
Bloomberg
NAMED
ENTITY
DISAMBIGUATION

AGENDA

1. Why bother?
2. How to disambiguate
3. Bloomberg NED
4. Results

All images from Wikipedia unless otherwise specified

1/ WHY BOTHER?

Why do we need Named Entity Disambiguation?

From www.bloomberg.com:

"Castro in a history-laden speech of more than 45 minutes, criticized the U.S. for past actions against Cuba, including the designation as a supporter of terrorism, while praising Obama for embracing the new policy"source

http://www.bloomberg.com/politics/articles/2015-04-11/obama-meets-cuba-s-castro-with-much-to-discuss-on-restoring-ties

Why do we need Named Entity Disambiguation?

?

From www.bloomberg.com:

*"**Castro** in a history-laden speech of more than 45 minutes, criticized the U.S. for past actions against Cuba, including the designation as a supporter of terrorism, while praising Obama for embracing the new policy"*

http://www.bloomberg.com/politics/articles/2015-04-11/obama-meets-cuba-s-castro-with-much-to-discuss-on-restoring-ties

Why do we need Named Entity Disambiguation?




?
?


?


From www.bloomberg.com:

*"**Castro** in a history-laden speech of more than 45 minutes, criticized the **U.S**. for past actions against **Cuba**, including the designation as a supporter of terrorism, while praising **Obama** for embracing the new policy"*

http://www.bloomberg.com/politics/articles/2015-04-11/obama-meets-cuba-s-castro-with-much-to-discuss-on-restoring-ties

Why do we need Named Entity Disambiguation?

- Bloomberg owns a vast news archive
- We processes > 1M new articles/day. Clearly, NED needs automation.
- Named Entity disambiguation can enable intelligent products by means of:
  - Semantic search
  - Machine Reading
  - Building knowledge graphs


2/
HOW TO DISAMBIGUATE

How do we disambiguate?

Let's consider the same example again:

"Castro in a history-laden speech of more than 45 minutes, criticized the U.S. for past actions against Cuba, including the designation as a supporter of terrorism, while praising Obama for embracing the new policy"

As human beings, we use both priors (our personal expectations about the entities) and context (text around the named entities)

How do we disambiguate?

Let's consider the same example again:

"Castro in a history-laden speech of more than 45 minutes, criticized the U.S. for past actions against Cuba, including the designation as a supporter of terrorism, while praising Obama for embracing the new policy"

As human beings, we use the context, i.e., text around the named entities.

Word Sense Disambiguation

Named Entity Disambiguation is a specialization of the problem of finding the sense (or meaning) of a string based on the context around the it, namely Word Sense Disambiguation.

A survey of Word Sense Disambiguation by Navigli (2009), offers a very comprehensive introduction to the problem.

Let's look at how machines can interpret context…

How do machines disambiguate?

Let's consider the same example again:

"Castro in a history-laden speech of more than 45 minutes, criticized the U.S. for past actions against Cuba, including the designation as a supporter of terrorism, while praising Obama for embracing the new policy"

Named Entity Recognition is the process of finding mentions of named entities in the text.

How do machines disambiguate?

Let's consider the same example again:

BAG OF WORDS

PART-OF-SPEECH TAGS

TOPIC: POLITICS

"[*Castro*]$_{NNP}$ *in a history-laden speech of more* [illegible]*45 minutes,* [*criticized*]$_{VBD}$ *the* [*U.S.*]$_{NNPS}$ *for past actions* [illegible] *the designation as a supporter of* [illegible] *while praising Obama for embracing the new policy*"

SUBJECT

VERB

ART

OBJECT

NOUN PHRASE

VERB PHRASE

And let's look at some features that machines can extract from the context.

SENTENCE

n-GRAMS

This looks suspiciously like Wikipedia

"Castro in a history-laden speech of more than 45 minutes, criticized the U.S. for past actions against Cuba, including the designation as a supporter of terrorism, while praising Obama for embracing the new policy"

Indeed

Cuban Thaw

From Wikipedia, the free encyclopedia

The **Cuban Thaw**[1] is an historic warming of U.S.-Cuba relations that began in December 2014.

On December 17, 2014, U.S. President Barack Obama and Cuban President Raúl Castro announced the beginning of a process of normalizing relations between Cuba and the United States. The normalization agreement was secretly negotiated in preceding months with the assistance of Pope Francis. Meetings were held in both Canada and Vatican City.[2] The agreement would see the lifting of some U.S. travel restrictions, fewer restrictions on remittances, U.S. banks' access to the Cuban financial system,[3] and the reopening of embassies in Havana and Washington, which closed in 1961 after the breakup of diplomatic relations following the establishment of Cuba's close alliance with the USSR.[4][5]

Some relevant related work

2003 – 1st Web-scale NED system (SemTag) proposed by Dill et al.

