

The IBM Systems for Trilingual Entity Discovery and Linking at TAC 2015

Avirup Sil, Georgiana Dinu and Radu Florian
IBM T.J. Watson Research Center
Yorktown Heights, NY



Outline

- General Architecture for the IBM Entity Discovery & Linking (EDL) System
 - Mention Detection
 - Entity Linking & Clustering
- Adjusting the system to the TAC Trilingual EDL Task
- Experiments and Results

Mention Detection

- Standard IOB sequence classifier, trained on the task
- 2 main classifiers: CRF and Neural Network-based
- The Spanish system was jointly trained on English and Spanish
- Chinese system is a character-based system

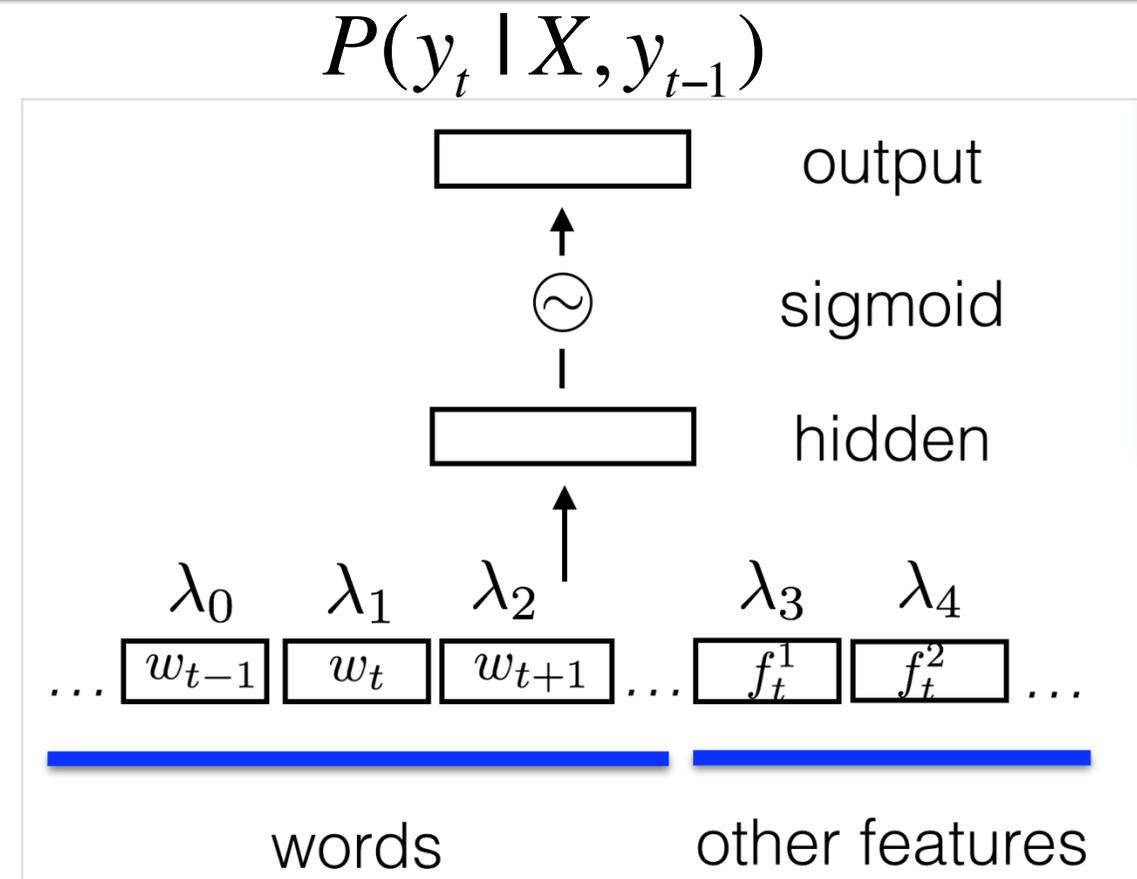
Mention Detection (NN)

- Computed the probability:

$$P(y_t | X, y_{t-1})$$

using a neural network

- Uses Viterbi to find the best tag sequence
- Contrary to popular belief, it does better when trained with linguistic features!



System Combination for Mention Detection

- Both systems are high precision
- We combine them as follows
 - Start with the “best” system
 - For each consequent system
 - Add any mentions that do not overlap with the current output

	CRF	NN	Combination
English	0.715	0.718	0.727
Spanish	0.703	0.698	0.752

Overlapping Mentions

- Some “interesting” examples

m.07_m9_ NILo1468
 HitlerWasASexyMofo

m.045m1_ NILo1371
 Jesus_was_a_Panda

m.o_6t_z8 m.o2jgz
 EU

- Some others

m.034ls (George H.W. Bush) m.019x9z (Jeb Bush)
 grandfather

TEDL15_EVAL_22905

m.019x9z m.019x9z
 Jeb Bush

TEDL15_EVAL_22905

m.019x9z m.019x9z
 Jeb Bush

TEDL15_EVAL_27473

NILo0009 NILo0009
 Dylann Roof

TEDL15_EVAL_03416 ... (21 of them)

m.0345h m.0345h NILo0929
 Germany

TEDL15_EVAL_04270

Entity Linking (EL)

- Reference Knowledge Base
- Preprocessing for IBM EL System
- Our Re-ranking model (and using the same model for other languages)
- Experiments

Reference Knowledge Base (KB)

- Information extraction from Wikipedia
 - April 2014 dump of the English corpus
 - ~4.3M Pages (unique KB ids/titles)
 - Text
 - Redirects
 - Inlinks
 - Outlinks
 - Categories
 - $\text{Pr}(\text{title}|\text{mention})$: prior probability

Reference Knowledge Base (KB)

