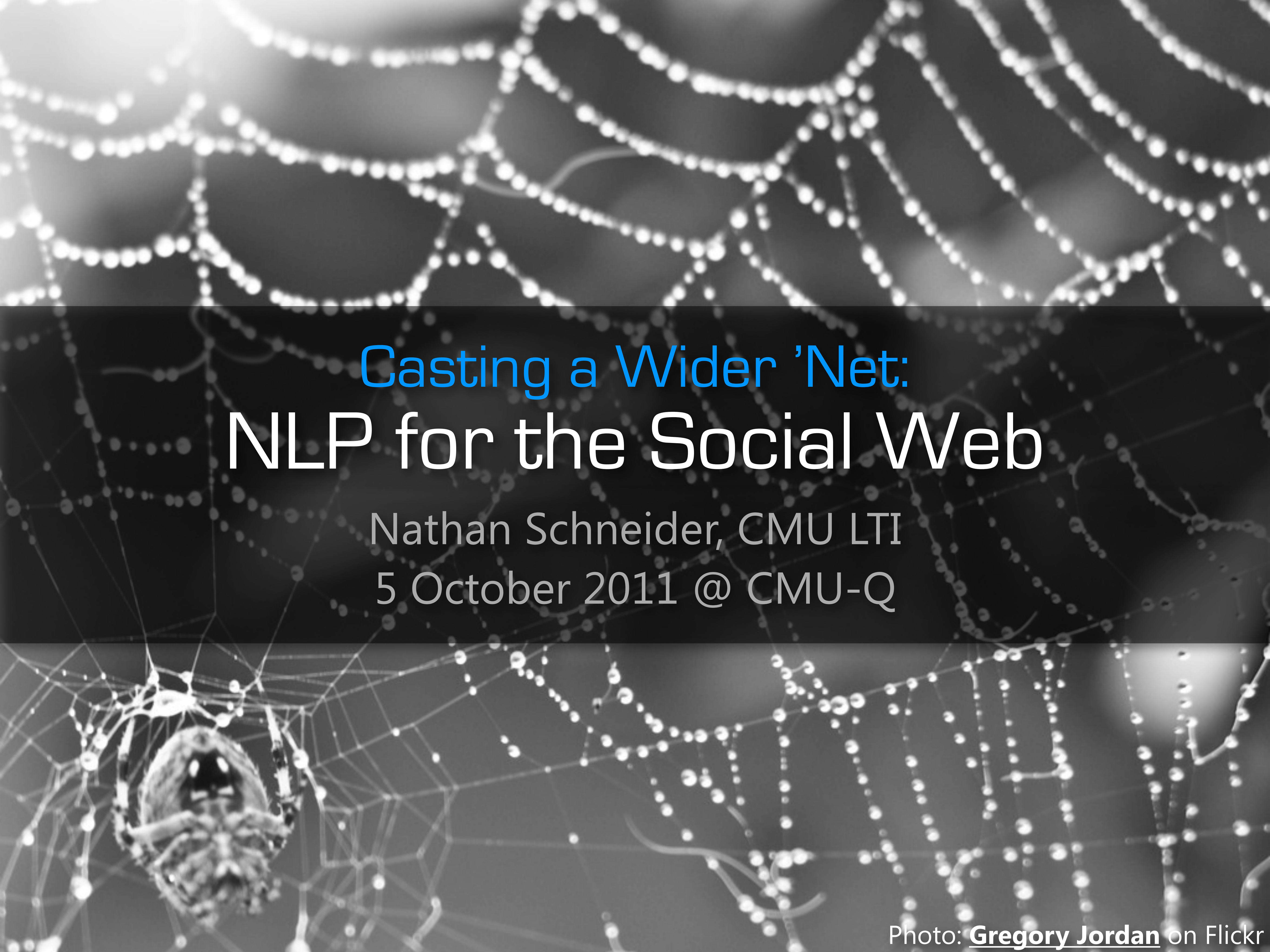


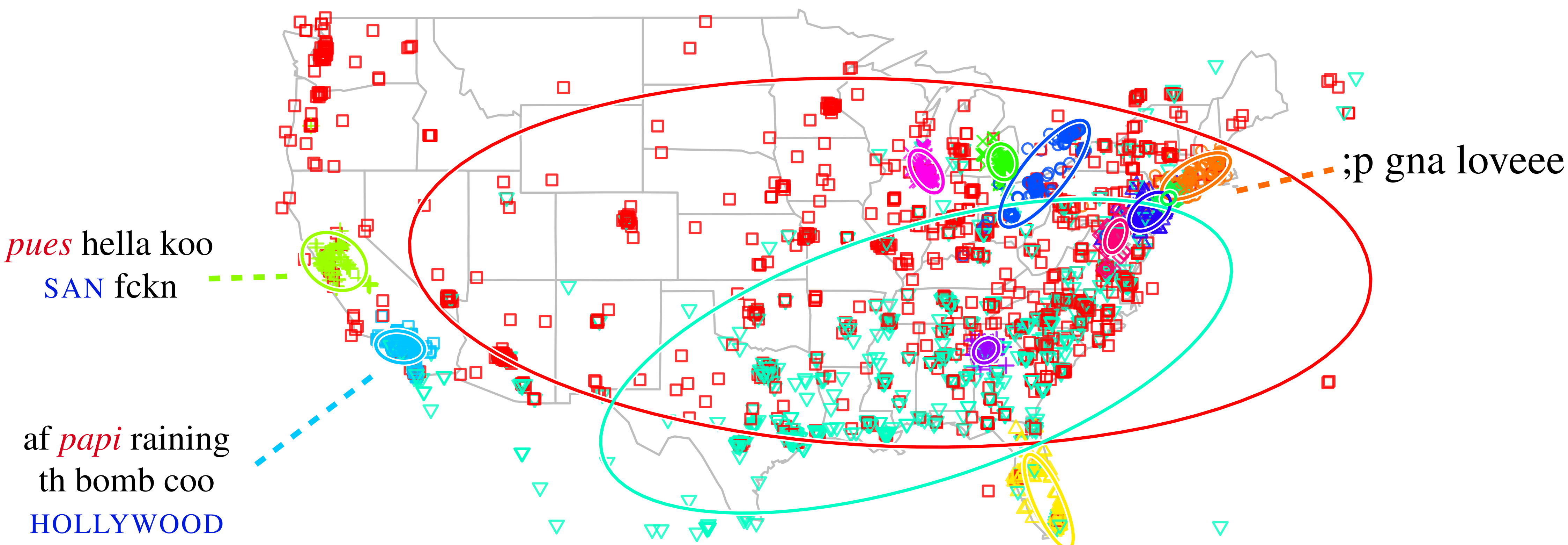
Casting a Wider 'Net:
NLP for the Social Web

