The IBM Systems for Trilingual Entity Discovery and Linking at TAC 2016

Avi Sil

Joint work with: Georgiana Dinu and Radu Florian
IBM T.J. Watson Research Center
Yorktown Heights, NY

Gaithersburg, MD

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
  - CRF: a standard model similar to most prior work
  - NN: next slide
- We do a classifier combination since the outputs are different

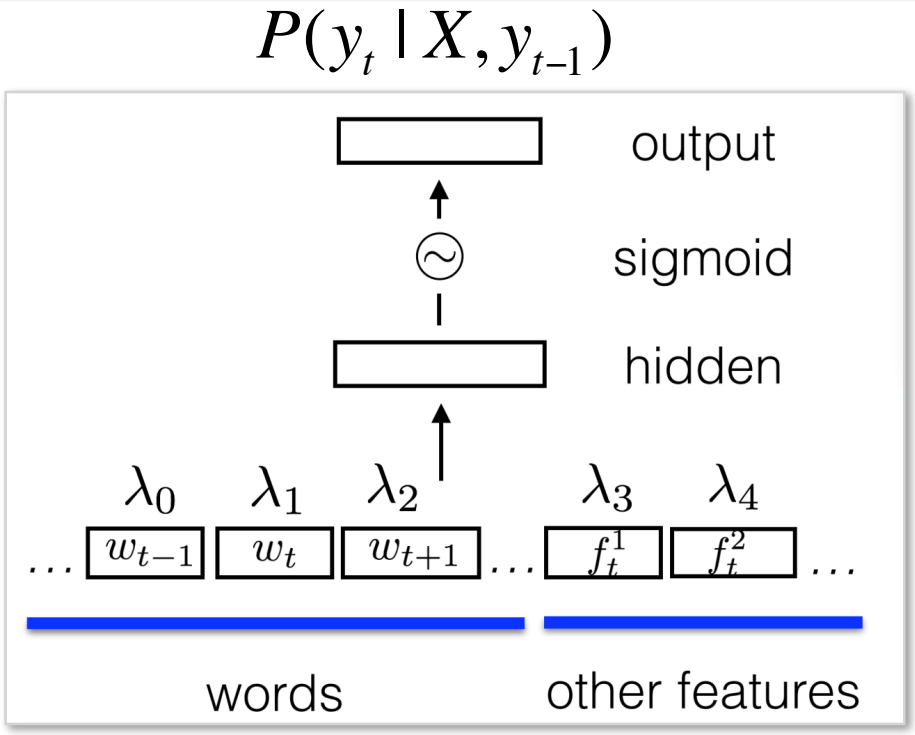
Mention Detection (NN)

- Computed the probability:

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

using a neural network

- It does better when trained with linguistic features!
- We use:
  - Capitalization features
  - Gazetteers
  - Character-level representations (bi-dir LSTMs)



Mention detection (NN): Chinese

- Chinese uses
  - Word (embeddings)
  - character (bi-LSTM)
  - Character and positional character embeddings (concatenation of character+position in the word) [Peng&Dredze,15]

- We perform 10 runs for each model
  - using different random initializations.
  - We combine them through voting.

System Combination for Mention Detection

- We combine the NN and CRF models as follows
  - Start with the "best" system
  - For each consequent system
    - Add any mentions that do not overlap with the current output

	CRF	Best/NN	Vote/NN	Combination
English	0.760	0.747	0.748	0.771
Spanish	0.785	0.766	0.750	0.800
Chinese		0.743	0.744	

TAC 2015 Guidelines: Per, Org, Loc, Fac. Nom: Per (only)

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

Entity Linking (EL)

- LIEL (**L**anguage **I**ndependent **E**ntity **L**inker)
  - Reference Knowledge Base
  - Preprocessing for IBM EL System
  - Training a Re-ranking model (and using the same model for other languages)
  - Experiments

ACL 2016 Paper (top score in previous TAC EDL years):
One for All: Towards Language Independent Named Entity Linking
Avi Sil & Radu Florian