2006 – (1st time) Wikipedia's textual context is used for NED by Bunescu and Pasca.

2009 – Joint disambiguation: disambiguating multiple mentions to entities (from Wikipedia) at the same time are used for an NED system (AIDA) by Kulkarni et al.

- Best accuracy, but NP-hard on # of mentions and candidate entities. Heuristics still slow.

2010-2011 – Fast Web-scale NED systems based on Wikipedia are proposed like TagMe (Ferragina and Scaiella) and DBpedia Spotlight (Mendes et al.).

- Simple features are used to assign entities to mentions (individually).

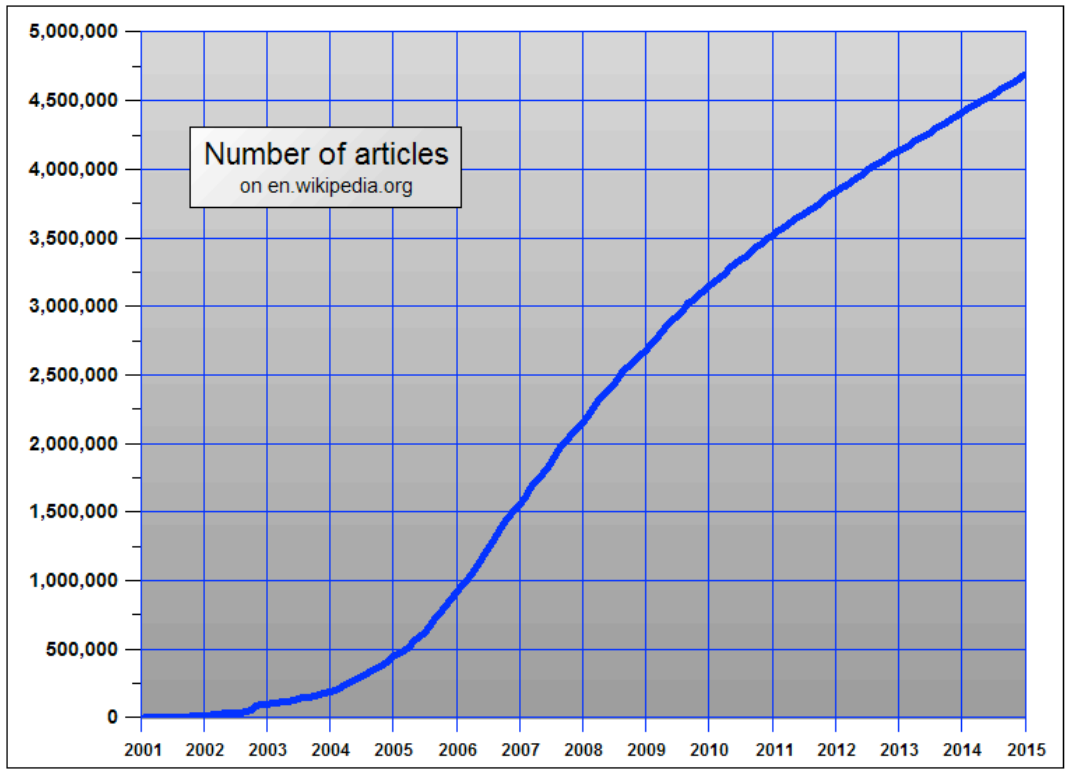
2011 – AIDA: Approximate solution for the joint disambiguation into Wikipedia models are proposed by Hoffart et al.

2014 – A Web-scale NED system (AIDA-light) based on Wikipedia with Joint disambiguation features is proposed by Nguyen et al.

- At least as fast the fastest NED systems (Dbpedia Spotlight and TagMe, at the time)
- At least as accurate as the most accurate systems (AIDA, at the time)

Wikipedia, as annotated set of Named Entities

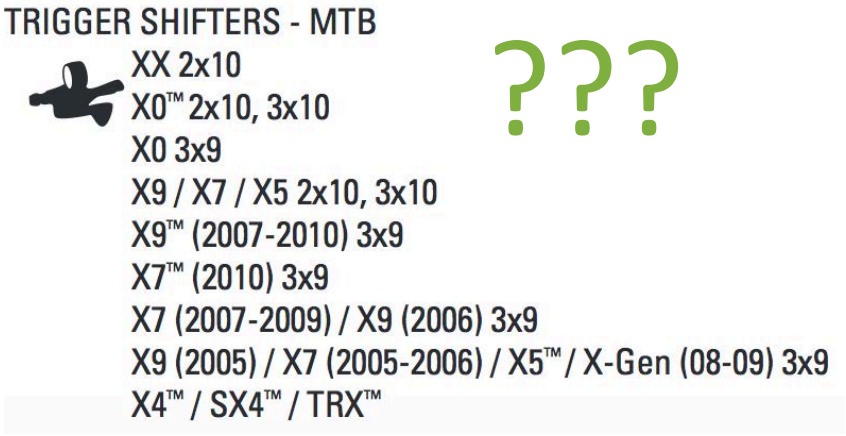
Bunescu and Pasca 2006: use Wikipedia for both training and evaluation, using both *links* and *categories*.