Nathan Schneider, CMU LTI

5 October 2011 @ CMU-Q

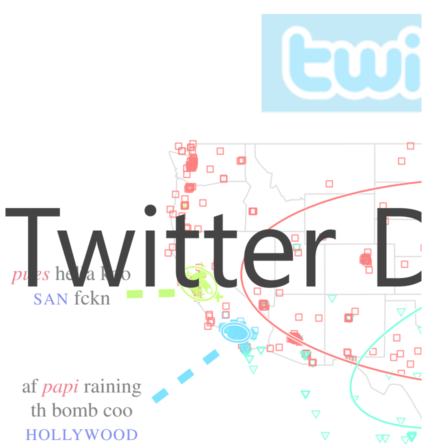
Photo: **Gregory Jordan** on Flickr

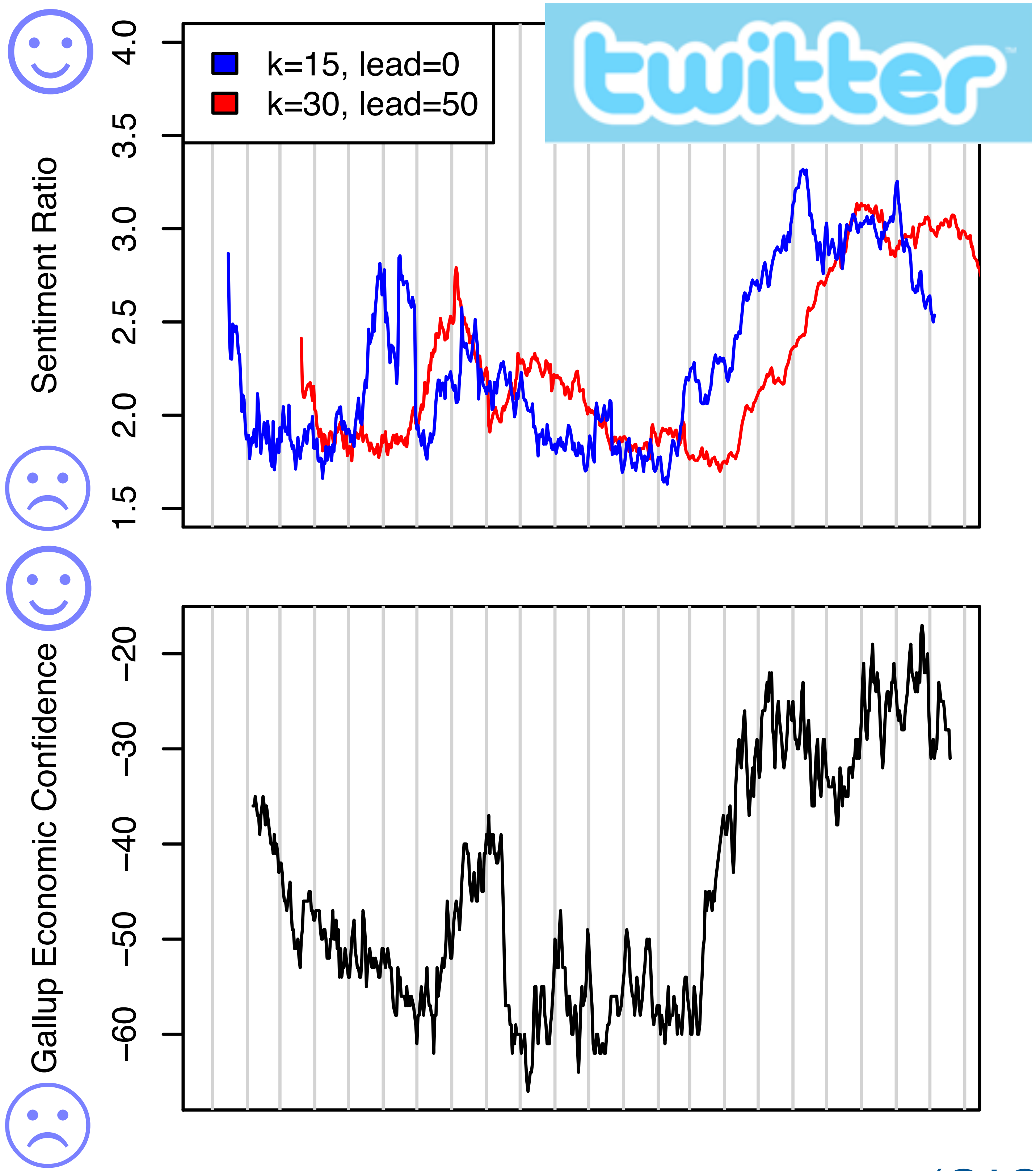
Social Media NLP



(Eisenstein et al. 2010)

Social Media NLP





(O'Connor et al. 2010)

edia NLP

DAILY KOS

▼ **she has told me personally that she is a** (2+ / 0-)

BDINO = Blue Dog In Name Only

So, yeah, i did know that. And she comes from a rich family so let her finance her own goddam elections from now on. But i hope she loses. Rumor has it Arizona's "Independent" Re-Districting Commission (but in reality, is a GOP shill group) is planning on tossing her and Grijalva into the same CD for next time. Well, hmmm, since i helped them BOTH in 2010, and now this backstabbing from Giffords, who do you think **i** will work for next time??? Hint: i **LOVE** Raul and he is a friend as well and is sooo strong on my signature issues of gay rights.

Not a single issue voter, but if I was, gay rights would be it. I just want Democrats to be tough. And I wish Obama were tougher. That's all. I'm a proud gay!

by **BoyBlue** *on* **Thu Jan 06, 2011 at 11:26:47 AM PST**
[**Parent**]

▼ **She Lied To You** (3+ / 0-)

ProgressivePunch rates Giffords almost dead last among Democrats when it comes to voting on the right side in the areas of Aid to the Less Advantaged, Fair Taxation, and Making the Government Work for Everybody, Not Just the Rich and Powerful.

http://progressivepunch.org/...

Too Folk For You

by **TooFolkGR** *on* **Thu Jan 06, 2011 at 11:29:19 AM PST**
[**Parent**]

▼ **okay, so she's a liar and i am an idiot.** (1+ / 0-)

she's **STILL** dead to me.