Reference Knowledge Base (KB)

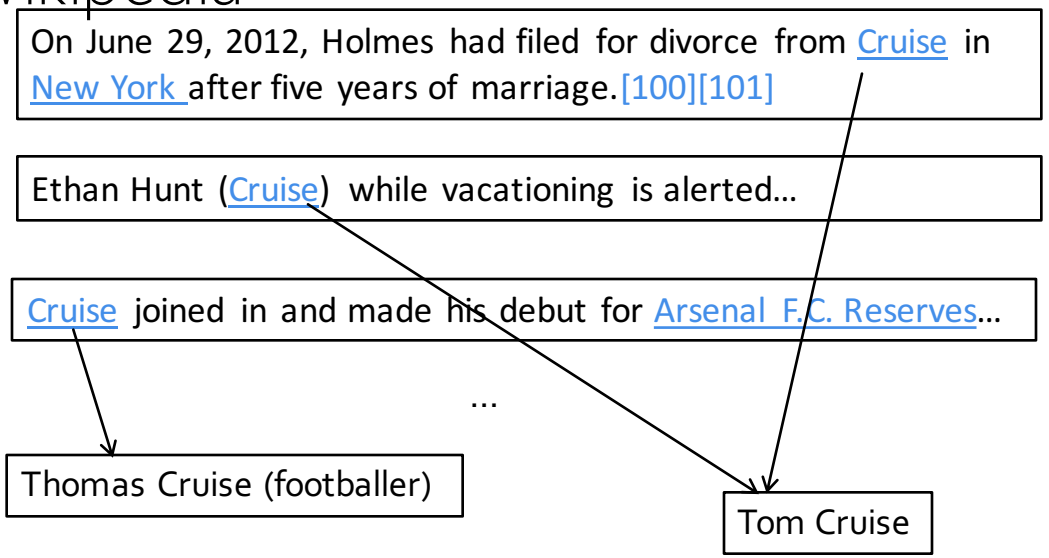
- Information extraction from Wikipedia
  - April 2014 dump of the English corpus
  - ~4.3M Pages (unique KB ids/titles)
  - Text
  - Redirects
  - Inlinks
  - Outlinks
  - Categories
  - Pr(title|mention) : prior probability

Reference Knowledge Base (KB)

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

Reference Knowledge Base (KB)

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



Outline

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

Preprocessing for the IBM EL System

Any Web Document

"..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

Preprocessing for the IBM EL System

Any Web Document

"..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; 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

"**Mention-Entity Link**" Tuples:

[Broad] ; [England] ; [Australia]

Stuart Broad | Neil Broad | Broad Ins. | England | England Cricket Team | England Rugby Team

[Tim Bresnan] ; [David Warner]

...

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}
5. ...

"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:**

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

- **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 [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, ...

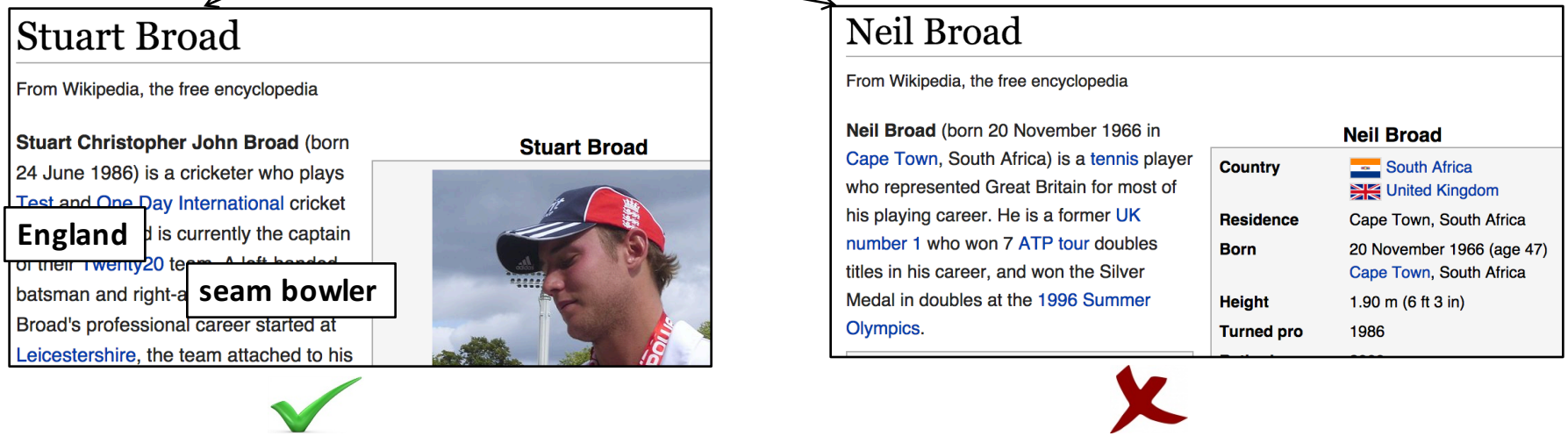
Count All (Outlinks) {([Broad], Stuart_Broad)}
= Count<Outlink_1> + Count<Outlink_2> + ..
= Count<England> + Count<Australia> +...
= 1 + 1 + 1 + 1 +.. = 4 ✓

