Low-Resource NLP

Stephen Mayhew

CIS530

April 11, 2018

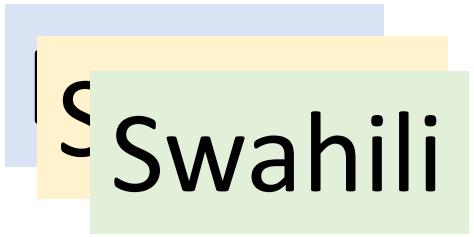
My Research

- Named Entity Recognition at different resource levels
- Many languages, few resources
- Recent work:
  - Cross-lingual NER via Wikification
  - Cheap translation for NER
  - Non-speaker annotation for NER
  - NER with Partial Annotations

Lecture Overview

1. What is and why talk about Low-Resource NLP?
2. Common approaches
    1. Parallel projection
    2. Direct transfer
    3. Strange hybrid of the two?
    4. Cross-lingual embeddings

What is “Low-Resource”?



Penn Treebank

Gutenberg Ebooks

Question Answering

Reddit

Sentiment Treebanks

WordNet

Scientific literature

Enron Email

Story Completion

NYT Corpus

FrameBank

Dialogue

Universal Dependencies

Parallel Text

NER Train data

Bible

Yelp Reviews

Thousands of words of text

PropBank

Wikipedia

Different things for different tasks

- NER
  - Only monolingual text (no name annotations)
- Machine Translation
  - “Only” 10,000 parallel sentences (instead of millions)
- Parsing
  - Only POS tagging (no parse trees in native text)
- Speech Recognition
  - Only 20 hours of recorded speech (instead of hundreds of hours)

Why are we interested?

- Most languages are low-resource
- About 7K languages
- About 3.5K are written
- Useful for English-speaking world
  - Understand what is happening in that area
- Useful for that area
  - Better tools, better management of information

What can we do?

- Unsupervised learning
  - This is a myth!

- All learning has some signal
  - Word2vec: explicitly supervised
  - Word alignments: supervision in the form of counts

Low Resource NLP:
Searching for Signals

What can we do?

Train on other language data

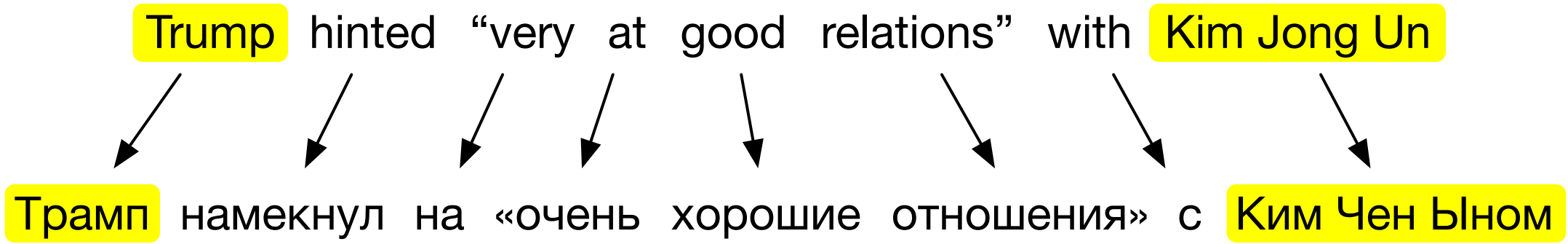
Create data in target language

Parallel Projection

The proposal will not now be implemented

Word alignments are free
(given enough parallel text)

Les propositions nes seront pas mises en application maintenant

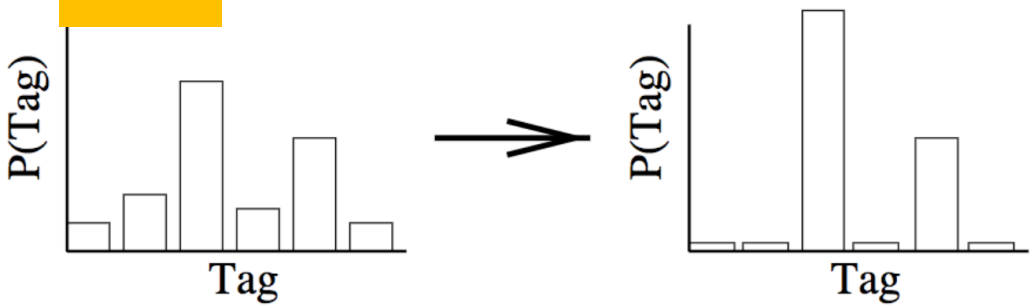


Yarowsky 2001, McDonald 2011, Zitouni 2008, Kim 2012, Wang 2014, Zhang 2016

Projection – cleaning up

- Alignments are often wrong

POS



NP Chunking

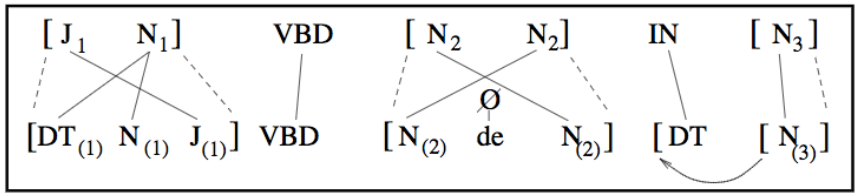


Figure 5: Standard NP projection scenarios.

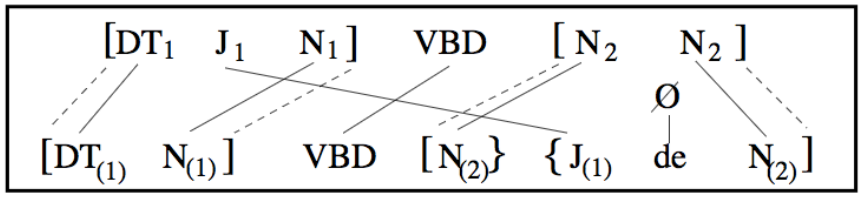


Figure 6: Problematic NP projection scenarios.

NER

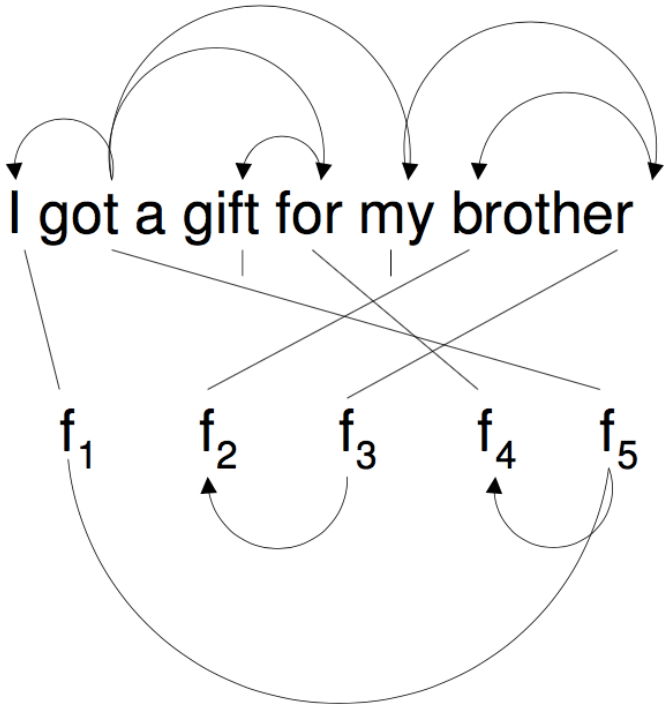
$$P(\text{NEclass}_j|\text{FW}) = \sum_i P(\text{NEclass}_j|\text{EW}_i)\, P_a(\text{EW}_i|\text{FW})$$

$$P(\text{PLACE}|\text{Corée}) = P(\text{PLACE}|\text{Korea})\, P_a(\text{Korea}|\text{Corée}) + \ldots$$

Yarowsky 2001, McDonald 2011, Zitouni 2008, Kim 2012, Wang 2014, Zhang 2016

Projecting a Parser

Direct Correspondence Assumption (DCA): Given a pair of sentences E and F that are (literal) translations of each other with syntactic structures $Tree_E$ and $Tree_F$, if nodes x_E and y_E of $Tree_E$ are aligned with nodes x_F and y_F of $Tree_F$, respectively, and if syntactic relationship $R(x_E, y_E)$ holds in $Tree_E$, then $R(x_F, y_F)$ holds in $Tree_F$.