Not a single issue voter, but if I was, gay rights would be it. I just want Democrats to be tough. And I wish Obama were tougher. That's all. I'm a proud gay!

by **BoyBlue** *on* **Thu Jan 06, 2011 at 11:31:19 AM PST**
[**Parent**]

YGLESIAS

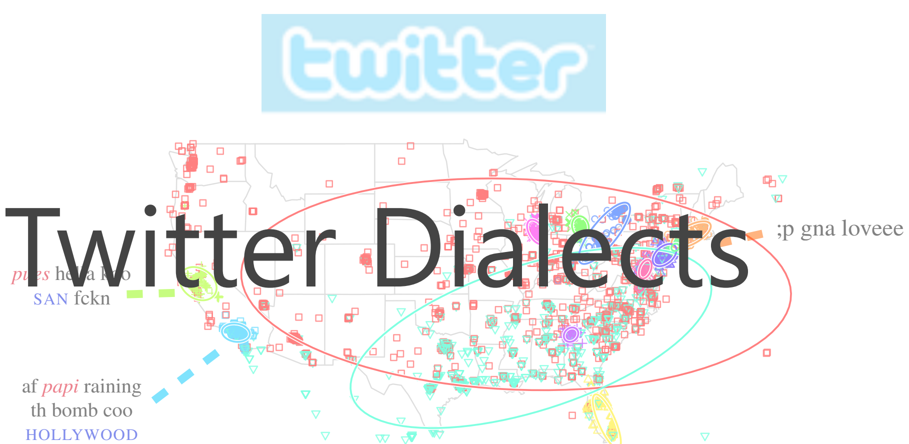
THE CARPETBAGGER REPORT Commentary and Analysis on Politics in America by Steve Benen

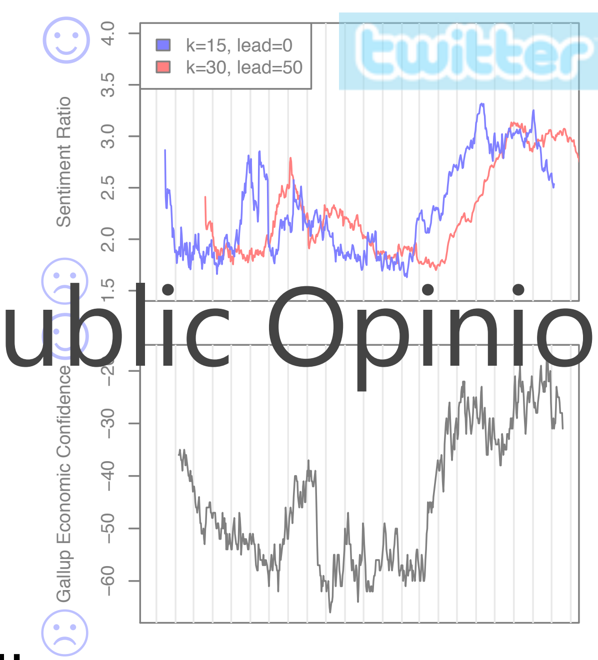
RIGHTWINGNEWS

RS REDSTATE

(Yano et al. 2009)

Social Media NLP





- Extracting news storylines (Shahaf & Guestrin 2010; Ahmed et al. 2011)
- Twitter sentiment (Barbosa & Feng 2010; Thelwall et al. 2011)
- Personalized recommendation of blog posts (El-Amini 2009)
- Predicting movie grosses from reviews (Joshi et al. 2010)

Linguistic Structure NLP

- Much of NLP is concerned with identifying aspects of **linguistic structure** in text, e.g.:

United Illuminating is based in New Haven , Conn. , and

Northeast is based in Hartford , Conn.

Linguistic Structure NLP

- Much of NLP is concerned with identifying aspects of **linguistic structure** in text, e.g.:
  - Part-of-speech tagging (/morphological analysis)

Noun	Noun	verb$_{pres}$	verb$_{pastpart}$	prep	Noun	Noun	,	Noun	,	conj
United	Illuminating	is	based	in	New	Haven	,	Conn.	,	and

Noun	verb$_{pres}$	verb$_{pastpart}$	prep	Noun	,	Noun
Northeast	is	based	in	Hartford	,	Conn.

