



HITS at TAC 2015

Entity Discovery and Linking

Benjamin Heinzerling^{1,2} and Michael Strube²

¹AIPHES, ²Heidelberg Institute for Theoretical Studies

Joint Global Disambiguation and NIL Clustering

- Our previous years' system (Fahrni et al., 2013, Judea et al., 2014).
- Jointly performs global disambiguation and NIL clustering using a Markov Logic Network (MLN).
- General, not trained on TAC data.
- Performed consistently well in various evaluations.
- Gets some easy decisions wrong.
- Mention detection not joint.

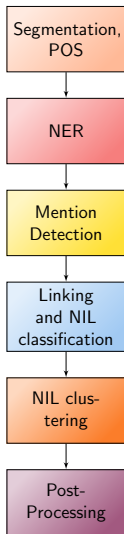
Goals

- Better integrate mention detection.
- Get the easy decisions right.
- See how far the new KB gets us.
- English EDL with mention detection.

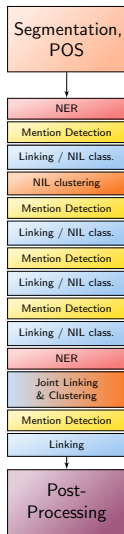
Architecture: Options

- Joint approach: Strong interaction, slow, development difficult.
- Pipeline of tasks: Weak interaction, fast, ordering difficult.
- Pipeline of decisions: Some interaction, fast, ordering feasible.

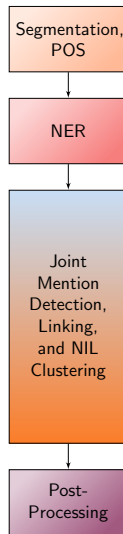
Architecture



Too simple!



Happy middle.



Too complex!

Sieves

- High-precision linking:
 - Unambiguous CrossWiki mentions
 - Entity type checking
 - Sense label mismatch filter
- Person name matching
- Salient semantic paths
- General and TAC-specific post-processing:
 - Dominant sense fallback
 - Country adjectivals mapping
 - Media organization filter

Salient Semantic Paths

- Previous work maximizes relatedness between candidate senses of two or more given mentions (Hoffart et al., 2011; Moro et al. 2014).

Nigeria became Africa's largest economy town of Daura .

Salient Semantic Paths

- Previous work maximizes relatedness between candidate senses of two or more given mentions (Hoffart et al., 2011; Moro et al. 2014).

Nigeria became Africa's largest economy town of Daura .

Nigeria (Africa)

-

Pierre Daura

Salient Semantic Paths

- Previous work maximizes relatedness between candidate senses of two or more given mentions (Hoffart et al., 2011; Moro et al. 2014).

Nigeria became Africa's largest economy town of Daura .

Nigeria (Africa) -

Pierre Daura

Nigeria (Jazz Album) -

Pierre Daura

Salient Semantic Paths

- Previous work maximizes relatedness between candidate senses of two or more given mentions (Hoffart et al., 2011; Moro et al. 2014).

Nigeria became Africa's largest economy town of Daura .

Nigeria (Africa)	-	Pierre Daura
Nigeria (Jazz Album)	-	Pierre Daura
Nigeria (Africa)	/location/location/contains	Daura (Nigeria)

Salient Semantic Paths

- Previous work maximizes relatedness between candidate senses of two or more given mentions (Hoffart et al., 2011; Moro et al. 2014).

Nigeria became Africa's largest economy town of Daura .

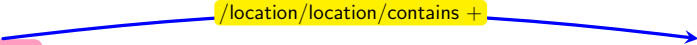
Nigeria (Africa)	-	Pierre Daura
Nigeria (Jazz Album)	-	Pierre Daura
Nigeria (Africa)	/location/location/contains	Daura (Nigeria)
Nigeria (Jazz Album)	-	Daura (Nigeria)

- Slow on Freebase.