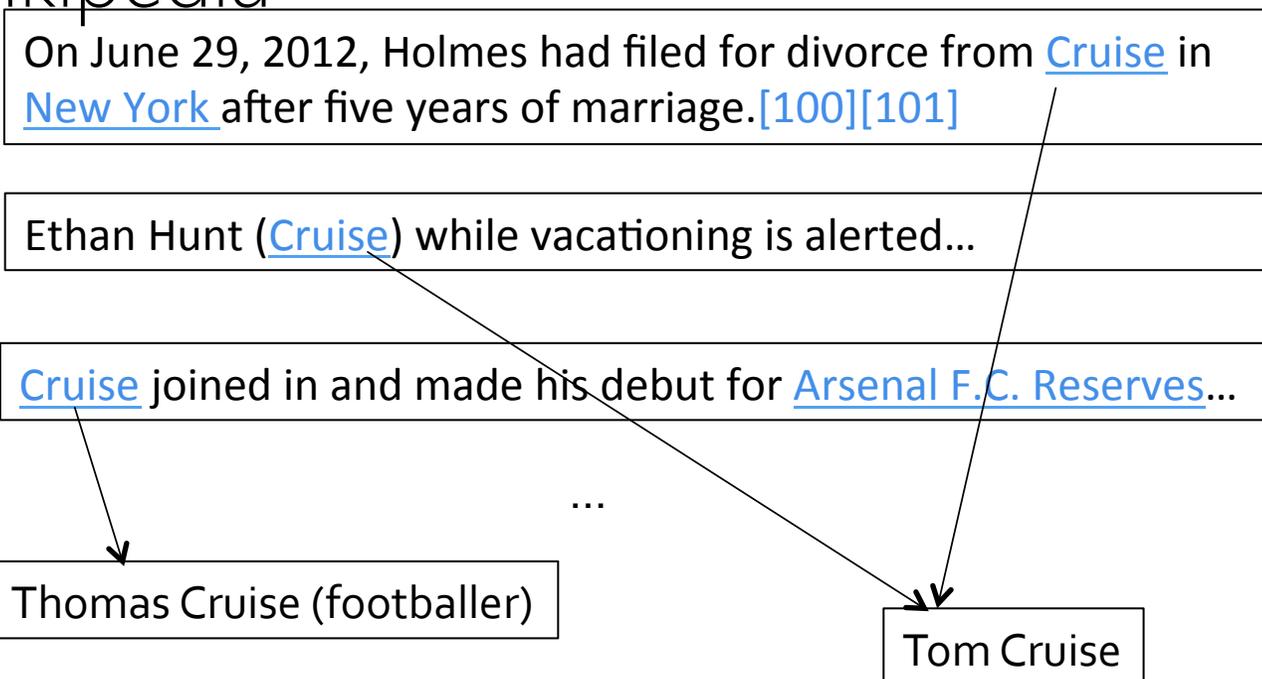
- Information extraction from Wikipedia
 - April 2014 dump
 - ~4.3M KB Ids
 - Text
 - Redirects
 - Inlinks
 - Outlinks
 - Categories
 - $\text{Pr}(\text{title}|\text{mention})$: prior probability

Categories: [Tom Cruise](#) | [1962 births](#) | [20th-century American male actors](#) | [21st-century American male actors](#) | [American expatriates in Canada](#) | [American male film actors](#) | [American film producers](#) | [American people of English descent](#) | [American people of Irish descent](#) | [American Scientologists](#) | [Best Actor Empire Award winners](#) | [Best Drama Actor Golden Globe \(film\) winners](#) | [Best Musical or Comedy Actor Golden Globe \(film\) winners](#) | [Best Supporting Actor Golden Globe \(film\) winners](#) | [Former Roman Catholics](#) | [Converts to Scientology from Roman Catholicism](#)

Reference Knowledge Base (KB)

■ Information extraction from Wikipedia

- April 2014 dump
- ~4.3M KB Ids
- Text
- Redirects
- Inlinks
- Outlinks
- Categories
- $\Pr(\text{title}|\text{mention})$: prior probability



Outline

- ✓ Reference Knowledge Base
 - Preprocessing for IBM EL System
 - Our Re-ranking model
 - Experiments

Preprocessing for the IBM EL System



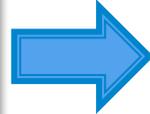
Ashes 2013: England win Ashes as Stuart Broad stars with ball

By Sam Sheringham
BBC Sport at Chester-le-Street

Comments (967)

Fourth Investec Test, Emirates Durham ICG (day four):
England (238 & 330) beat Australia (270 & 224) by 74 runs
[Match scorecard](#)

An inspired spell of fast bowling from Stuart Broad catapulted England to a 74-run win over Australia in the fourth Test and sealed victory in the Ashes series.



*"..Broad catapulted England to a 74-run win over Australia...
...
Tim Bresnan had opener David Warner.."*

Extracted Text

IBM SIRE



1. Mention Detection
2. In-Doc Coref

*"[Broad] catapulted [England] to a 74-run win over [Australia]...
...
[Tim Bresnan] had opener [David Warner].."*

Text with mentions



Partition the mentions into sets of mentions

Any Web Document

Preprocessing for the IBM EL System



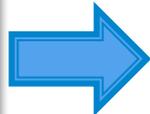
Ashes 2013: England win Ashes as Stuart Broad stars with ball

By Sam Sheringham
BBC Sport at Chester-le-Street

Comments (967)

Fourth Investec Test, Emirates Durham ICG (day four):
England (238 & 330) beat Australia (270 & 224) by 74 runs
[Match scorecard](#)

An inspired spell of fast bowling from Stuart Broad catapulted England to a 74-run win over Australia in the fourth Test and sealed victory in the Ashes series.



“..Broad catapulted England to a 74-run win over Australia...
...
Tim Bresnan had opener David Warner..”

Extracted Text



1. Mention Detection
2. In-Doc Coref

Broad; England; Australia

Tim Bresnan; David Warner

Text with mentions

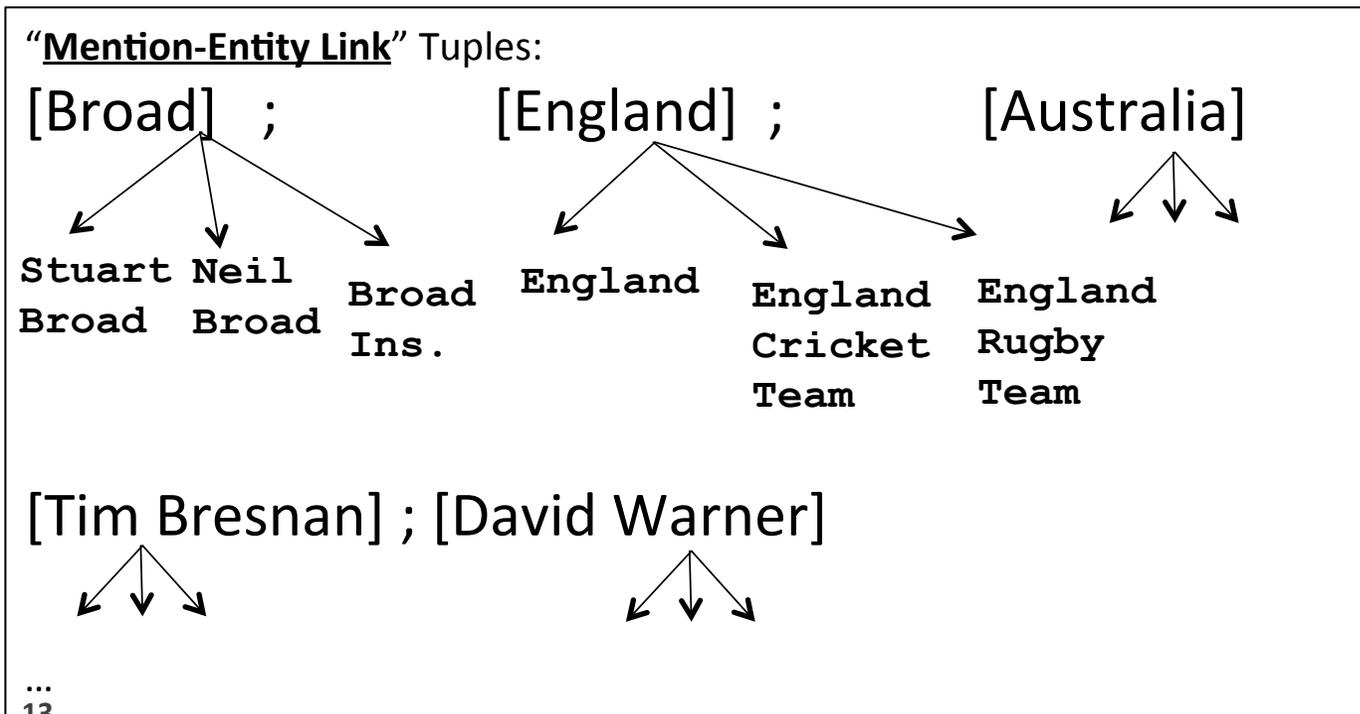


Partition the mentions into sets of mentions

- Connected Component 1
 - Mentions:
 - Broad; England; Australia
- Connected Component 2
 - Mentions:
 - Tim Bresnan; David Warner
- ...