Source: https://en.wikipedia.org/wiki/Wikipedia:Statistics

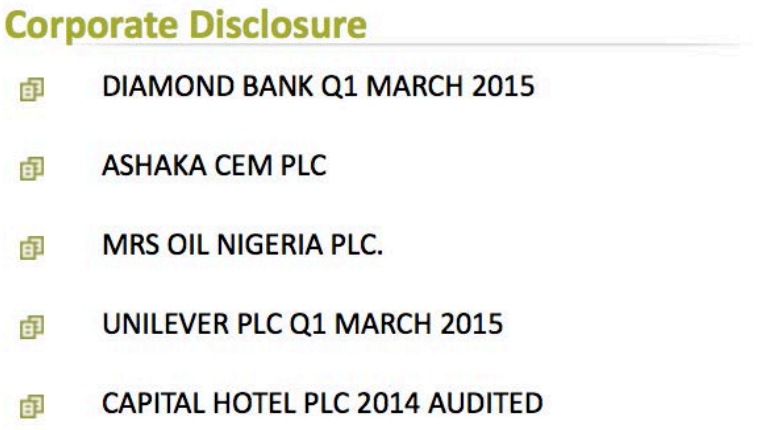
But what if Wikipedia is not enough?

Wikipedia (or its derivatives) may not contain pages about the specific domain at hand.



Or many smaller publicly traded companies…

Source: http://www.nse.com.ng



We need appropriate external knowledge

Wikipedia is not sufficient for Bloomberg.

But we have much knowledge that can be used for similar purposes.

All we need to do is extract the same features from our data as we would from Wikipedia!

Companies and people, for now

In order to prove that we could build this technology we restricted our attention to companies and people disambiguation.

Let's see how we managed to build the Bloomberg Named Entity Disambiguation system for Bloomberg entities: BNED


3/ BLOOMBERG NED

Recipe for a Named Entity Disambiguation System

1. Obtain a set of entities for the domain, and aliases that represent the entities in text
2. Produce a description or context for the entities
3. Find data to extract features used for disambiguation
4. Train disambiguation model
5. Find or create data to evaluate quality

Recipe for a Named Entity Disambiguation System

1. Obtain a set of entities for the domain, and aliases that represent the entities in text

Bloomberg has huge databases with companies, people of interest, and names and aliases.

Yay!

Aliases are hard to enumerate, and may require both automatic generation as well as fuzzy matching such as LSH (Nguyen et al 2014).

Blue links in Bloomberg News

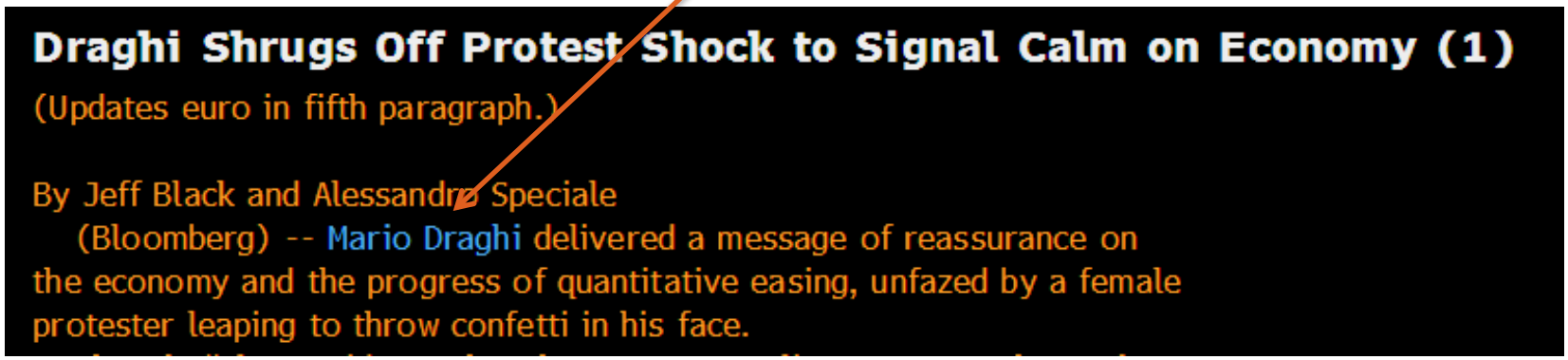
We can identify entities:

some_bloomberg_id

Blue links in Bloomberg News

We can identify people:

some_bloomberg_id



Draghi Shrugs Off Protest Shock to Signal Calm on Economy (1)

(Updates euro in fifth paragraph.)