English dependencies

English sentence

Alignment

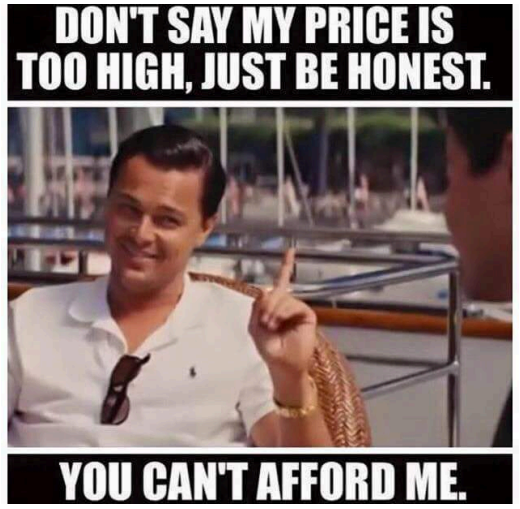
Foreign language sentence

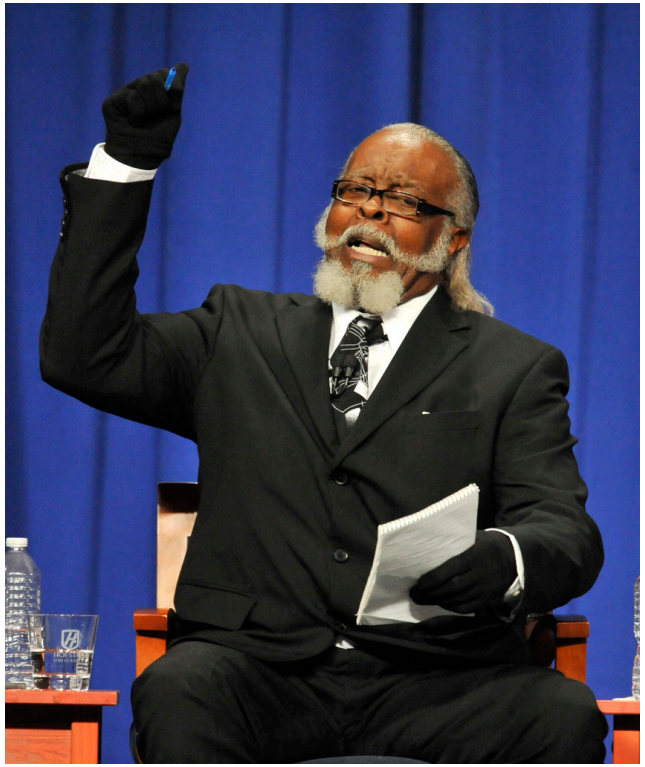
Projected dependencies

Hwa et al. 2004

Parallel projection

THAT'S TOO EXPENSIVE





Direct Transfer

- Stupid because you train in one language and test in another.
- For NER:
  - Might work for English to Spanish
  - Doesn't work for English to Tamil
  - Language similarity matters.
- Parsing:
  - Delexicalization (assume part of speech tags)

Direct Transfer for Parsing

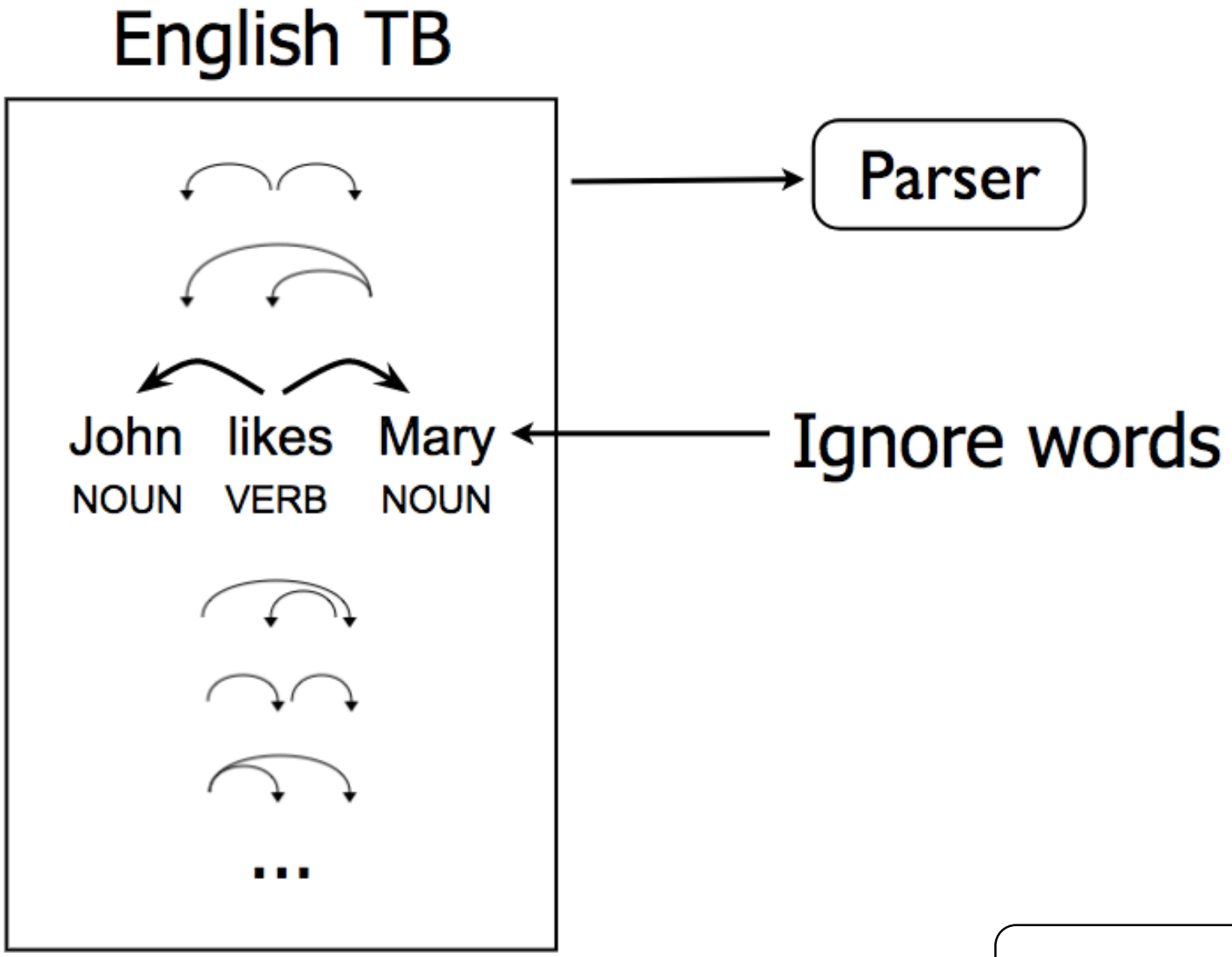
Multi-Source Transfer of Delexicalized Dependency Parsers

Ryan McDonald
Google
New York, NY
ryanmcd@google.com

Slav Petrov
Google
New York, NY
slav@google.com

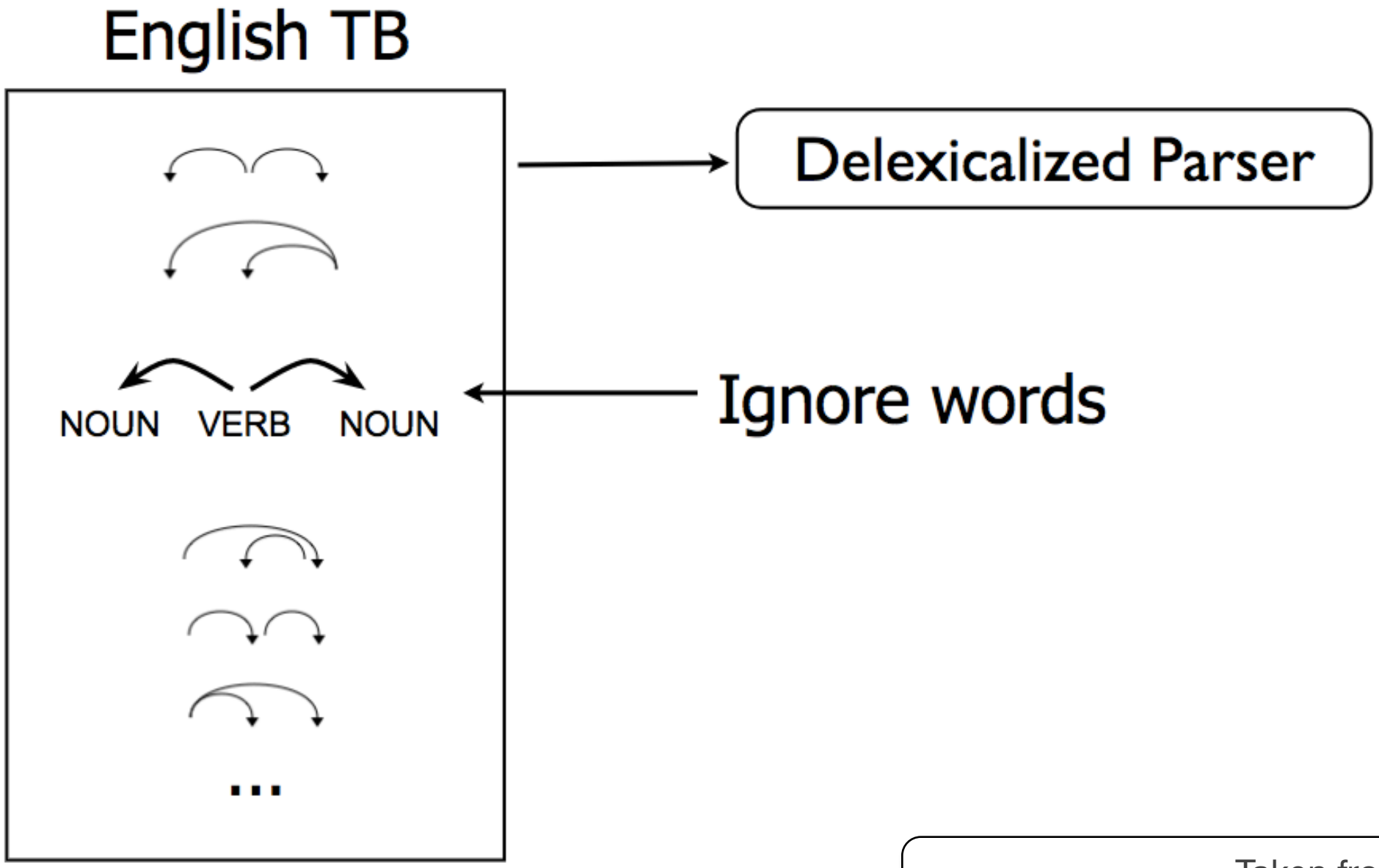
Keith Hall
Google
Zürich
kbhall@google.com