Connected Components

Extract top-K Candidate Entity Links



IBM EL Model: Re-ranking the "Mention-Entity Link" Tuples

"Broad; England; Australia"
Connected Component

Mention-Entity Link Tuples:

1. { [Broad], Stuart_Broad, [England], England_Cricket_Team, [Australia], Australia_Cricket_Team }
2. { [Broad], Neil Broad, [England], England, [Australia], Australia }
3. ...
4. { [Broad], Neil Broad, [England], England, [Australia], Australia_Cricket_Team }



"Tim Bresnan; David Warner"
Connected Component

Mention-Entity Link Tuples:

1. { [Tim Bresnan], Tim_Bresnan, [David Warner], David_Warner_(actor) }
2. { [Tim Bresnan], Tim_Bresnan, [David Warner], David_Warner_(cricketer) }
3. ...



Re-ranking model:

$$P(b|d, cc, \mathbf{w}) = \frac{\exp(\mathbf{w} \cdot \mathbf{f}(b, d, cc))}{\sum_{b' \in B(cc)} \exp(\mathbf{w} \cdot \mathbf{f}(b', d, cc))}$$

Classifier:

- Maximum Entropy

IBM EL Feature Functions

- Local Features
 - Cosine Similarity
 - Domain Independent features
 - Count All (Category, Redirect Links, InLinks, Outlinks,..)
 - Count Unique (Category, Redirect Links, InLinks, Outlinks,..)
- Global Features
 - Features from Entity Links
 - Categorical Relation Count
 - Entity-Type-PMI
 - NIL Detector Features
 - Token-level features
 - Link Overlap

Local Features

- Knowledge-base Independent features from Sil et.al. 2012 are ported to Wikipedia
- Example of such a feature: Count All (OutLinks)

Text: "...[Broad] catapulted [England] to a 74-run win over [Australia] in the [Ashes] Test series thanks to [Tim Bresnan]..."



ID	Name	Outlinks
Stuart_Broad	Stuart Broad	England; Australia; Ashes; Tim Bresnan, ...



ID	Name	Outlinks
Neil_Broad	Neil Broad	Australia, Grand Slam, ...

Count All (Outlinks) {[Broad], Stuart_Broad}
 = Count<Outlink_1> + Count<Outlink_2> + ..
 = Count<England> + Count<Australia> + ...
 = 1 + 1 + 1 + 1 + .. = 4 

Count All (Outlinks) {[Broad], Neil_Broad}
 = Count<Outlink_1> + Count<Outlink_2> + ..
 = Count<Australia> + Count<Grad Slam> + ...
 = 1 + 0 + .. = 1 

Local Features: Cosine Similarity (Word Embeddings)

" ..seam bowler [Broad] catapulted [England] to a 74-run win "

Stuart Broad

From Wikipedia, the free encyclopedia

Stuart Christopher John Broad (born 24 June 1986) is a cricketer who plays [Test](#) and [One Day International](#) cricket and is currently the captain of their [Twenty20](#) team. A left-handed batsman and right-arm **seam bowler**, Broad's professional career started at [Leicestershire](#), the team attached to his

England

Stuart Broad





Neil Broad

From Wikipedia, the free encyclopedia

Neil Broad (born 20 November 1966 in [Cape Town](#), South Africa) is a [tennis](#) player who represented Great Britain for most of his playing career. He is a former [UK number 1](#) who won 7 [ATP tour](#) doubles titles in his career, and won the Silver Medal in doubles at the [1996 Summer Olympics](#).

Neil Broad	
Country	 South Africa  United Kingdom
Residence	Cape Town, South Africa
Born	20 November 1966 (age 47) Cape Town , South Africa
Height	1.90 m (6 ft 3 in)
Turned pro	1986



1. Obtain the embeddings [Mikolov13] of words from input and Wiki target
2. Sum up all the embeddings from input and Wiki target
3. Compute:
 - $\text{Cosine_Similarity}(\text{InputDoc}, \text{Wiki}(\text{Stuart_Broad})) > \text{Cosine_Similarity}(\text{InputDoc}, \text{Wiki}(\text{Neil_Broad}))$

Local Features: Cosine Similarity (Raw)

" ..seam bowler [Broad] catapulted [England] to a 74-run win "

Stuart Broad

From Wikipedia, the free encyclopedia

Stuart Christopher John Broad (born 24 June 1986) is a cricketer who plays [Test](#) and [One Day International](#) cricket for **England** and is currently the captain of their [Twenty20](#) team. A left-handed batsman and right-arm **seam bowler**, Broad's professional career started at [Leicestershire](#), the team attached to his

Stuart Broad




Neil Broad

From Wikipedia, the free encyclopedia

Neil Broad (born 20 November 1966 in [Cape Town](#), South Africa) is a [tennis](#) player who represented Great Britain for most of his playing career. He is a former [UK number 1](#) who won 7 [ATP tour](#) doubles titles in his career, and won the Silver Medal in doubles at the [1996 Summer Olympics](#).