By Jeff Black and Alessandro Speciale

(Bloomberg) -- Mario Draghi delivered a message of reassurance on the economy and the progress of quantitative easing, unfazed by a female protester leaping to throw confetti in his face.

Recipe for a Named Entity Disambiguation System

1. Obtain a set of entities for the domain, and aliases that represent the entities in text
2. Produce a description or context for the entities
3. Find data to extract features used for disambiguation
4. Train disambiguation model
5. Find or create data to evaluate quality

Recipe for a Named Entity Disambiguation System

Once again, Bloomberg company description snippets provided a useful start.

For people, (and in general) use what you have. In this case, education, career, public holdings…

If entities within an ontology or taxonomy, use the classes or categories to which they belong.

Recipe for a Named Entity Disambiguation System

1. Obtain a set of entities for the domain, and aliases that represent the entities in text
2. Produce a description or context for the entities
3. Find data to extract features used for disambiguation
4. Train disambiguation model
5. Find or create data to evaluate quality

Recipe for a Named Entity Disambiguation System

Popularity: probability that mention m is the surface form for entity e.

- Annotated data required: remember those blue links? Otherwise, manual annotation or crowdsourcing.
- "**Kashmir**"




?


Milne et al. 2008

Entity-entity coherence: semantic similarity between two entities:

- Based on entity context similarity, but can include other measures such as entity co-occurrence. **E.g., "Rome and Manchester"**

Recipe for a Named Entity Disambiguation System

1. Obtain a set of entities for the domain, and aliases that represent the entities in text
2. Produce a description or context for the entities
3. Find data to extract features used for disambiguation
4. Train disambiguation model
5. Find or create data to evaluate quality

Recipe for a Named Entity Disambiguation System

Evaluation data was our first concern: data sets in the wild are all annotated with Wikipedia!

We could reuse the blue links of Bloomberg News as well for evaluation.

But we started with manual annotation first! (Annotation pipeline).

Caveats:

(Bloomberg) -- Mario Draghi delivered a message of reassurance on the economy and the progress of quantitative easing, unfazed by a female protester leaping to throw confetti in his face.

- Annotations are for mentions of entities in their *canonical form*.
- Sparse annotations

Text corpora used for NED at Bloomberg

Bloomberg articles from the last 10 years (in the hundreds of millions).

Two sets

- Training set: 100,000 articles selected uniformly at random
- Evaluation set: 100 articles/month (12,000), from which we selected 995 articles.

Training
100,000 articles
uniform random sampling

Evaluation
995 articles
6,244 annotated mentions
85% people
15% companies
62% unambiguous

Recipe for a Named Entity Disambiguation System

1. Obtain a set of entities for the domain, and aliases that represent the entities in text
2. Produce a description or context for the entities
3. Find data to extract features used for disambiguation
4. Train disambiguation model
5. Find or create data to evaluate quality

Single-entity disambiguation

Single-entity disambiguation: disambiguate mentions in the text one by one.

Let's first talk about context similarity.

Define window of text around the mention, and compare it with the *description* of each entity.

For example:

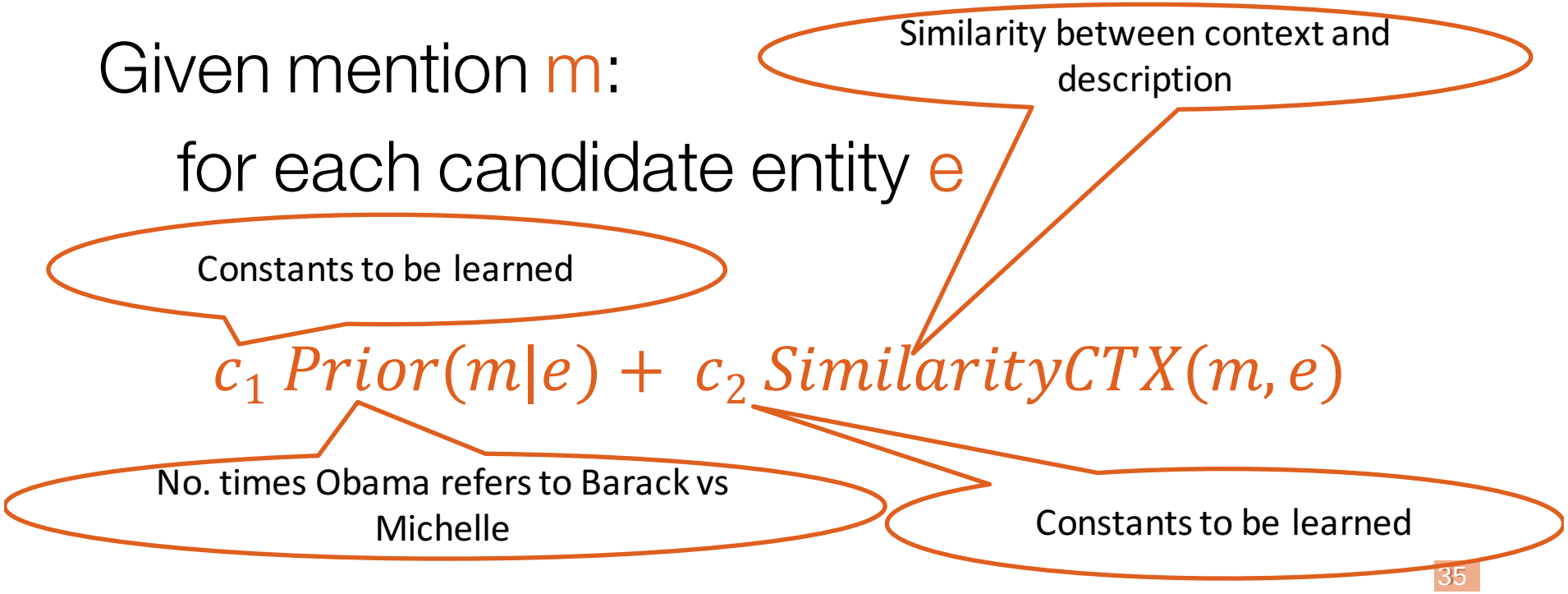
- Jaccard
- Cosine
- Overlap

Single-entity disambiguation

Single-entity disambiguation: disambiguate mentions in the text one by one.

Given mention m:

for each candidate entity e



Single-entity disambiguation

Single-entity disambiguation: disambiguate mentions in the text one by one.

Made up example: "Castro in San Francisco, surprise visit for Obama".

Joint-entity disambiguation

Joint disambiguation: not one by one, but assign entities all together using adding the *semantic relatedness* of chosen entities.

Pros: Impressive accuracy in difficult mentions. It models better how human beings disambiguate.

Cons: The complexity of finding a solution becomes intractable very fast (*NP-HARD* see Kulkarni et al. 2009). Instead of having a bound on the search space of (single-entity disambiguation), we now have one of .

M: number of mentions
E: upper bound on the number of candidate entities per mention

Joint-entity disambiguation

Joint disambiguation: not one by one, but assign entities all together using adding the *semantic relatedness* of chosen entities.

AIDA (Hoffart et al. 2011) and AIDA-light (Nguyen et al. 2014) :

- greedy approximate algorithms (heuristics)
- allow the addition of new features.
- features combined into weights of a graph with edges from mentions-to-entities and entities-to-entities.

Joint-entity disambiguation

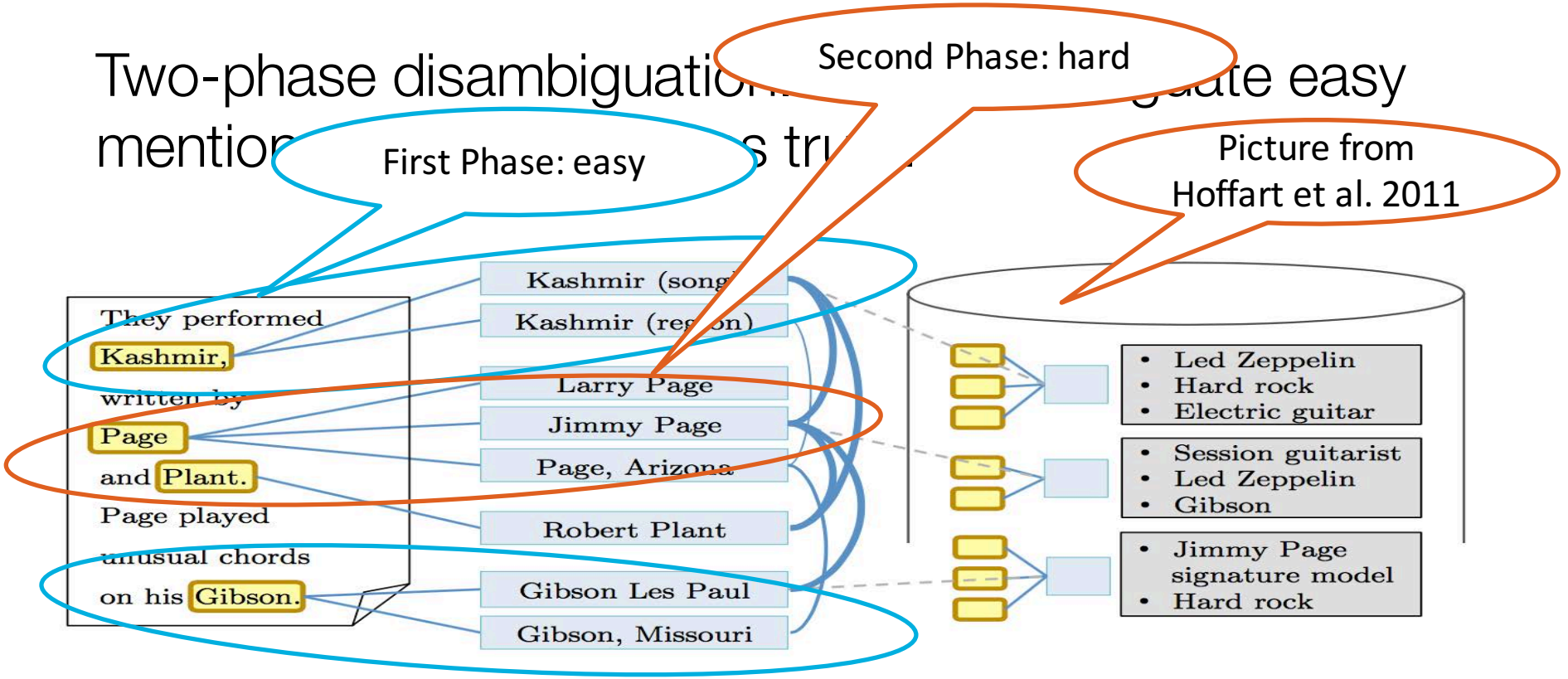
Single-entity disambiguation: disambiguate mentions in the text one by one.