In EMNLP 2011



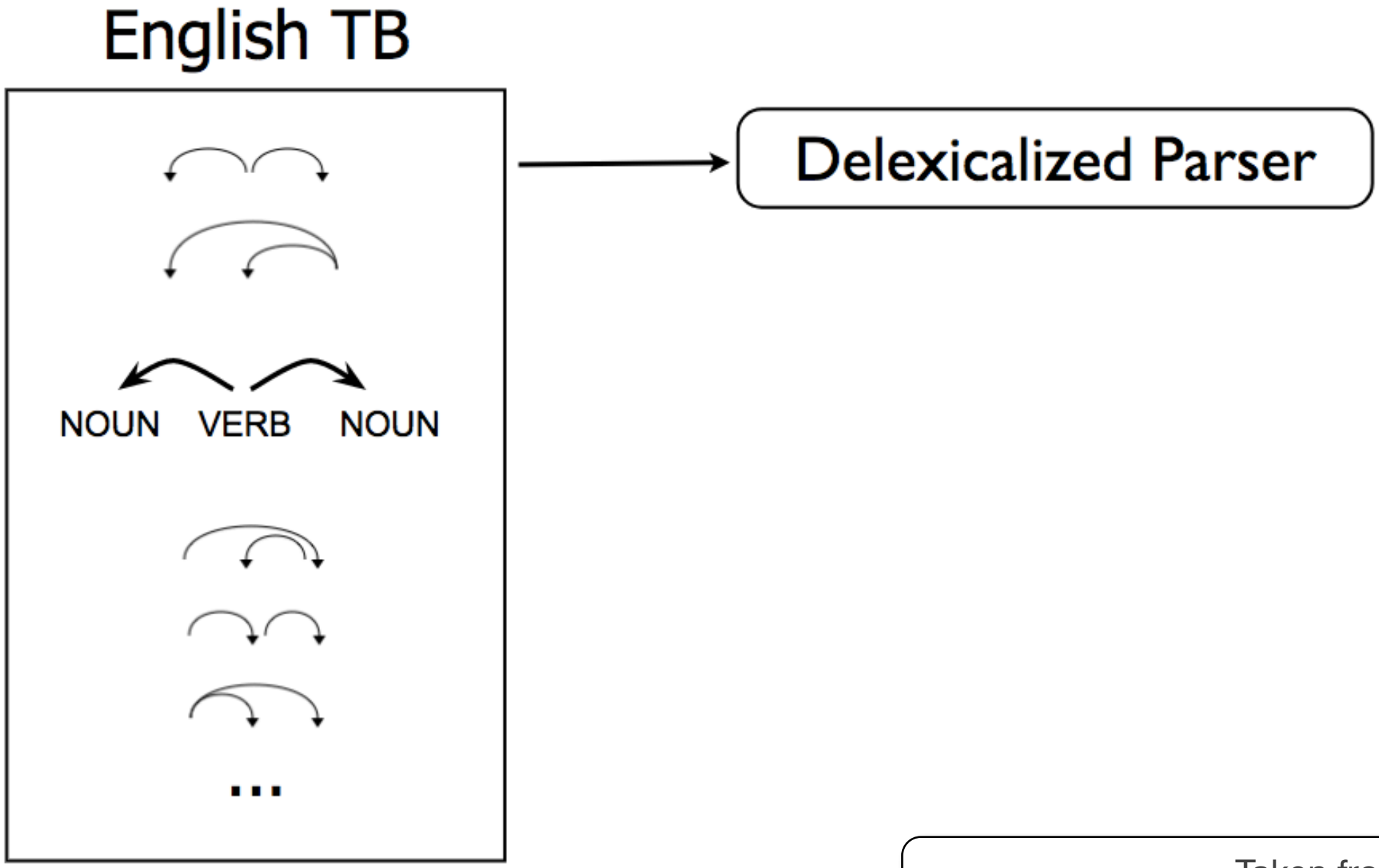
McDonald, Petrov, Hall 2011

Taken from
http://www.petrovi.de/data/emnlp11aslides.pdf



McDonald, Petrov, Hall 2011

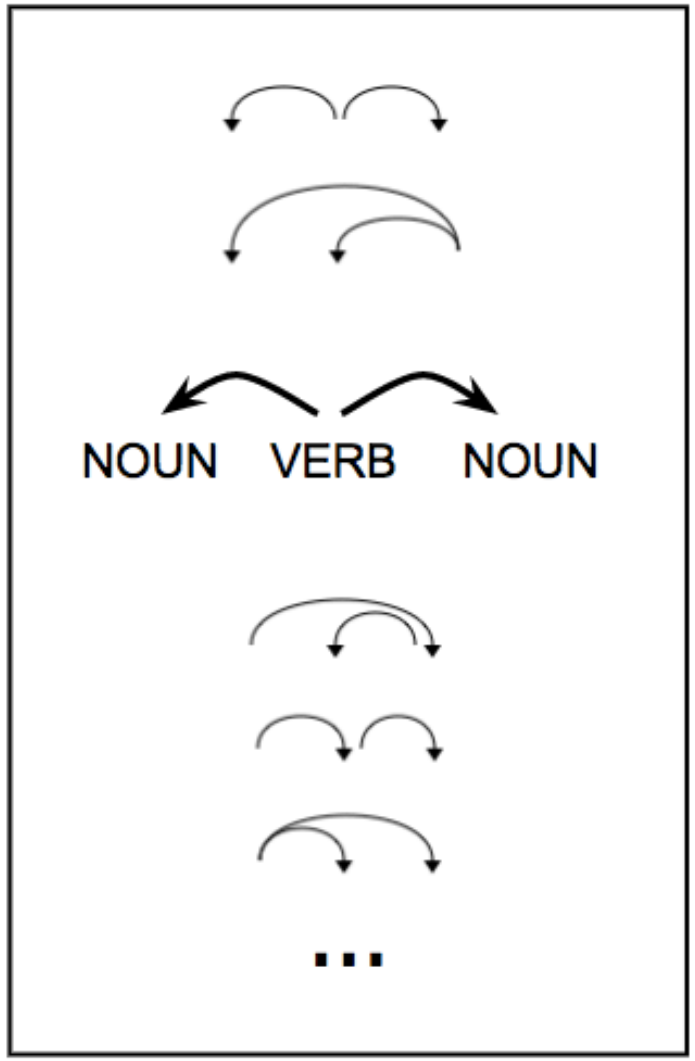
Taken from
http://www.petrovi.de/data/emnlp11aslides.pdf