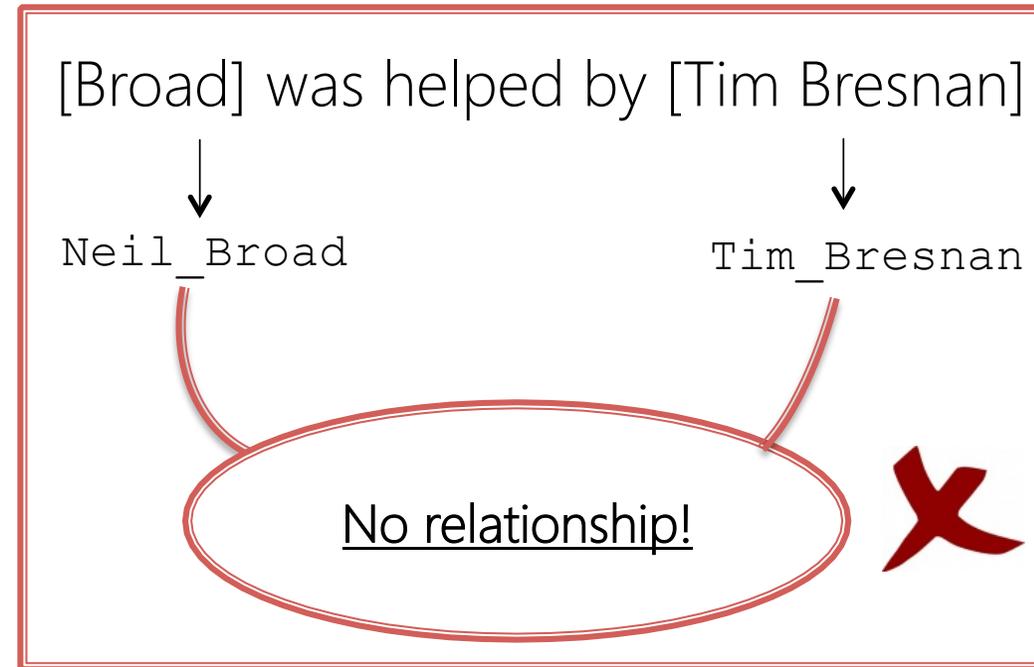
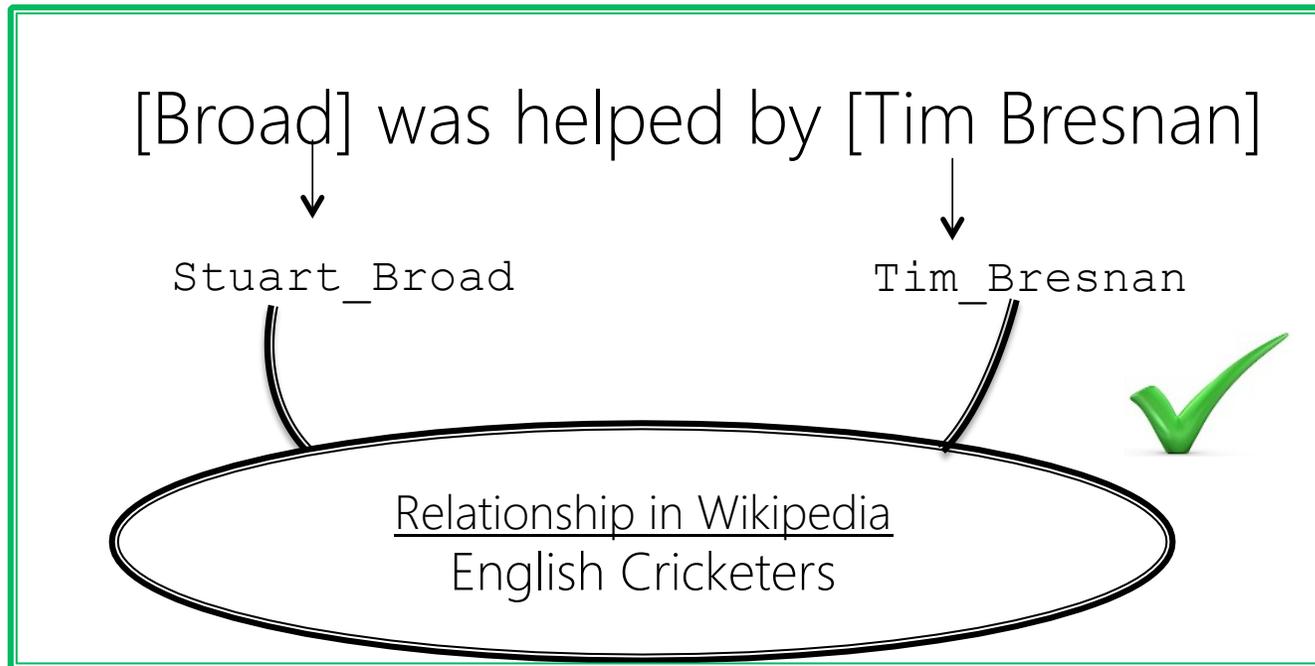
Neil Broad	
Country	 South Africa  United Kingdom
Residence	Cape Town, South Africa
Born	20 November 1966 (age 47) Cape Town , South Africa
Height	1.90 m (6 ft 3 in)
Turned pro	1986



$\text{Cosine_Similarity}(\text{InputDoc}, \text{Wiki}(\text{Stuart_Broad})) > \text{Cosine_Similarity}(\text{InputDoc}, \text{Wiki}(\text{Neil_Broad}))$

Global Features: Categorical Relation Count

- Use Category Relations between entities in Wikipedia
- Example:



Indicates: A Poor Match!

Global Features: NIL Detector Features (NDF)

*“Local journalist **[Michael Jordan]** reported, “[Martin O'Malley], meanwhile, offered his prayers and solidarity with the president”.*

=> CC = {Martin O'Malley, Michael Jordan}

- NDF1: Count #OutLinks overlap
 - NDF1 (Martin_O'Malley, Michael_Jordan_(basketball_player)) = 0
- NDF2: Count #RoleName
 - NDF2 (journalist, Michael_Jordan_(basketball_player)) = 0



Extending to Spanish & Chinese

- The IBM EL system is Language-Independent
 - The same EL model has been ported for the Spanish & Chinese EL Task without the need for re-training
 - Only requirement:
 - Preprocess the Spanish & Chinese WP corpus to build our own internal Spanish & Chinese KB
 - Prior probabilities, Inlinks, Outlinks, Categories, etc.

Adapting the System for TEDL (Demo)

- IBM Statistical Information and Relation Extraction (SIRE) system:

INPUT

Singer Madonna 'can't stop crying' over Jackson

Los Angeles, June 25, 2009 (AFP)

Pop diva Madonna revealed she was left in tears over the death of Michael Jackson on Thursday, saying the music world had lost ..

IBM OUTPUT

PERSON DATE EVENT_COMMUNICATION GPE ORGANIZATION

P Singer Madonna 'can't stop crying' over Jackson

P Los Angeles, June 25, 2009 (AFP)

P Pop diva Madonna revealed she was left in tears over the death of Michael Jackson on Thursday, saying the music world had lost ..

Madonna (entertainer)

From Wikipedia, the free encyclopedia

Madonna Louise Ciccone^[2] (/ˈmɑːdoʊni/; born August 16, 1958) is an American singer, songwriter, actress, and businesswoman. She achieved popularity by pushing the boundaries of lyrical content in mainstream popular music and imagery in her [music videos](#), which became a fixture on [MTV](#).

Madonna



Michael Jackson

From Wikipedia, the free encyclopedia
(Redirected from [Michael Jackson](#))

For other people named Michael Jackson, see [Michael Jackson \(disambiguation\)](#).