Made up example: "Castro in San Francisco, surprise visit for Obama".

BNED core is based on AIDA-light

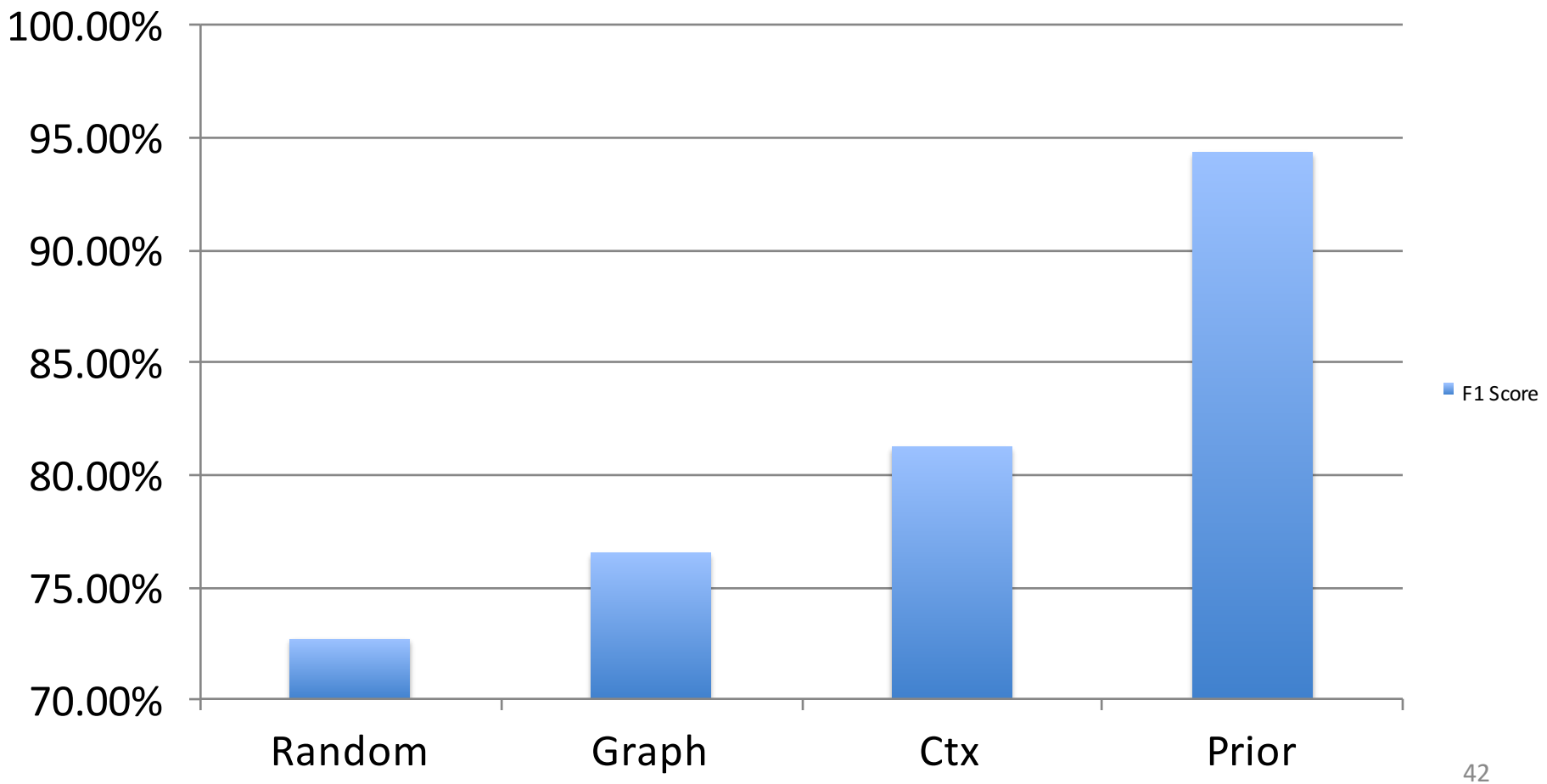
From Nguyen et al. 2014: fast system for performing joint disambiguation.