Taken from
http://www.petrovi.de/data/emnlp11aslides.pdf

McDonald, Petrov, Hall 2011

English TB



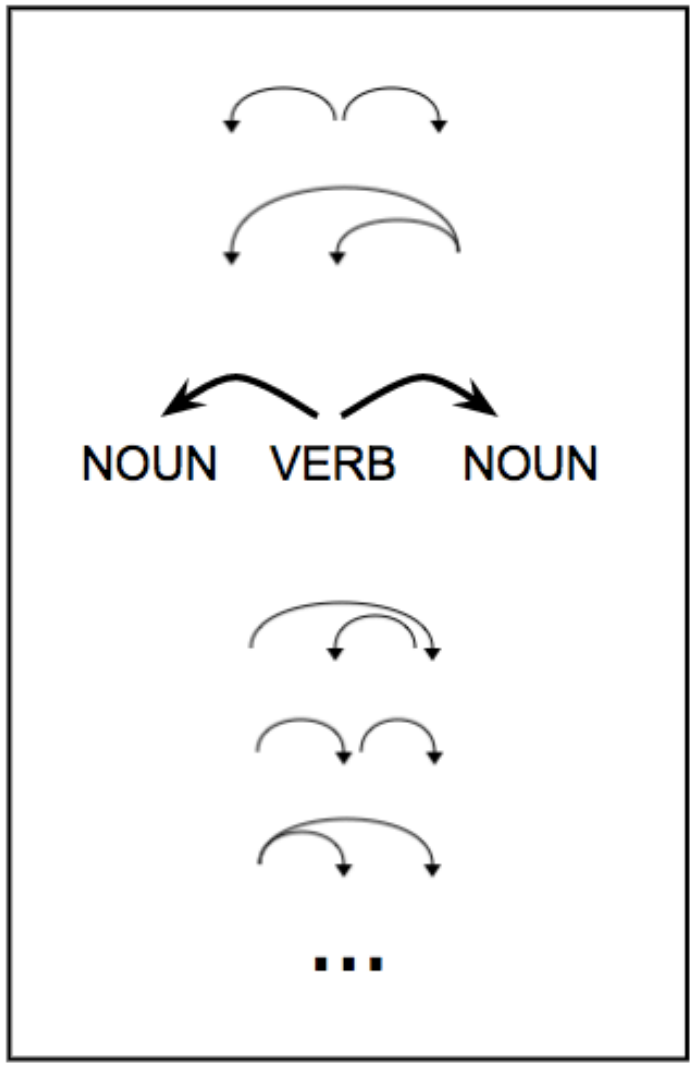
Delexicalized Parser

Ο Γιαννις βλεπει την Μαρια
DET NOUN VERB DET NOUN

Taken from
http://www.petrovi.de/data/emnlp11aslides.pdf

McDonald, Petrov, Hall 2011

English TB



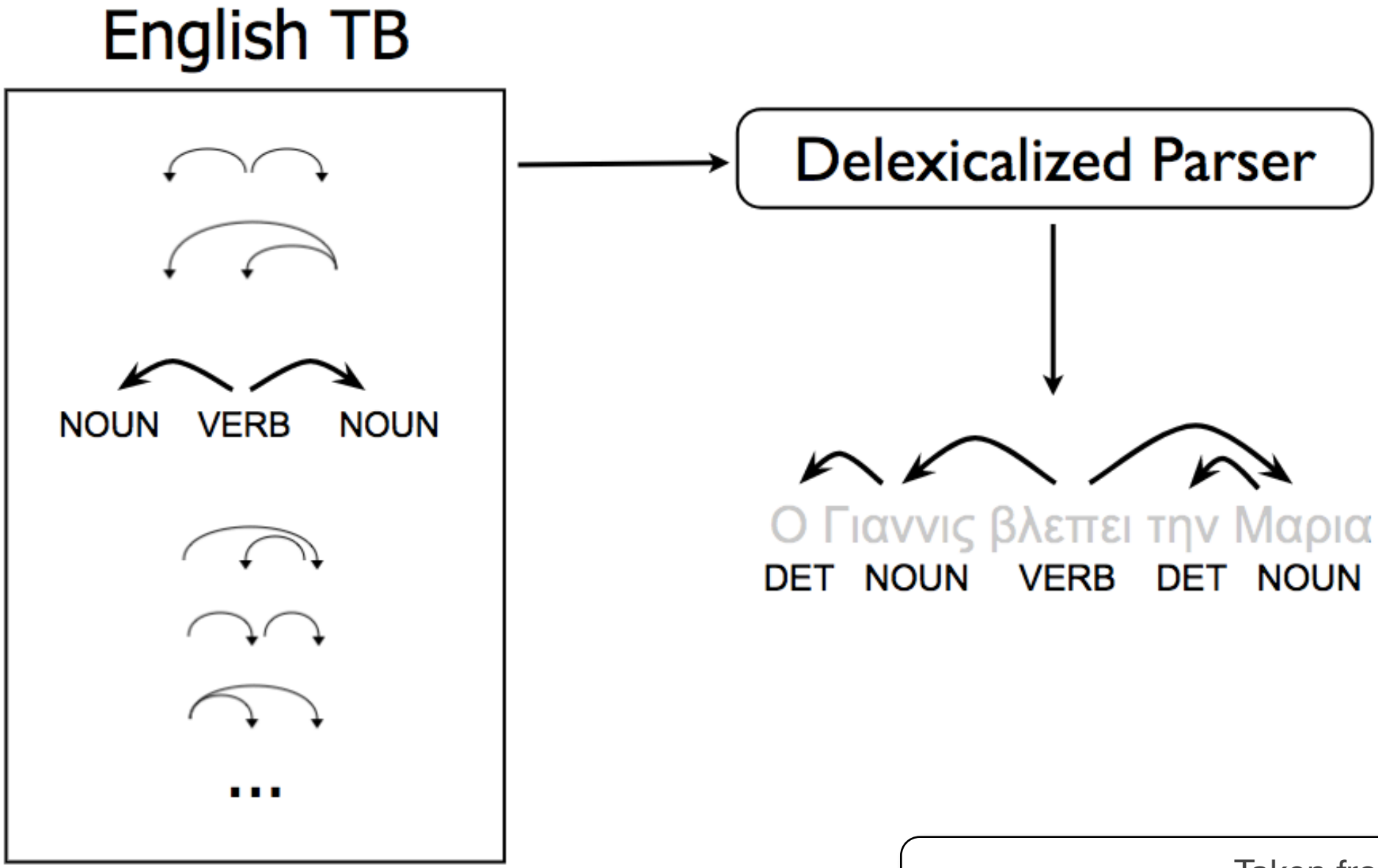
Delexicalized Parser

Ο Γιαννις βλεπει την Μαρια

DET NOUN VERB DET NOUN

Taken from
http://www.petrovi.de/data/emnlp11aslides.pdf

McDonald, Petrov, Hall 2011



McDonald, Petrov, Hall 2011

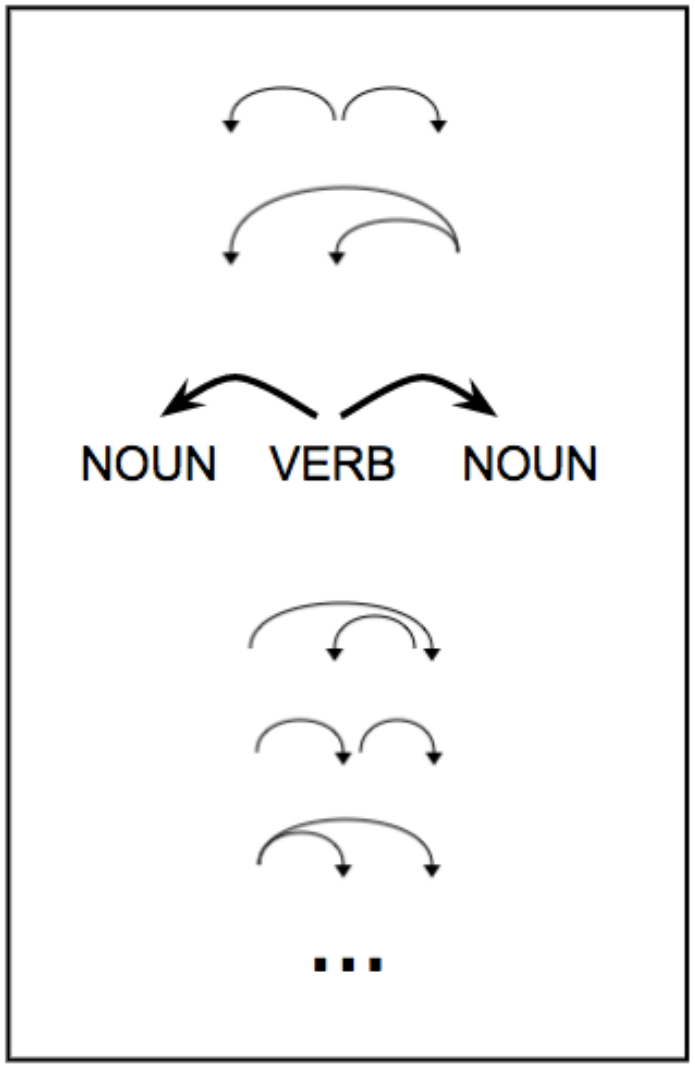
Taken from
http://www.petrovi.de/data/emnlp11aslides.pdf

Results

	gold-POS			pred-POS		
	DMV	en-dir.	en-proj.	DMV	en-dir.	en-proj.
da	33.4	45.9	48.2	18.4	44.0	45.5
de	18.0	47.2	50.9	30.3	44.7	47.4
el	39.9	63.9	66.8	21.2	63.0	65.2
es	28.5	53.3	55.8	19.9	50.2	52.4
it	43.1	57.7	60.8	37.7	53.7	56.3
nl	38.5	60.8	67.8	19.9	62.1	66.5
pt	20.1	69.2	71.3	21.0	66.2	67.7
sv	44.0	58.3	61.3	33.8	56.5	59.7
avg	33.2	57.0	60.4	25.3	55.0	57.6

McDonald, Petrov, Hall 2011

German TB



Delexicalized Parser

Ο Γιαννις βλεπει την Μαρια

DET NOUN VERB DET NOUN

McDonald, Petrov, Hall 2011

Multisource Results

Diagonal is always the best.

Target Test Language	Source Training Language: da	de	el	en	es	it	nl	pt	sv
da	**79.2**	45.2	44.0	45.9	45.0	48.6	46.1	48.1	47.8
de	34.3	**83.9**	53.2	47.2	45.8	53.4	55.8	55.5	46.2
el	33.3	52.5	**77.5**	63.9	41.6	59.3	57.3	58.6	47.5
en	34.4	37.9	45.7	**82.5**	28.5	38.6	43.7	42.3	43.7
es	38.1	49.4	57.3	53.3	**79.7**	68.4	51.2	66.7	41.4
it	44.8	56.7	66.8	57.7	64.7	**79.3**	57.6	69.1	50.9
nl	38.7	43.7	62.1	60.8	40.9	50.4	**73.6**	58.5	44.2
pt	42.5	52.0	66.6	69.2	68.5	74.7	67.1	**84.6**	52.1
sv	44.5	57.0	57.8	58.3	46.3	53.4	54.5	66.8	**84.8**

English is rarely the second best.