Michael Joseph Jackson^{[2][3]} (August 29, 1958 – June 25, 2009) was an American singer, songwriter, dancer, and actor. Called the [King of Pop](#),^{[4][5]} his contributions to music and dance, along with his publicized personal life, made him a global figure in [popular culture](#) for

Michael Jackson



NIL Clustering

- Mentions are linked to the 2014 Wikipedia

Mentions	Wikipedia 2014	TAC KB
Tsarnaev	Dzhokhar_Tsarnaev	NILxxx0
	Tamerlan_Tsarnaev	NILxxx1
Steenkamp	Reeva_Steenkamp	m.0qtngg8
	June_Steenkamp_(NIL)	NILxxx2

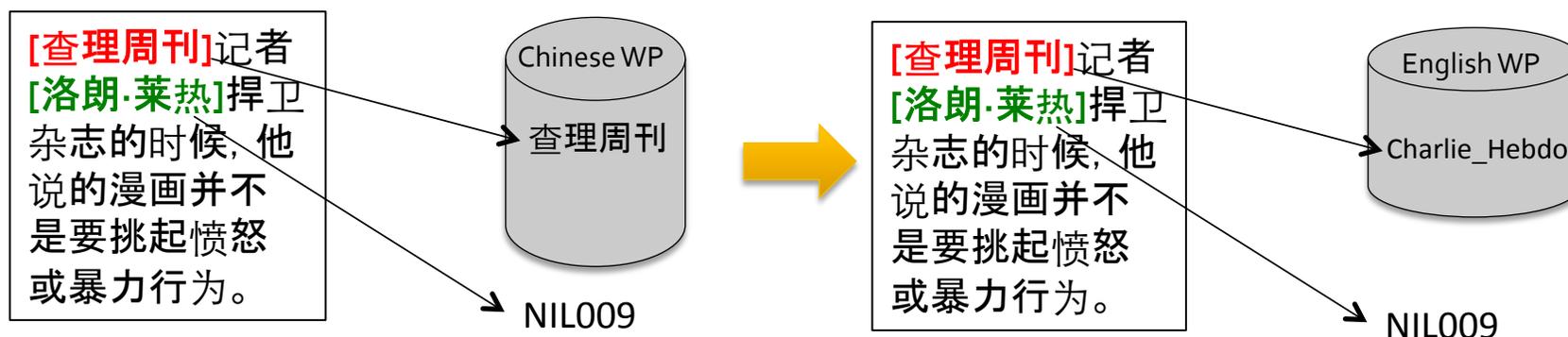
- We also use our in-Doc Coreference component
 - Steenkamp-> June_Steenkamp-> NILxxx2

NIL Clustering (contd.)

- Mapping back to Freebase/ TAC KB :

- Follow [Sil & Florian'14]:

- Map back all non-English titles to the English WP titles (thanks! To WP inter-language links) ☺

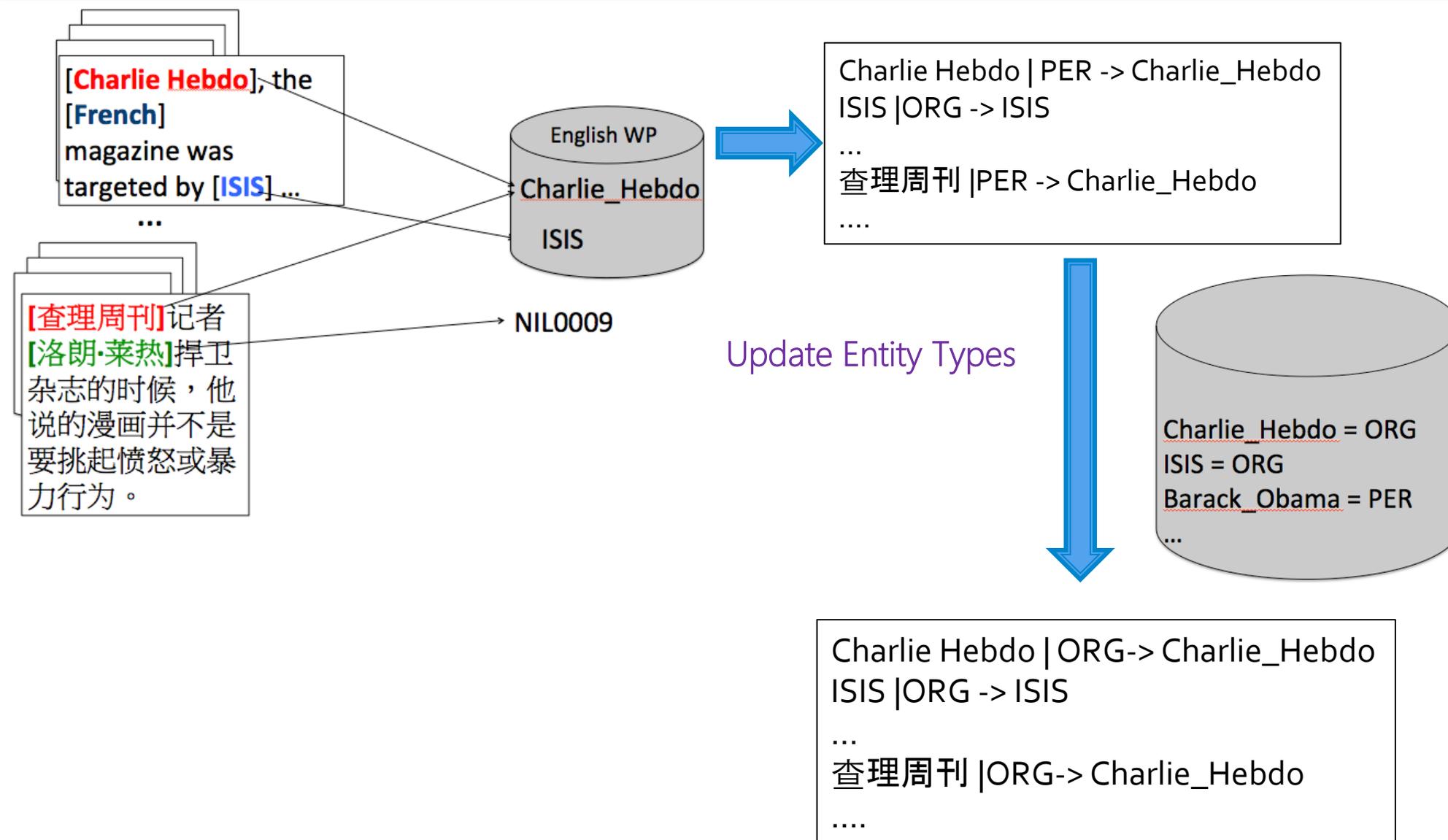


- Map the English WP titles to TAC KB using Freebase to WP redirects



- We use the set of all Wikipedia redirects for clustering entities for NIL or obtaining their KB ids.