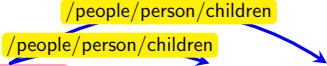
Salient Semantic Paths

- Our approach: Given one linked mention, follow salient semantic paths through the knowledge base graph.
- Faster than looking for paths between candidate senses.

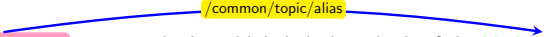
Nigeria became Africa's largest economy town of **Daura**.



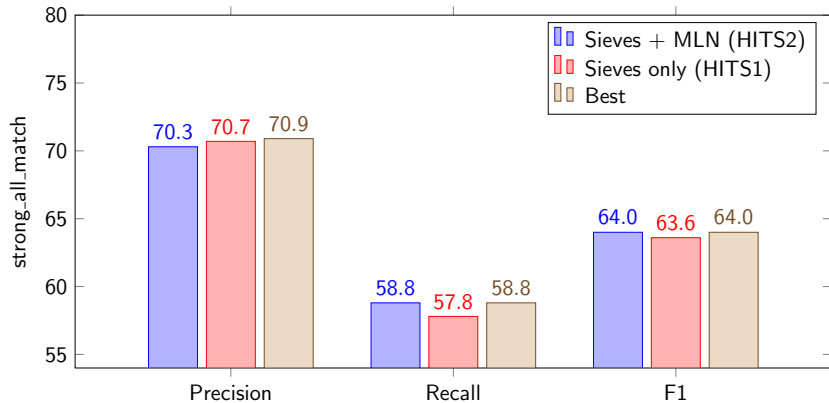
Netanyahu 's sons, **Avner** and **Yair** , were chosen



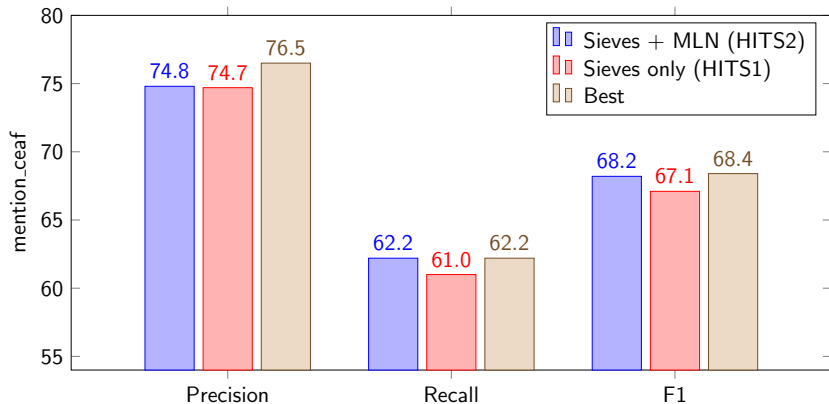
Think of it as **Oscar Pistorius** on steroids. I couldn't help but think of the **blade runner**.



Results: Linking



Results: Clustering



Nominal Coreference?

Hilary Clinton 's latest book ... the *author* ...

Netanyahu will be busy being a leader, unlike Obama the golfer!

- Can common noun surface forms (NOM) be reached in the Kb?
- Tested for paths up to length two.
- Only 30 percent of non-NIL NOMs reachable.

Limitations

- Fails for nominal coreference.
- Hand-picked, genre-specific semantic paths.
- Only works for non-NILs.

Conclusions

- State-of-the-art monolingual linking and clustering.
- Simple symbolic approach: large-scale resources, string match, KB queries.
- Global disambiguation has only a small effect on linking performance.
- Joint disambiguation and NIL clustering improves clustering performance.

Future Work

- Ordering and granularity of decisions.
- Top three systems' linking performance similar: All solving the same easy problems?
- Hard: Metonymy, cohyperonymal lists, nominal coreference.
- Symbolic approaches unlikely to solve these.
- Need to combine with distributional methods.

Thank You