McDonald, Petrov, Hall 2011

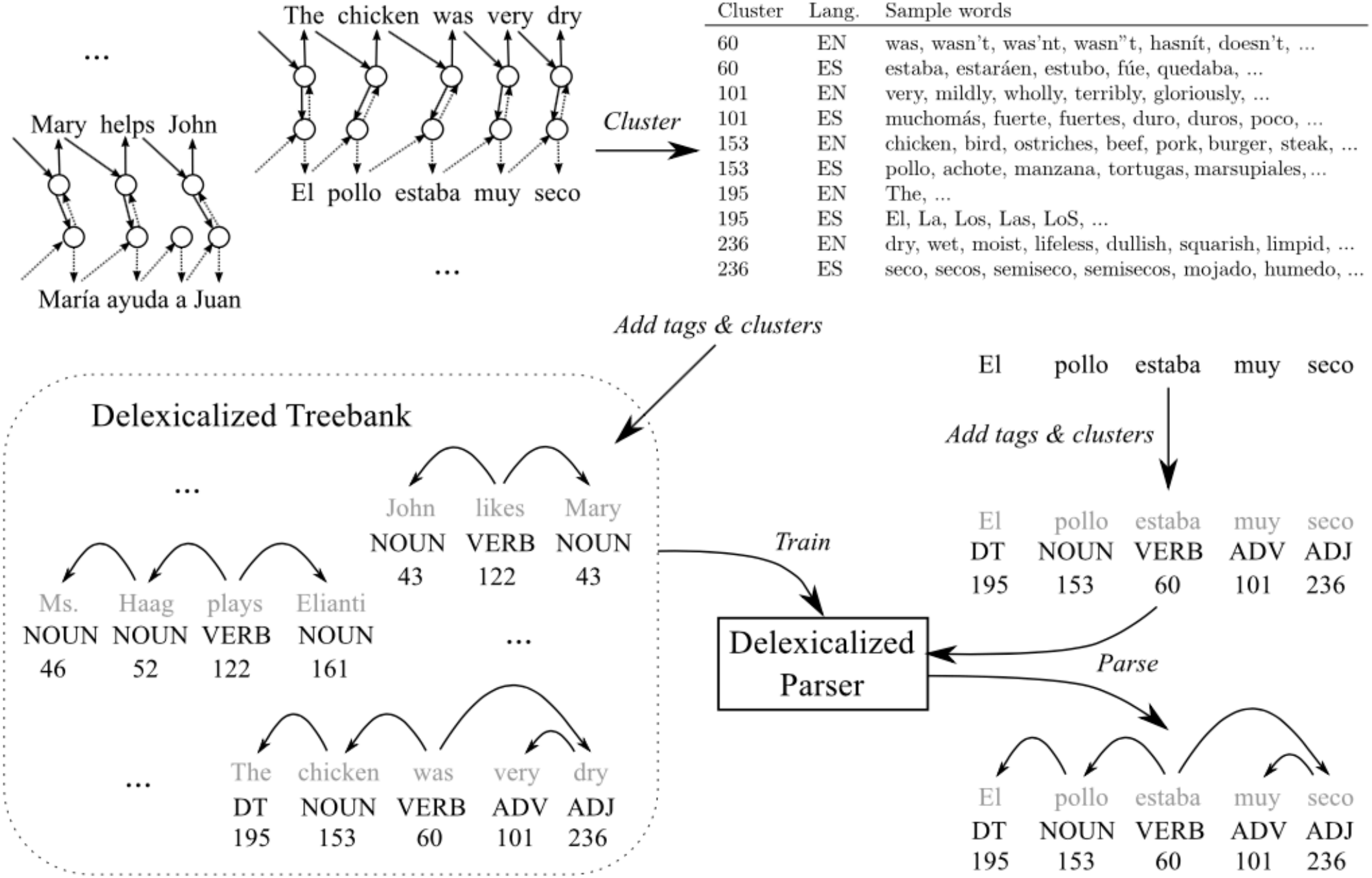
Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure

Oscar Täckström*
SICS / Uppsala University
Kista / Uppsala, Sweden
oscar@sics.se

Ryan McDonald
Google
New York, NY
ryanmcd@google.com

Jakob Uszkoreit
Google
Mountain View, CA
uszkoreit@google.com

In NAACL 2012

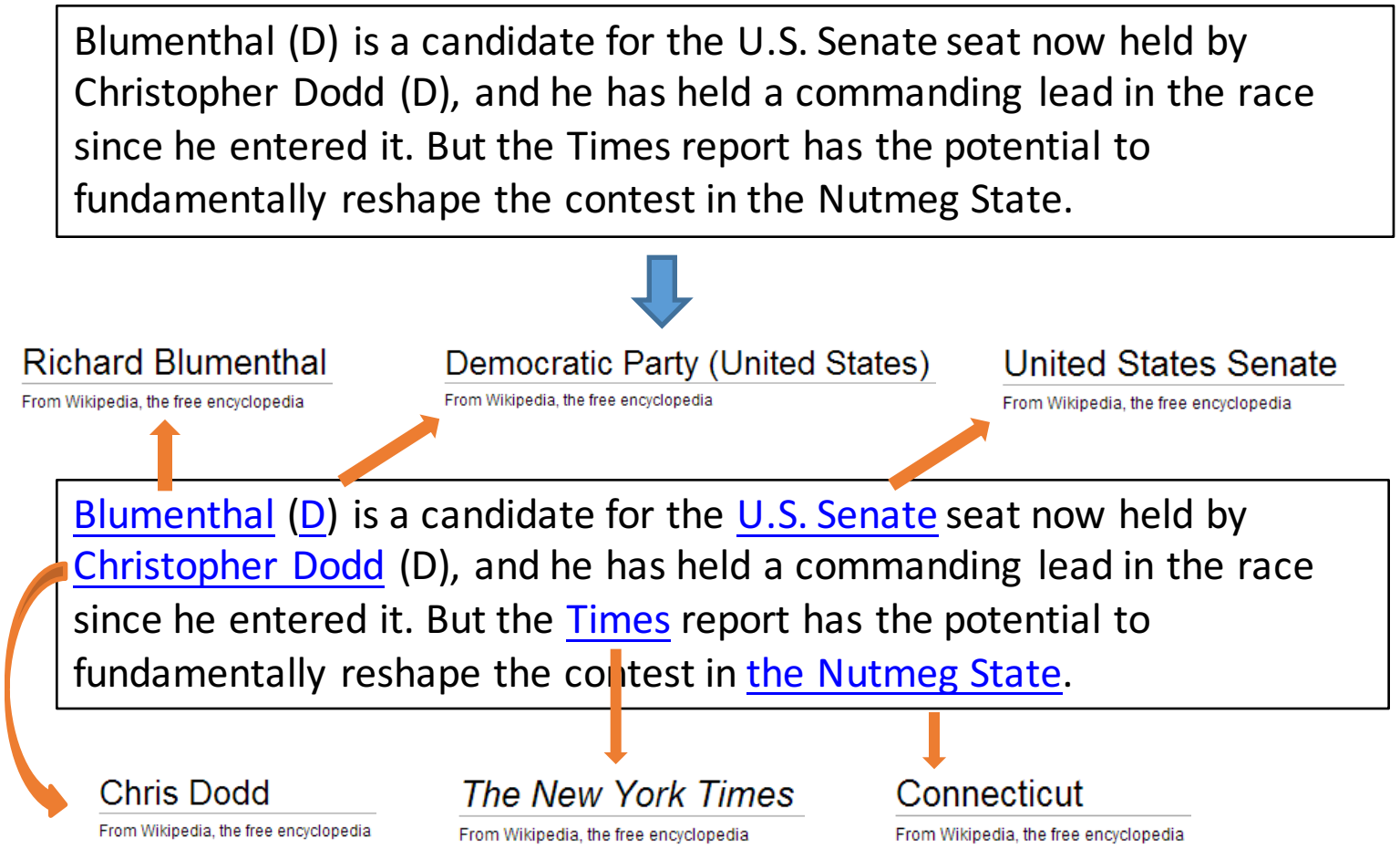

The chicken was very dry
El pollo estaba muy seco
Mary helps John
María ayuda a Juan
Cluster
Cluster Lang. Sample words
60 EN was, wasn't, was'nt, wasn"t, hasnít, doesn't, ...
60 ES estaba, estaráen, estubo, fúe, quedaba, ...
101 EN very, mildly, wholly, terribly, gloriously, ...
101 ES muchomás, fuerte, fuertes, duro, duros, poco, ...
153 EN chicken, bird, ostriches, beef, pork, burger, steak, ...
153 ES pollo, achote, manzana, tortugas, marsupiales, ...
195 EN The, ...
195 ES El, La, Los, Las, LoS, ...
236 EN dry, wet, moist, lifeless, dullish, squarish, limpid, ...
236 ES seco, secos, semiseco, semisecos, mojado, humedo, ...
Add tags & clusters
Delexicalized Treebank
John likes Mary
NOUN VERB NOUN
43 122 43
Ms. Haag plays Elianti
NOUN NOUN VERB NOUN
46 52 122 161
The chicken was very dry
DT NOUN VERB ADV ADJ
195 153 60 101 236
Train
Delexicalized Parser
El pollo estaba muy seco
Add tags & clusters
El pollo estaba muy seco
DT NOUN VERB ADV ADJ
195 153 60 101 236
Parse
El pollo estaba muy seco
DT NOUN VERB ADV ADJ
195 153 60 101 236