Entity Type Prediction



Outline

- ✓ Reference Knowledge Base
- ✓ Preprocessing for IBM EL System
- ✓ Our Re-ranking model
 - Experiments

Experiments (Datasets)

- MD Training
 - Dataset: TAC 2015 (444 docs)
- EL Training
 - Dataset:
 - WikiTrain (Ratinov et.al'11_UIUC): ~10k docs
 - CoNLL 2003 (Hoffart et.al'11_MPI train) ~ 900 docs

EL Results (2013 & 2014)

■ English

System	B3+F1
IBM	0.824
Rank 1	0.821



****Un-official****
2014

■ Spanish

System	B3+F1
IBM	0.80
Rank 2	0.66



****Official****
2014

■ Chinese

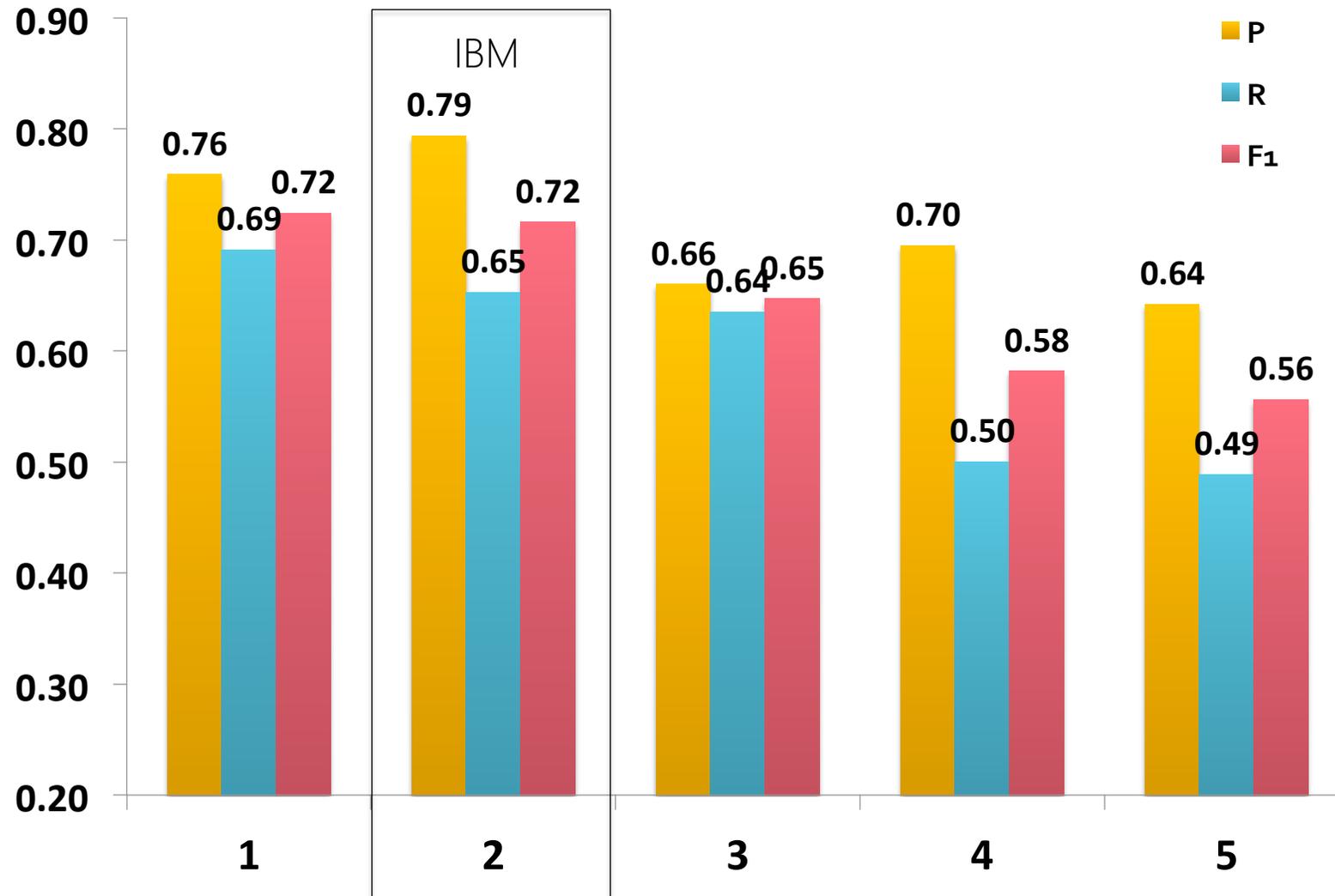
Systems	B3+F1
System 1	0.6
IBM	0.602
System 2	0.63
System 3	0.631



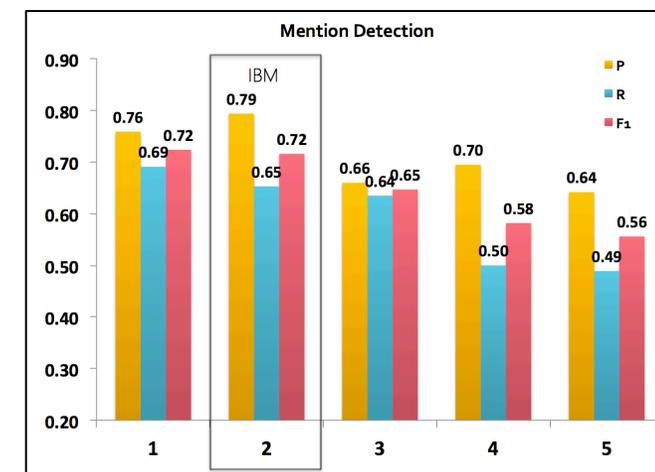
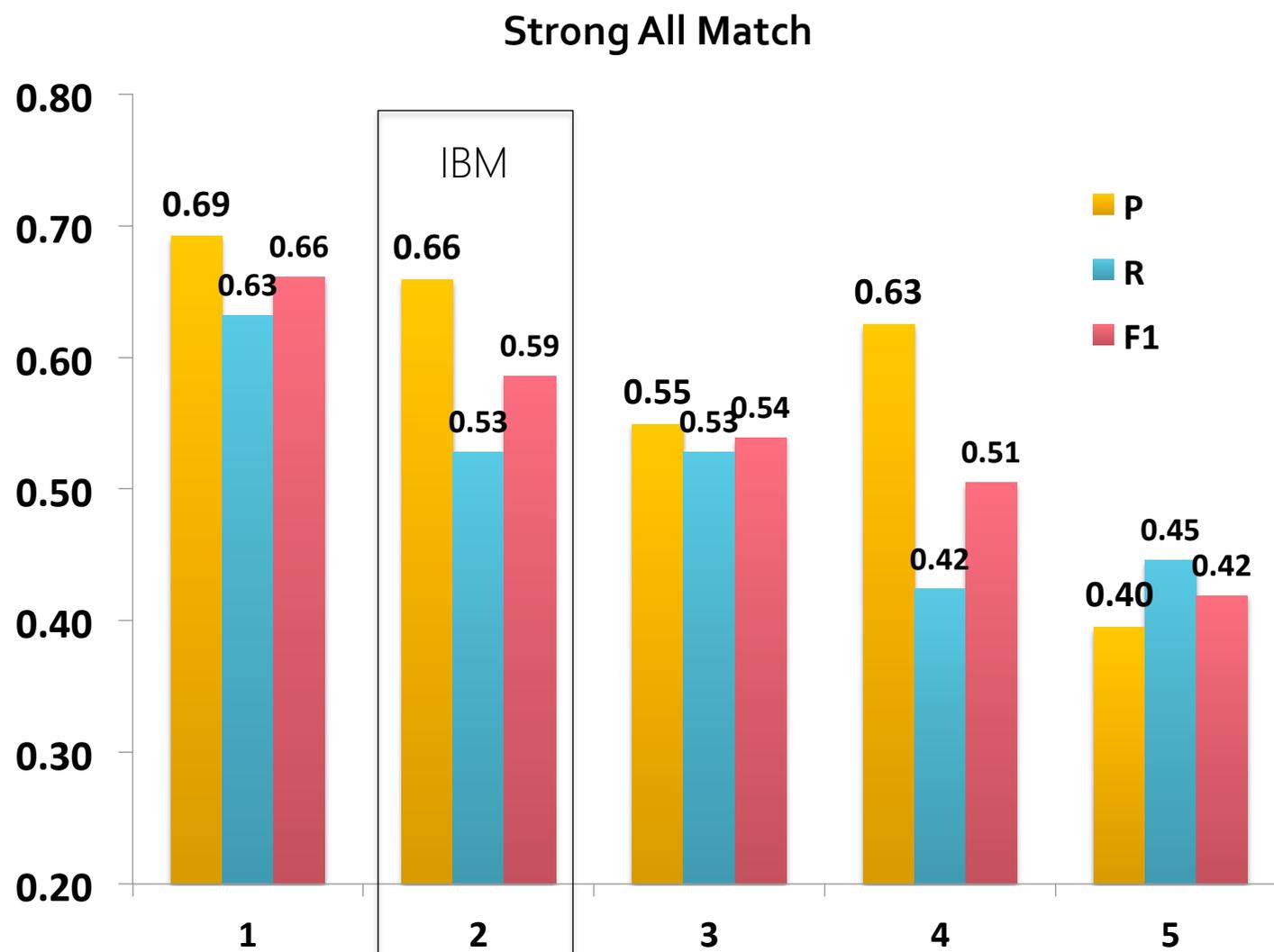
****Un-official****
2013