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

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)

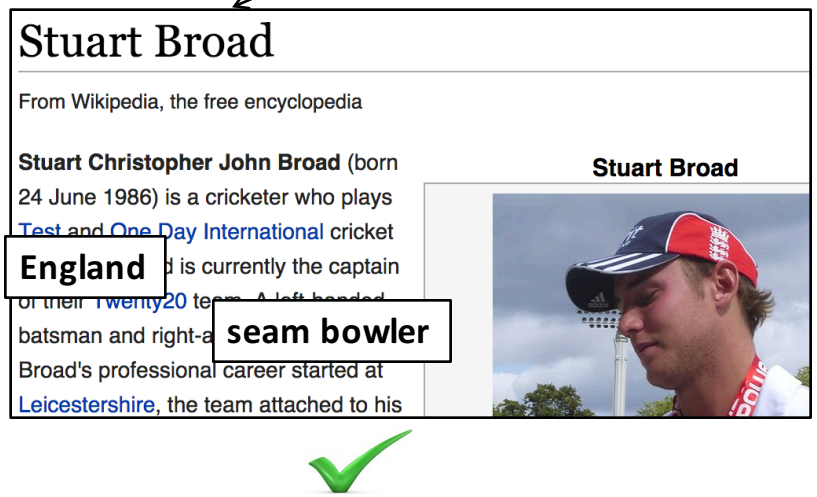
" ..seam bowler **[Broad]** *catapulted* **[England]** *to a 74-run win "*



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:
   - Cosine_Similarity (InputDoc, Wiki (Stuart_Broad)) > Cosine_Similarity (InputDoc, Wiki (Neil_Broad))

Local Features: Cosine Similarity (Raw)

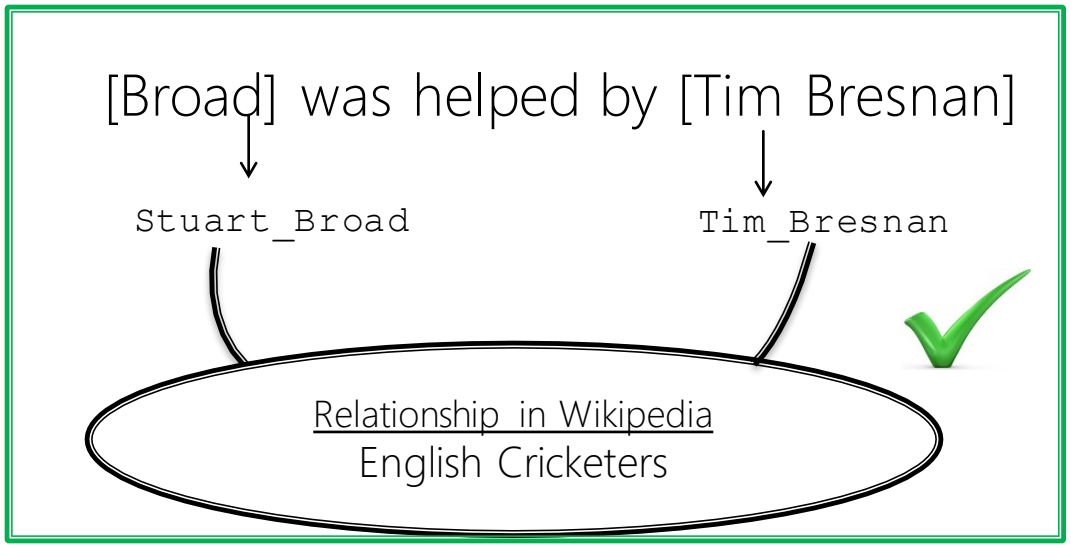
" ..seam bowler **[Broad]** *catapulted* **[England]** *to a 74-run win "*