Two-phase disambiguation: ... easy mentions tr...

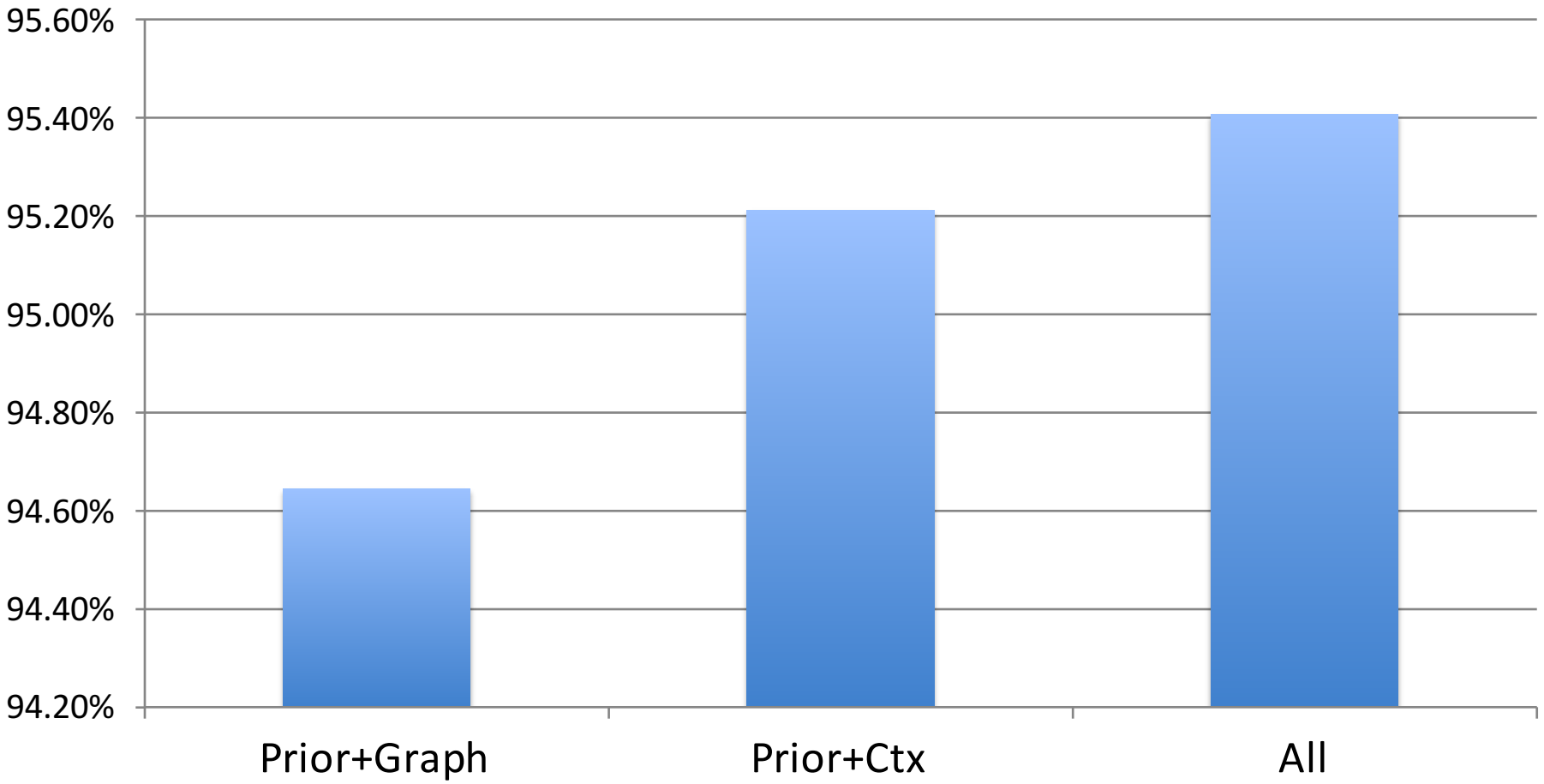


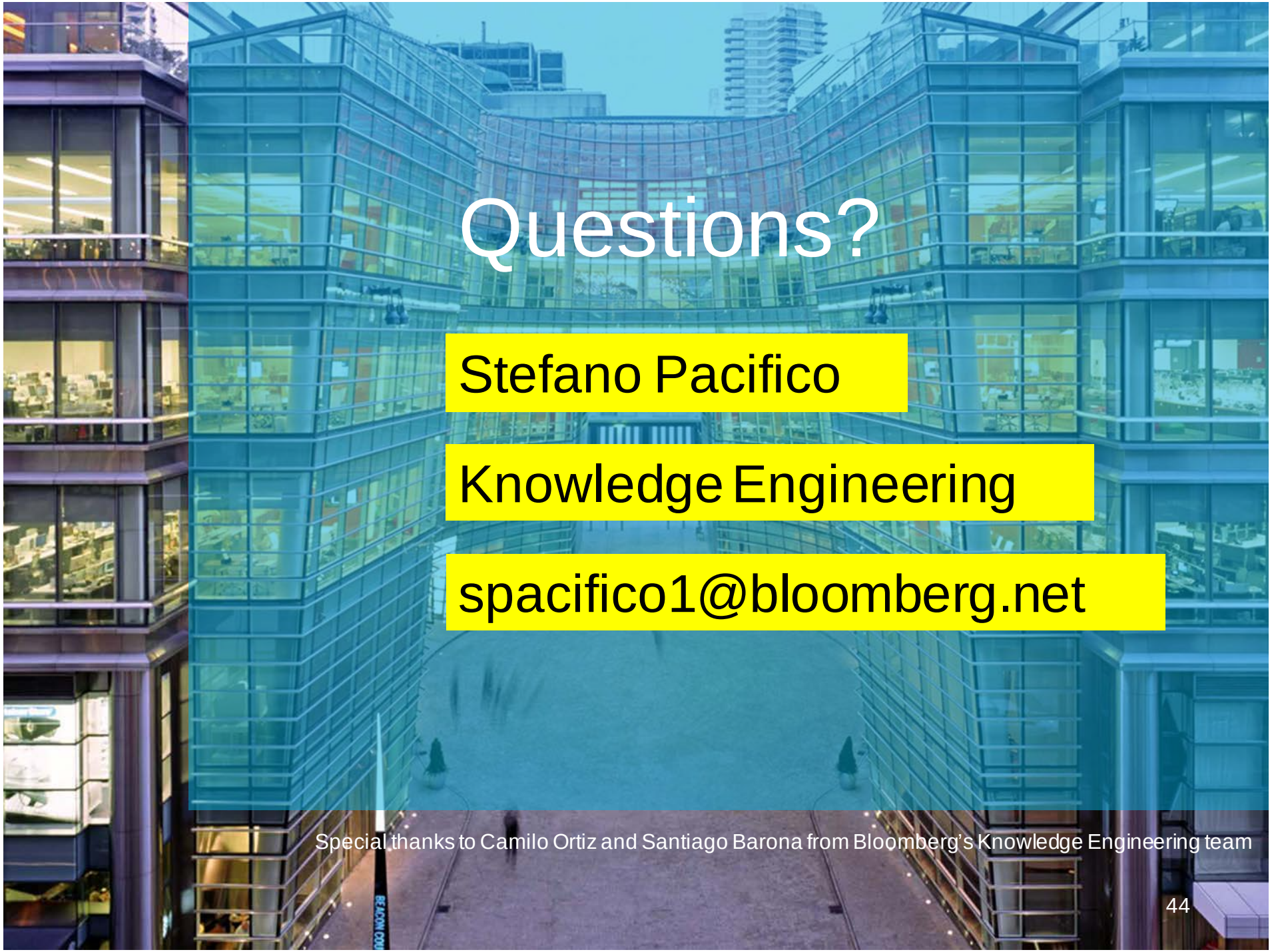

4/
RESULTS

SINGLE	**Random**	Baseline
	Ctx	Mention – Entity Context Similarity
	Prior	Prior probability (Popularity)
JOINT	**Graph**	Entity-Entity Context Similarity + Co-occurrence
	All	All the features together



SINGLE	**Random**	Baseline
	Ctx	Mention – Entity Context Similarity
	Prior	Prior probability (Popularity)
JOINT	**Graph**	Entity-Entity Context Similarity + Co-occurrence
	All	All the features together





Special thanks to Camilo Ortiz and Santiago Barona from Bloomberg's Knowledge Engineering team