TEDL Results (2015): MD

Strong Type Mention Match

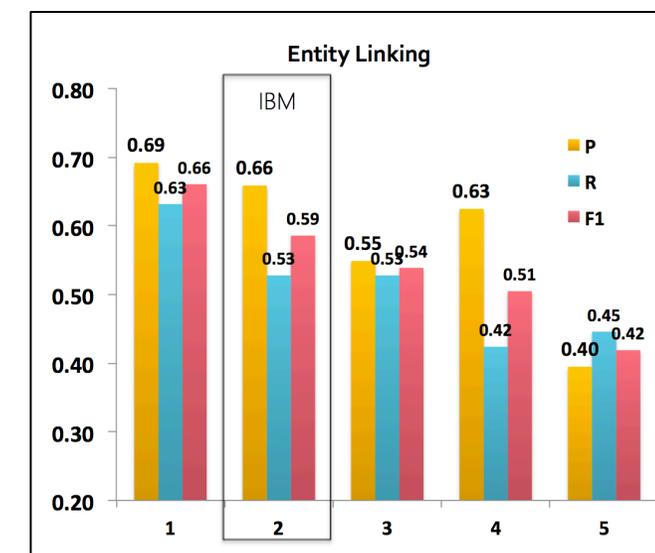
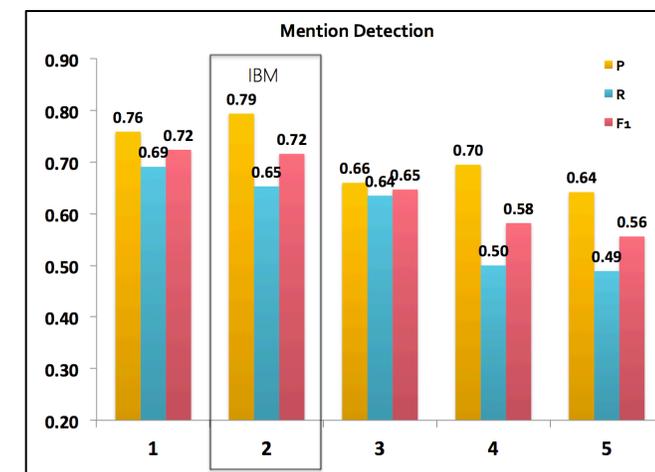
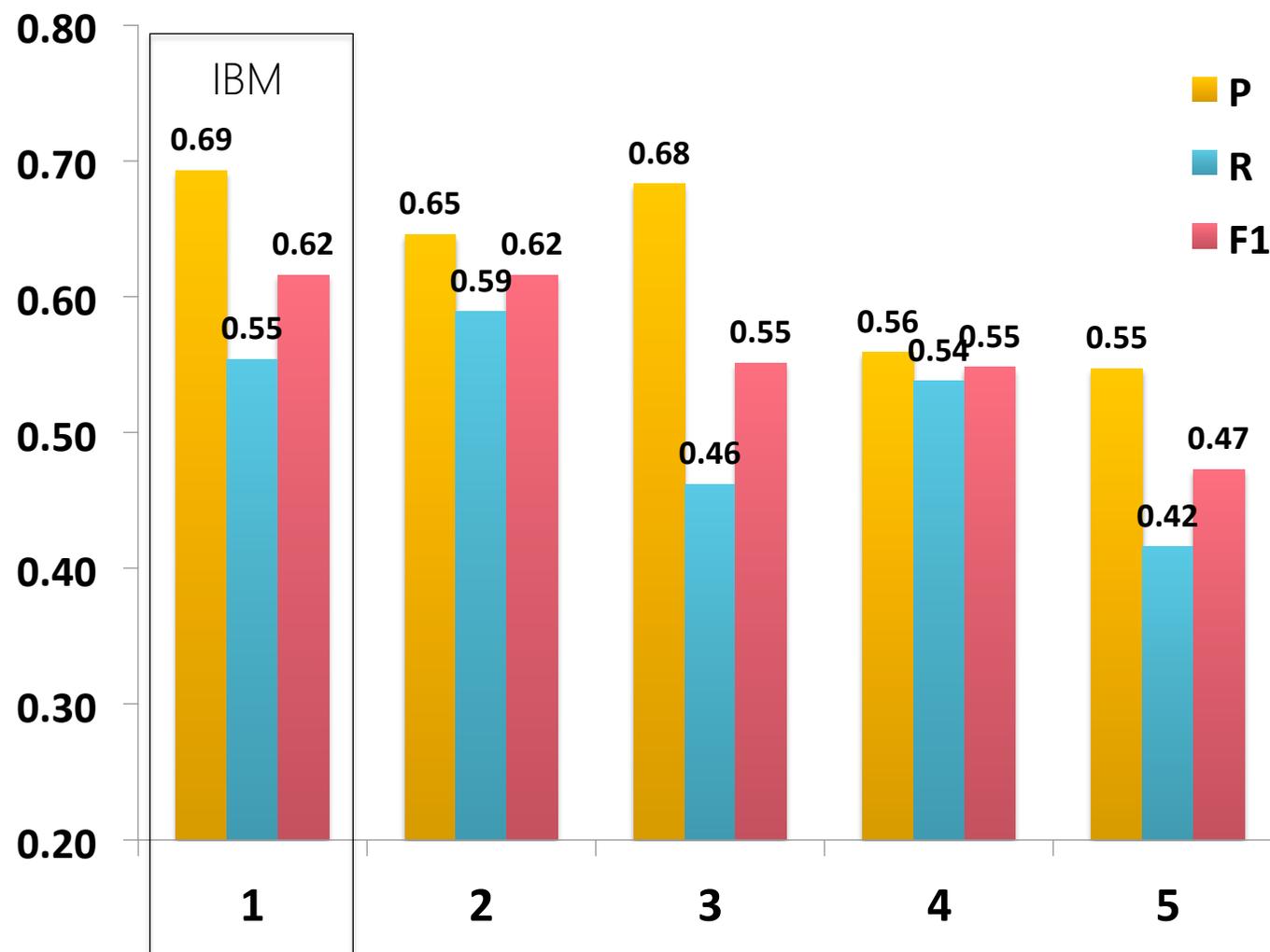