Cross-Lingual Named Entity Recognition via Wikification

Chen-Tse Tsai, Stephen Mayhew, and **Dan Roth**
University of Illinois, Urbana-Champaign
201 N. Goodwin, Urbana, Illinois, 61801
{ctsai12, mayhew2, danr}@illinois.edu

In CoNLL 2016

Wikification [Ratinov and Roth 2011, Cheng and Roth 2013]



Cross-Lingual Wikification [Tsai and Roth, 2016]

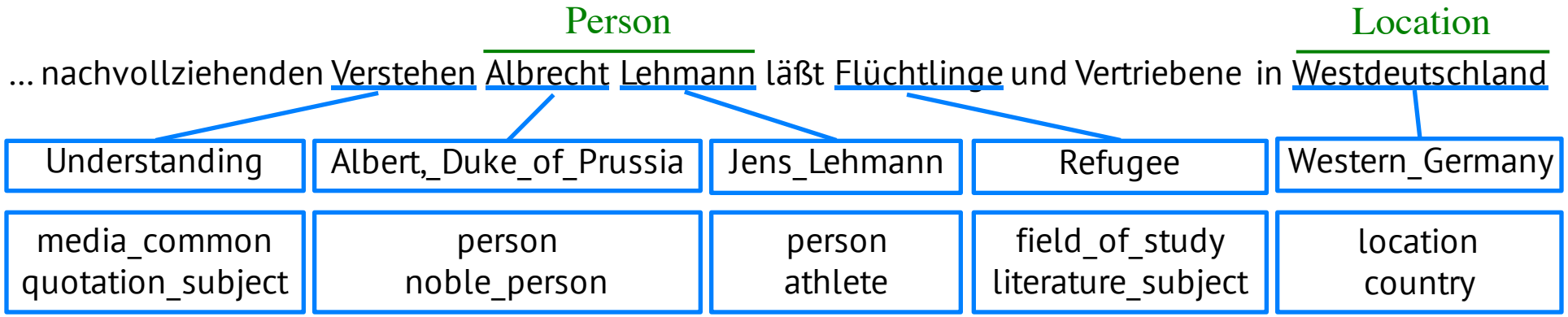
- Given mentions in a non-English document, find the corresponding titles in the English Wikipedia

  Tayvan, ABD ve İngiltere'de hukuk okuması, Tsai'ye bir LL.B. kazandırdı

- Grounding to the titles in the intersection of the target language and English Wikipedia
  - Smaller Wikipedia, poorer coverage

Key Idea

- Cross-lingual wikifier generates good language-independent features for NER by grounding n-grams



- Words in any language are grounded to the English Wikipedia.

Wikifier Features

- We ground every n-grams (n<5) to the English Wikipedia

- Features for each word
  - The FreeBase types and Wikipedia categories of the top two titles of the current, previous, and next word
  - The FreeBase types of the n-grams covering the word

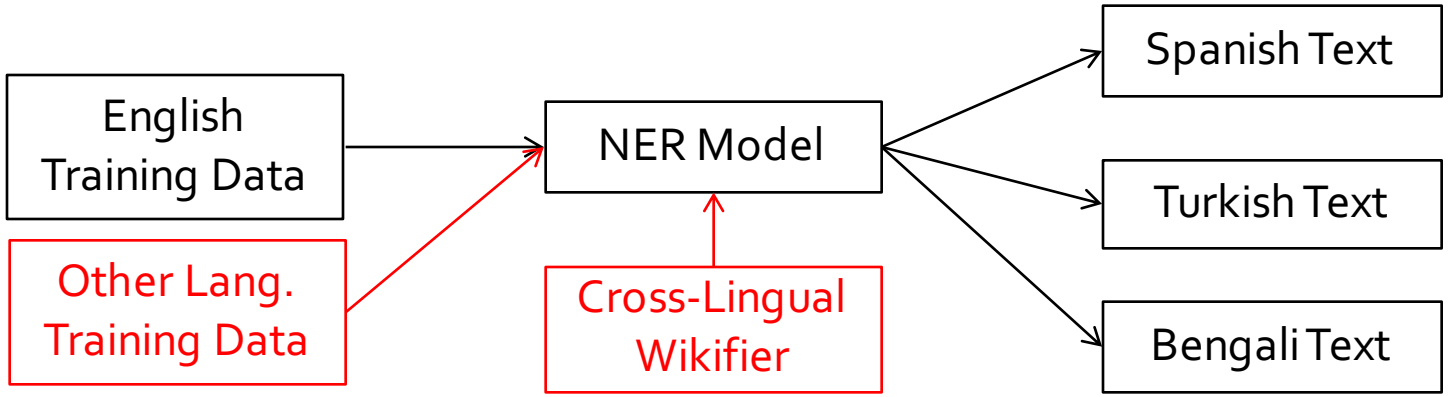
Algorithm

Train:

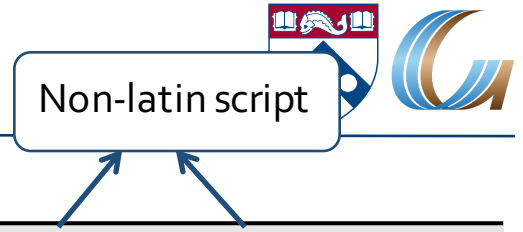
1. Wikify English training data.
2. Train NER model with wikifier results as language independent features

Test:

1. Wikify target language test data
2. Predict with annotated NER model, using wikifier results as features



Results



Approach	English	Dutch	German	Spanish	Turkish	Tagalog	Yoruba	Bengali	Tamil	Avg
Wiki Size	5.1M	1.9M	1.9M	1.3M	269K	64K	31K	42K	85K	
Eng. Intersec.	-	755K	964K	757K	169K	49K	30K	34K	51K	
Monolingual Experiments										
Wikifier only	71.57	57.02	49.74	60.13	52.84	51.02	29.35	47.78	38.05	50.8
Base + Gazet.	89.49	82.41	69.31	83.62	70.41	76.71	57.12	69.51	57.10	72.9
+Wikifier	89.92	84.49	73.13	83.87	73.86	77.64	57.60	71.15	60.02	74.7
Wikifier only		40.44	39.83	43.82	41.79	42.11	27.91	43.27	29.64	38.0
Base + Gazet.		50.26	34.47	54.59	30.21	64.06	34.37	3.25	0.30	33.8
+Wikifier		61.56	48.12	60.55	47.12	65.44	36.65	18.18	5.65	41.4
Täckström'12		58.4	40.4	59.3	-	-	-	-	-	-
Zhang'16		-	-	-	43.6	51.3	36.0	34.8	26.0	-