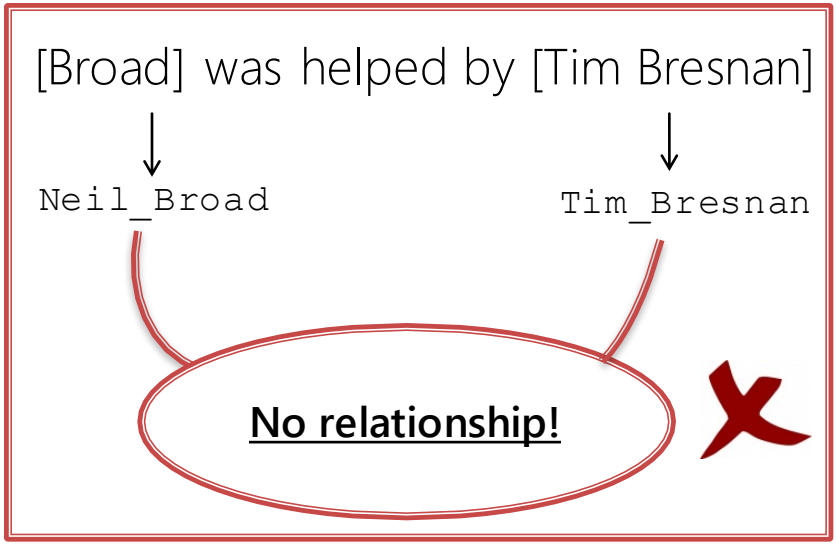


Cosine_Similarity (InputDoc, Wiki (Stuart_Broad)) > Cosine_Similarity (InputDoc, Wiki (Neil_Broad))

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] (/tʃɪˈkoʊni/; 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 (disami

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

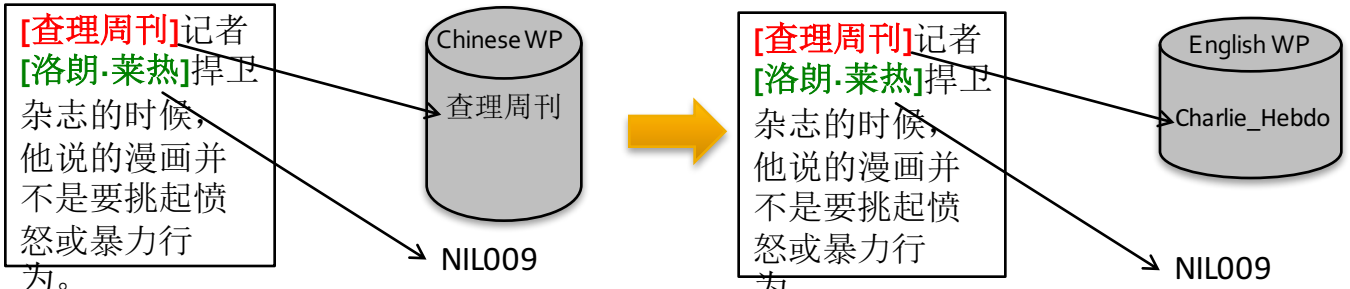
- Mentions are linked to the 2014 Wikipedia