TEDL Results (2015): MD + EL (no clustering)



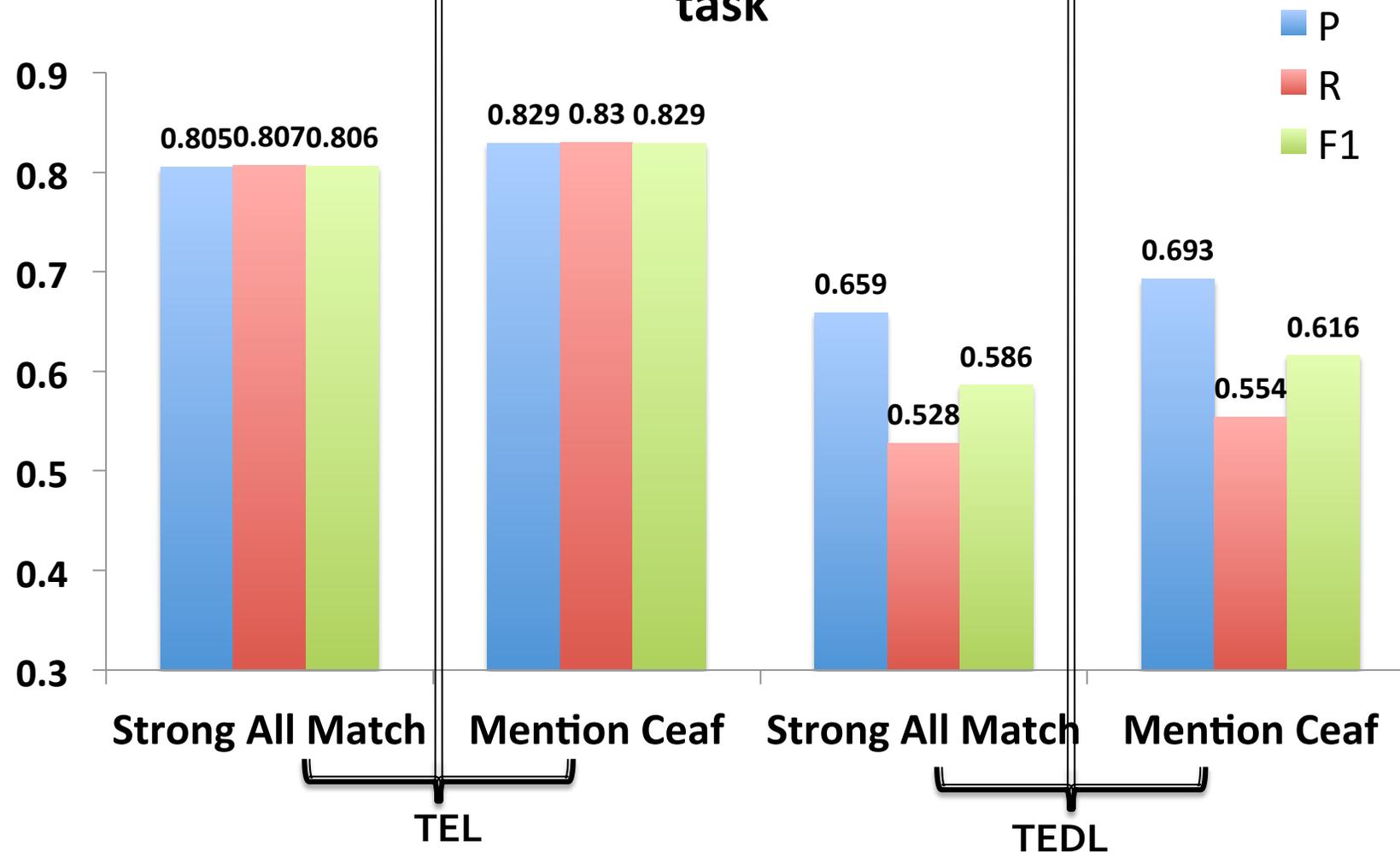
TEDL Results (2015) : MD + EL + Clustering

End to End (Mention CEF)



TEL Results (2015)

Comparison of IBM systems in the TEDL vs. TEL task



1. Difference of :

- A. 0.22 points for Linking
- B. 0.213 points for Linking + Clustering

2. Scope of improvements:

- A. In Doc Coref
- B. Chinese MD

Conclusion

- We presented the IBM Language-Independent EL system
 - The English EL system is used for both Spanish and Chinese
 - Performs joint entity disambiguation using local and global features
- The Mention Detection System obtained the top score in English & Spanish
 - A system combination of NNs and CRFs proved to be a robust solution for the task
- The EL system obtained the top score in the end-to-end metric

Poster/Demo of the IBM system at 15.50

Thanks! Questions?

Gratitude

From Wikipedia, the free encyclopedia

For other uses, see [Gratitude \(disambiguation\)](#).

"Thank" redirects here. For the protein symbol, see [THANK](#). For other uses, see [Thank You \(disambiguation\)](#) and [Thanks \(disambiguation\)](#).

"You're Welcome" redirects here. For the Angel episode, see [You're Welcome \(Angel\)](#).

See also: the Wiktionary entries [thank](#), [thanks](#), [thank you](#), and [you're welcome](#).

Gratitude, **thankfulness**, **gratefulness**, or **appreciation** is a feeling or attitude in acknowledgment of a benefit that one has received or will receive. The experience of gratitude has historically been a focus of several world religions,^[1] and has been considered extensively by moral philosophers such as [Lee Clement](#).^[2] The systematic study of gratitude within psychology only began around the year 2000, possibly because psychology



Question

From Wikipedia, the free encyclopedia
(Redirected from [Questions](#))

For other uses, see [Question \(disambiguation\)](#). To ask questions about Wikipedia, see [Wikipedia:Questions](#).

A **question** is a linguistic expression used to make a request for [information](#), or the request made using such an expression. The information requested may be provided in the form of an [answer](#).

Questions have developed a range of uses that go beyond the simple eliciting of information from another party. [Rhetorical questions](#), for example, are used to make a point, and are not expected to be answered. Many languages have special [grammatical](#) forms for questions (for example, in the English [sentence](#) "Are you happy?", the [inversion](#) of the subject *you* and the verb *are* shows it to be a question rather than a statement). However questions can

“ There are these four ways of answering questions. Which four? There are questions that should be answered categorically [straightforwardly yes, no, this, that]. There are

Email: avi@us.ibm.com