Multiple Training Languages

- Previous transfer model is trained on English

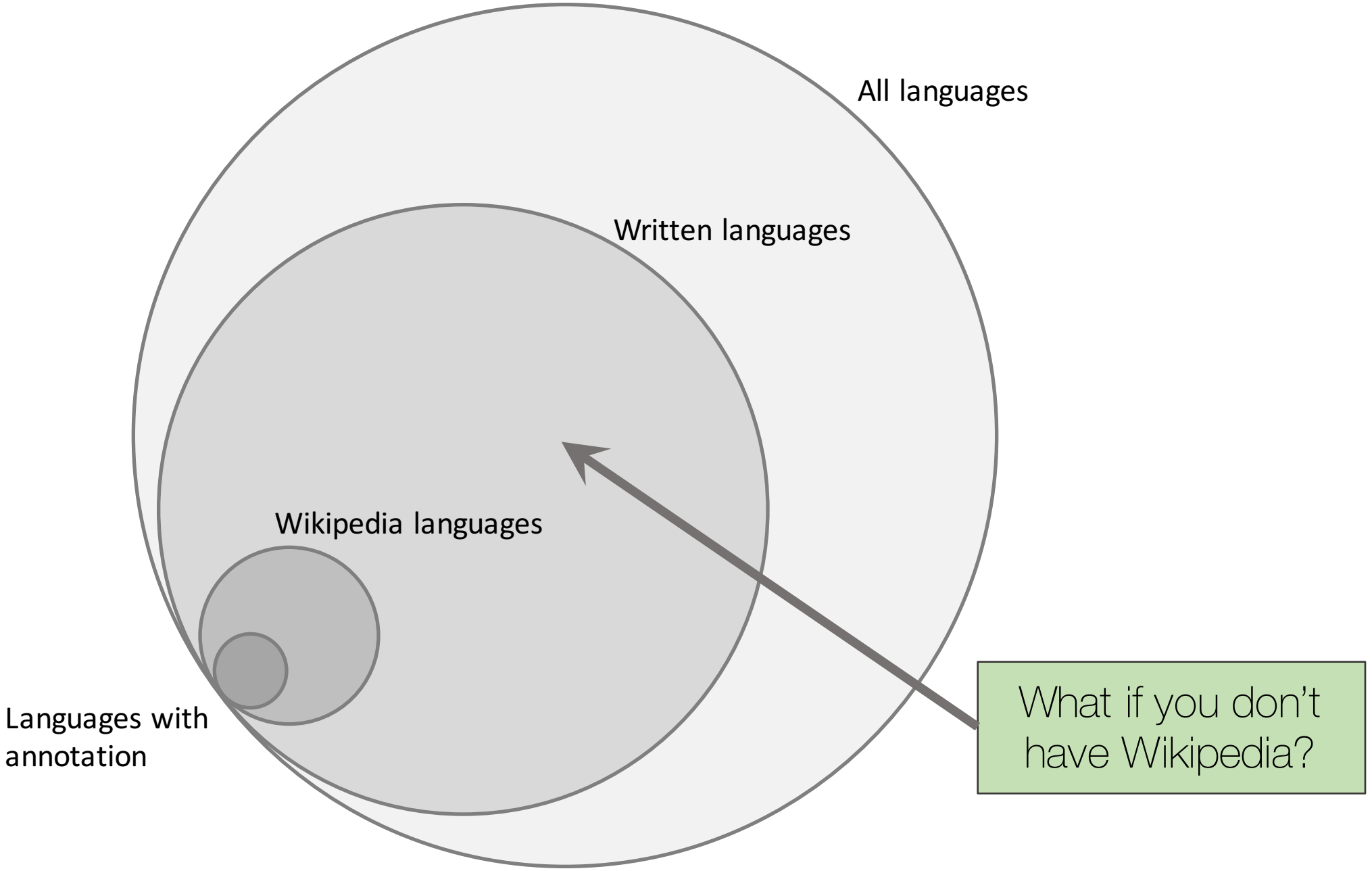
Training Languages	Turkish	Tagalog	Yoruba	Average
EN	47.12	65.44	36.65	49.74
EN+ES	44.85	66.61	37.57	49.68
EN+NL	48.34	66.09	36.87	50.43
EN+DE	49.47	64.10	35.14	49.57
EN+ES+NL+DE	49.00	66.37	38.02	51.13
ALL – Test Lang	49.83	67.12	37.56	51.50

  - EN: English, ES: Spanish, NL: Dutch, DE: German
  - ALL: EN, ES, NL, DE, Turkish, Tagalog, Yoruba

Cheap Translation for Cross-Lingual Named Entity Recognition

Stephen Mayhew Chen-Tse Tsai Dan Roth
University of Illinois, Urbana-Champaign
201 N. Goodwin
Urbana, Illinois, 61801
{mayhew2, ctsai12, danr}@illinois.edu

In EMNLP 2017


All languages
Written languages
Wikipedia languages
Languages with annotation
What if you don't have Wikipedia?

Lower-resource setting

We don't have:

- Annotators
- Parallel text
- Wikipedia

But we do have:

- High-resource data
- Bilingual lexicons
- Monolingual target text

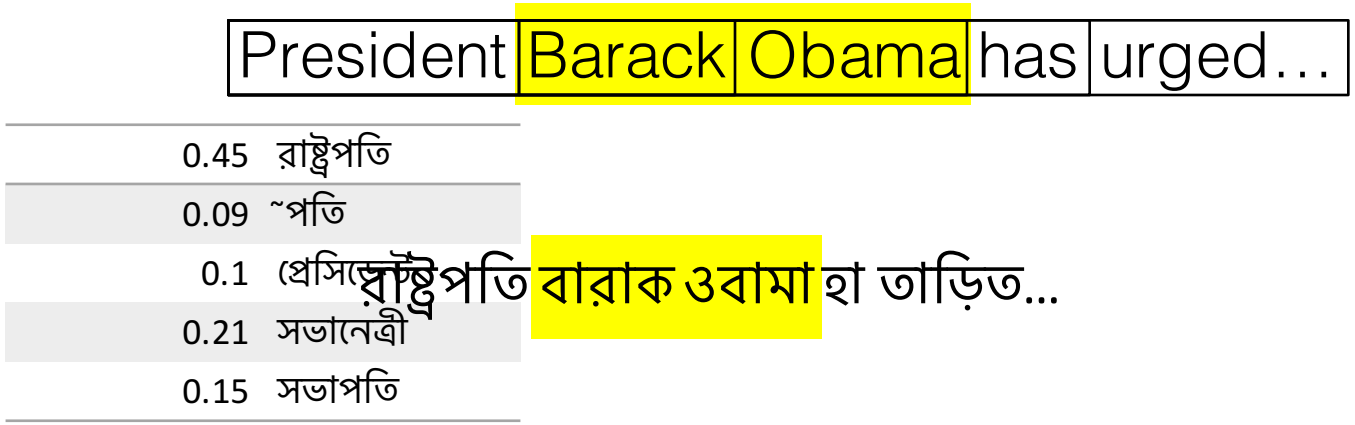
Cheap Translation!

Cheap Translation

Nicaraguan President Violeta Chamorro was due to fly to the United States

- Incorrect translations
- Incorrect morphology
- Incorrect word order

Algorithm


President Barack Obama has urged…
0.45 রাষ্ট্রপতি
0.09 ~পতি
0.1 প্রেসিডেন্ট
0.21 সভানেত্রী
0.15 সভাপতি
রাষ্ট্রপতি বারাক ওবামা হা তাড়িত…

Translation quality

Original

President Barack Obama has urged Democrats of all ethnic backgrounds to get out and vote for Hillary Clinton

Translated in Bengali

রাষ্ট্রপতি বারাক ওবামা হা তাড়িত সাম্যবাদী সবার জাতিগত পটভূমি পাওয়া বাইরে ও ভোট করে হিলারি ক্লিন্টন

Translate back (word level) with Google Translate

President Barack Obama Ha Struck Communist Public Ethnic Background Get Out Re Vote For Clinton Clinton

Translate back (sentence level) with Google

President Barack Obama Everything Ethnic Background Democrats To Call The Alight And Hillary Clinton For Vote

System	Task	BLEU
Online B	en-hi	17.31
(Durrani et al., 2014)	en-hi	12.47
Google Translate	en-hi	15.74
Ours	en-hi	2.43
(Durrani et al., 2014)	en-fr	36.62
Online B	en-fr	35.12
Google Translate	en-fr	28.35
Ours	en-fr	8.55

Multi-source translation

- Starting point: similar language.
- Good chance that there exists annotated data for a related language.
- Just need lexicons from related-target.
  - Transitive lexicons through English.
- We call this “Best Combination”

Target	Train languages
Dutch	English, German
German	English, Dutch
Spanish	English, Dutch
Turkish	English, Uzbek
Bengali	English, Hindi
Tamil	English, Malayalam
Yoruba	English, Hausa

Algorithm

Train

1. Translate Source language(s) into Target using lexicon
2. (Optional: extract wikifier features from Source, project to Target)
3. Train a standard NER model on new annotated Target data

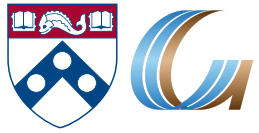
Test

1. (Optional: extract wikifier features on Target)
2. Predict with NER model as though it was trained on Gold Target data