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

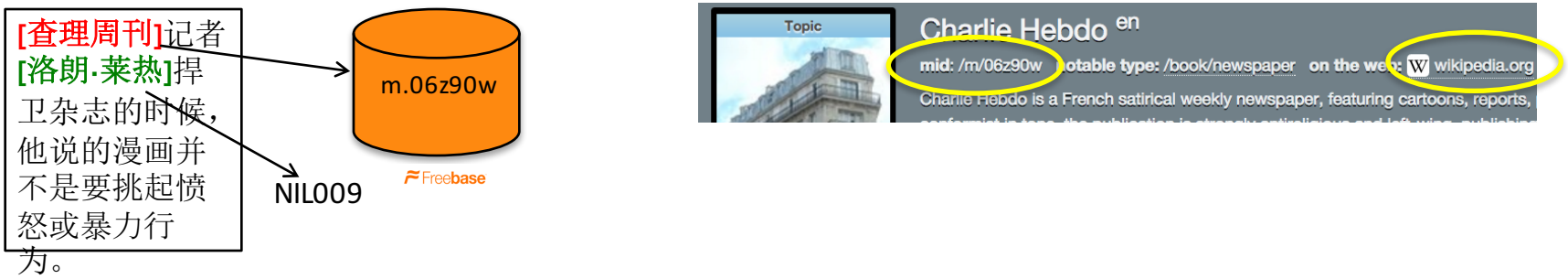
- We also use our in-Doc Coreference component
  - Steenkamp-> June_Steenkamp-> NILxxx2

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.

Outline

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

Experiments (Datasets)

- MD Training
  - Dataset: TAC 2015 train & test
  - Dev: subset of the test data of TAC 2015 (more details in the paper)
  - IBM Klue model used as an input for English
- EL Training
  - Dataset:
    - (Ratinov et.al'11_UIUC): ~10k docs
    - Wikipedia 2014 dataset

MD Results (Dev data)

	NN	Best/NN	Vote/NN	CRF	Combo
English	74.0(±0.4)	74.7	74.8	76.0	77.1
Spanish	75.2(±0.9)	76.6	75.0	78.5	80.0
Chinese	73.4(±0.6)	74.3	74.4	-	-

TEDL Results (2016): MD (best runs)

	Strong Typed Mention Match		
Run ID	Prec	Rec	F1
IBM3	0.829	0.602	0.697
IBM1	0.83	0.599	0.696
IBM2	0.83	0.599	0.696

No NOM mentions other than for PERSON entities

TEDL Results (2016): MD (across languages)

Strong Typed Mention Match			
Language	Prec	Rec	F1
English	0.877	0.665	0.756
Spanish	0.847	0.595	0.699
Chinese	0.761	0.541	0.633

No NOM mentions other than for PERSON entities

TEDL Results (2016): End-to-End (MD+EL)

	Typed Mention CEAF		
Run ID	Prec	Rec	F1
IBM3	0.708	0.511	0.593
IBM1	0.692	0.5	0.58
IBM2	0.687	0.499	0.578

TEDL Results (2016): End-to-End (MD+EL)

Typed Mention CEAF			
Language	Prec	Rec	F1
English	0.734	0.548	0.628
Spanish	0.731	0.514	0.603
Chinese	0.725	0.516	0.603

EL Results (non-TAC datasets)

	Cheng&Roth	LIEL	LIEL+ more Data
ACE	0.853	0.862	0.868
MSNBC	0.812	0.850	0.860

- More training data helps LIEL

Conclusion

- We presented the IBM Language-Independent EL (LIEL) 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
  - A system combination of NNs and CRFs were used
  - A bug was discovered: no NOMs extracted (other than PERSON)

Thanks!

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 T
You (disambiguation) and Thanks (disambiguation).

"You're Welcome" redirects here. For the Angel *episode, see You're Welcome (An*

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.

> 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

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

Email: avi@us.ibm.com