man]$_3$, [Dubliners]$_2$, and [Ulysses]$_7$]$_1$.

Interesting examples from our dataset

Example in which a mention is coreferent to a mention nested in it

Identify [this poem in which [Yevgeniy]$_2$ imagines [the title equestrian statue of [Peter the Great]$_4$]$_1$ chasing [him]$_2$ through the streets of St. Petersburg]$_1$, [a poem by [Alexander Pushkin]$_3$]$_1$.

Our end to end system:
A mention detector and a pairwise classifier

- Introduction
- Dataset
- System
- Results and Conclusion

A Simple Mention Detector

- Sequence labelling, as detecting mentions is detecting starts and stops.
- We use the MALLET implementation of CRF. We use nested BIO markers.
- The features are 1) the token itself, 2) the part of speech, 3) the named entity type, and 4) a dependency relation concatenated with the parent token.
- We obtain 76.1% precision, 69.6% recall, and 72.7% F1 measure.

Features and BIO labels

Identify	VB	*	root-ROOT	**lvl0**
this	DT	*	det-poem	**lvl1-Start**
poem	NN	*	dobj-identify	**lvl1**
in	IN	*	prep-imagines	**lvl1**
which	WDT	*	pobj-in	**lvl1**
Yevgeniy	NNP	**PERSON**	nsubj-imagines	**lvl2-Singleton**
imagines	VBZ	*	rcmod-poem	**lvl1**
the	DT	*	det-statue	**lvl2-Start**
title	NN	*	nn-statue	**lvl2**
equestrian	NN	*	nn-statue	**lvl2**
statue	NN	*	dobj-imagines	**lvl2**

A Simple Coreference Classifier

Given a mention pair A and B, the features extracted from them which the LR model uses are:

- A, B →binary vector over vocabulary * POS taglist
- A, B + n token window → binary vector over vocabulary * POS taglist
- Number of tokens between A and B
- Number of sentences between A and B
- Cosine similarity of the vector representations of the mentions

Word2vec and GloVe

For B: the title equestrian statue of Peter the Great

the: [0.24 0.32 … 0.28]

title: [0.11 0.49 … 0.16]

...

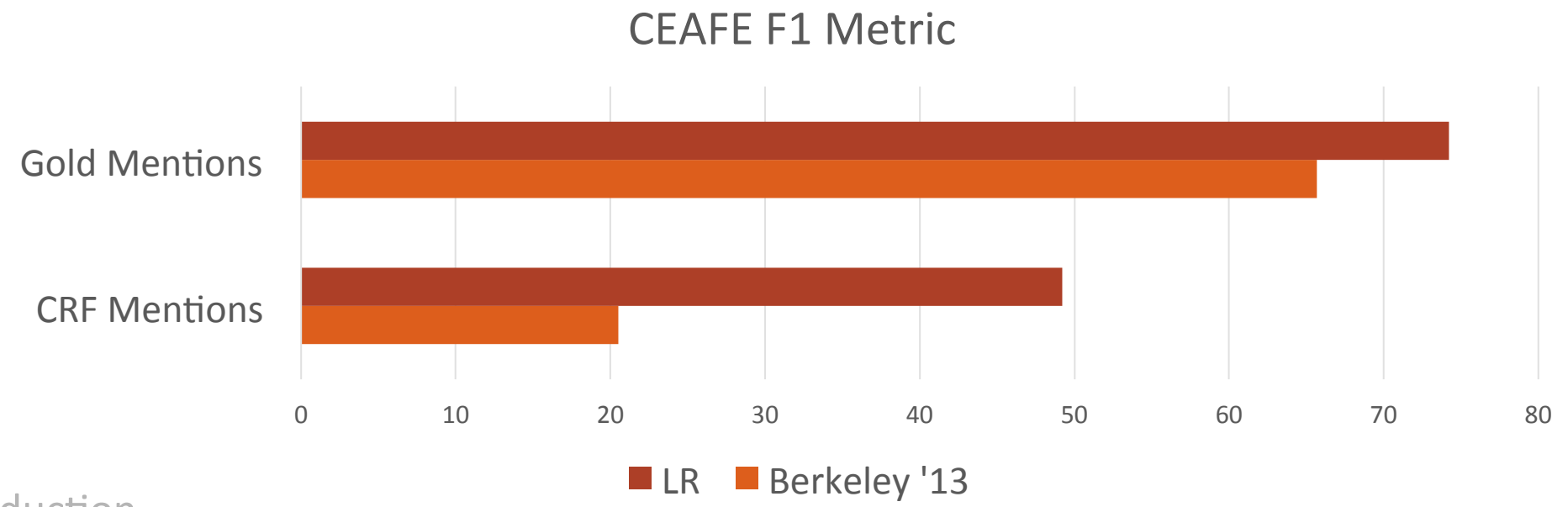
Great: [0.63 0.78 … 0.24]

Average : [0.86 0.24 … 0.71]

For A: [0.38 0.69 … 0.27]

Cosine distance of A and B

The simple LR model matches the state-of-the-art!



- Introduction
- Dataset
- System
- Results and Conclusion

Why does this happen?

We hypothesise

- Some of Berkeley's features are not tuned to our domain
- Our data contains a larger percentage of complex coreference types that require world knowledge to resolve. The word embeddings allow us to infuse world knowledge into our model.
- Our model is doing pairwise classification instead of mention ranking. Mention ranking may be optimised for sparser text.

Conclusion

- We present a new, naturally-occurring coreference dataset that is easy to annotate but difficult for computers to solve.
- We develop an end-to-end coreference system using very simple models that matches and outperforms traditional systems on our dataset.
- We incorporate active learning to select which documents to annotate and present an annotation framework.

Embracing harder coreference!

- Systems should be able to distinguish who is likely to marry whom, identify the titles of books from roundabout descriptions, and intuit family relationships.
- These kind of challenges are not present in traditional coreference problems or datasets, but we believe this is the way to approach human ability in this task.
- This paper gives a few building blocks to approach the problem in that fashion.

Future work?

- Expand the dataset to all quiz bowl categories
- Joint models between mention detection, coreference, and related tasks
- Joint models with other domains, primarily vision

Thank you!
Questions?

Paper and data present at

`http://www.cs.umd.edu/~aguha/qbcoreference`