Results

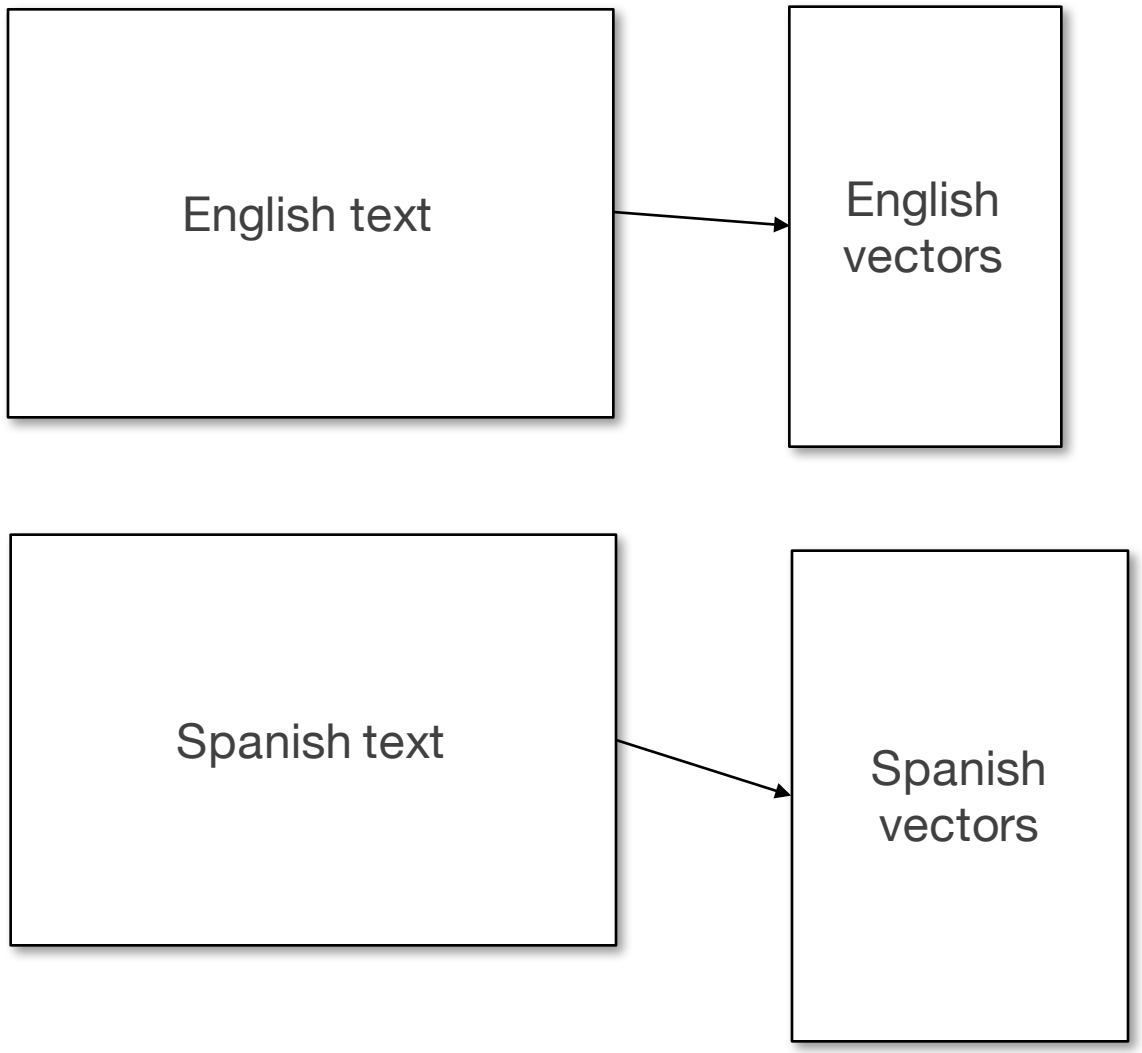
Method	Dutch	German	Spanish	Turkish	Bengali	Tamil	Yoruba	Avg
Lexicon Coverage	88.01	89.97	90.94	83.80	83.34	73.84	74.60	-
E-L Dict size	961K	1.36M	1.25M	578K	217K	182K	334K	-
Baseline	43.10	22.61	45.77	34.63	6.40	4.60	37.70	27.83
Google Translate ceiling	65.71	56.65	53.65	45.63	37.84	29.11	39.18	46.82
Wiki (Tsai et al., 2016)	61.56	48.12	60.55	47.12	43.27	29.64	36.65	46.70
Cheap Translation	53.94	50.96	51.82	46.37	30.47†	25.91†	37.58	42.43
Cheap Translation + Wiki	63.37	57.23	64.10	51.79	**46.28**†	33.10†	38.52	50.62
Best Combination	64.48	57.53	**65.95**	48.50	31.70†	27.63†	39.12	47.84
Best Combination + Wiki	**66.50**	**59.11**	65.43	**53.44**	45.70†	**34.90**†	**40.88**	**52.28**

Cross-lingual Word Vectors



- Dictionaries are cool, but also restrictive
- Surely the Embeddings Revolution extends even here?
- Key idea: words in every language projected into the same semantic space
- Thus, dist(*dog*, *perro*) = small

Cross-lingual Word Vectors



Problem!

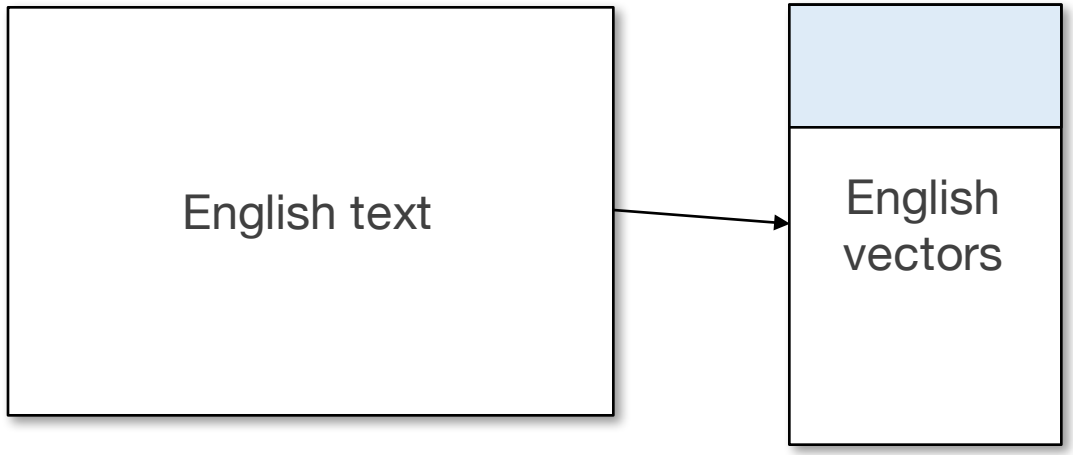
Mismatching dimensions?

Mismatching semantic space?

No progress…

Faruqui and Dyer, 2014

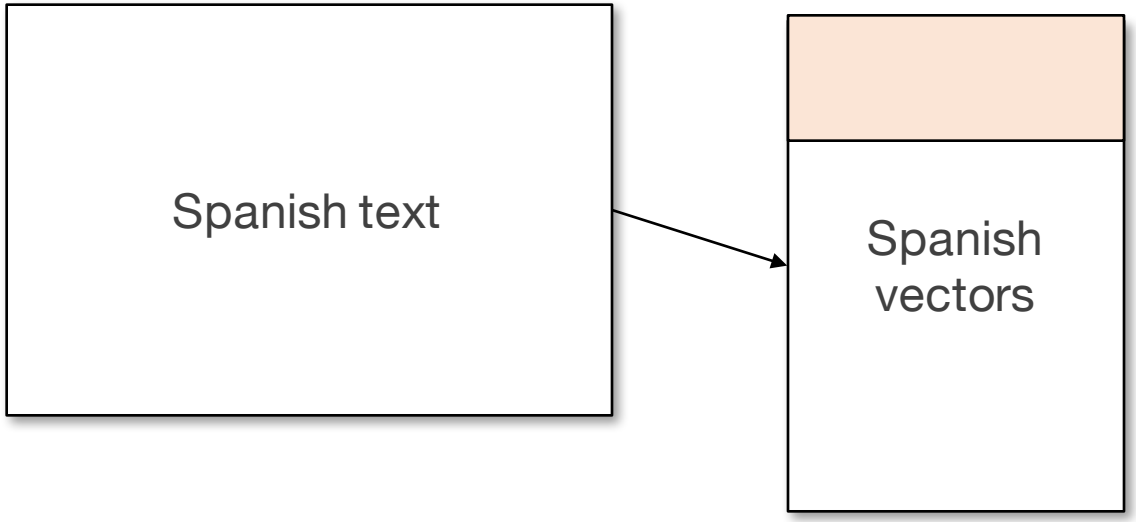
Cross-lingual Word Vectors

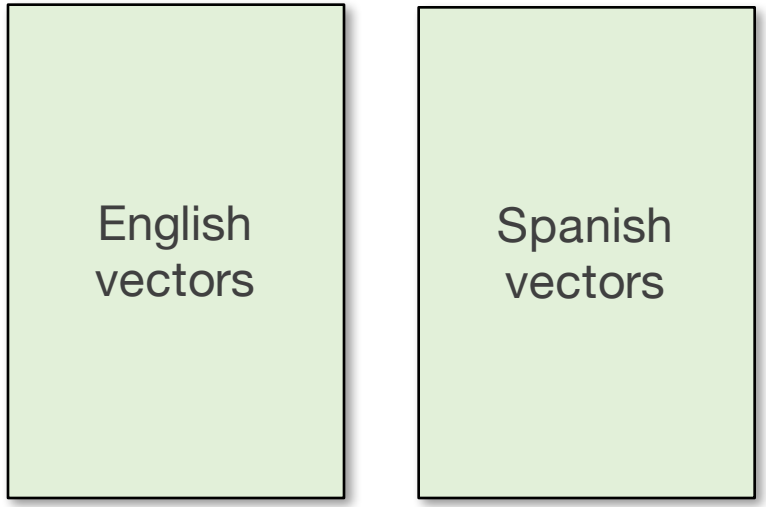


English-Spanish Lexicon

Dog 1.1 -3.9 0.3 1.0 0.23	Perro -2.1 -1 1.3 0.0 1.2
...	...

$$\boldsymbol{v}, \boldsymbol{w} = \mathrm{CCA}(\boldsymbol{x}, \boldsymbol{y})$$
$$= \arg\max_{\boldsymbol{v},\boldsymbol{w}} \rho(\boldsymbol{x}\boldsymbol{v}, \boldsymbol{y}\boldsymbol{w})$$





Faruqui and Dyer, 2014

8 Conclusions

We have unbelievable numbers. I've talked to some of the best science guys, you know, the people who know about this stuff, the best in the world, and they all say this is a really, really, tremendous work. And by the way, it's not easy work. You have to know about computers really well to do this stuff. A lot of people say to us, "you know, this work is just spectacular. We were dying for this kind of technique." And we're happy to help because that's what we like to do, and we'll keep doing it. Incredible.