Linguistic Structure NLP

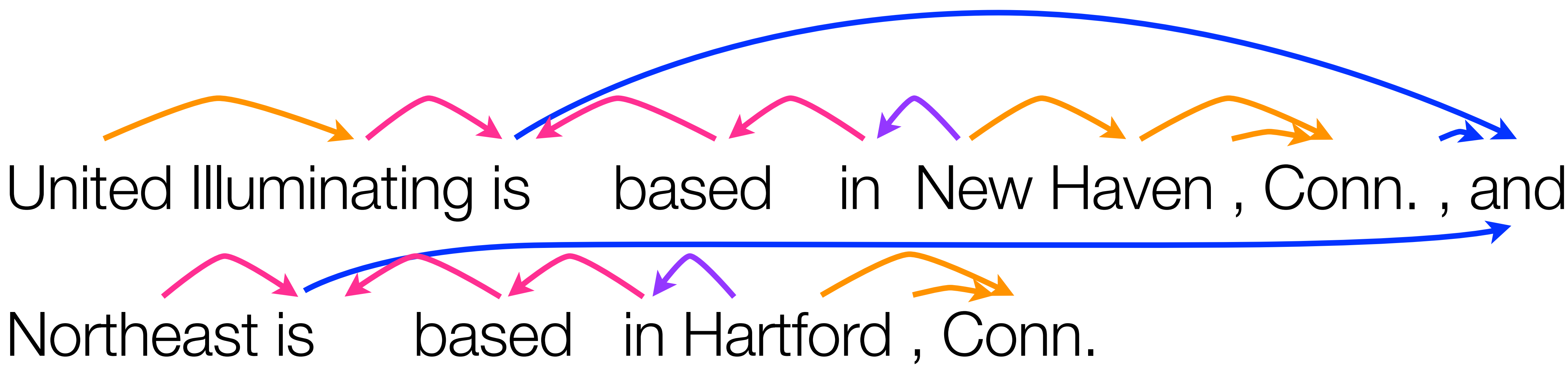
- Much of NLP is concerned with identifying aspects of **linguistic structure** in text, e.g.:
  - Part-of-speech tagging (/morphological analysis)
  - Named entity recognition

ORG________________ LOC__________________
United Illuminating is based in New Haven , Conn. , and

ORG_______ LOC______________
Northeast is based in Hartford , Conn.

Linguistic Structure NLP

- Much of NLP is concerned with identifying aspects of **linguistic structure** in text, e.g.:
  - Part-of-speech tagging (/morphological analysis)
  - Named entity recognition
  - Syntactic parsing



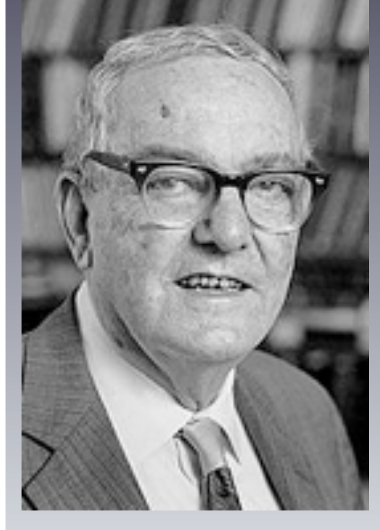
Social media language ≠ newspaper language

Applications of NLP

- Information extraction
  - *List songs people are talking about along with the album, artist(s), genre, sales, lyrics, etc.*
- Sentiment analysis
  - *Which songs do people like best?*
- Personalization/recommendation
  - *Which songs should I buy (given my past preferences and my friends' preferences)?*
- Machine translation
  - *Translate people's reviews into another language*

Noun adv noun+pos adj noun punc verb Noun URL

#**Qatar** now world's richest nation , says IMF bit.ly/pDLGVQ



… وهربرت سيمون الذي اسس …

General approach

- Supervised machine learning of a discriminative sequence model
    - **data-driven:** general-purpose algorithms for processing input examples and making statistical generalizations
    - **supervised:** (i) a *learning algorithm* uses labeled *training examples* produces a *model*; (ii) a *decoding algorithm* then uses the model to predict labels for new data at test time
    - **sequence model:** since context matters in language, we allow reasoning about neighboring decisions to influence each other

Noun adv noun+pos adj noun punc verb Noun URL

#**Qatar** now world's richest nation , says IMF bit.ly/pDLGVQ

Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments

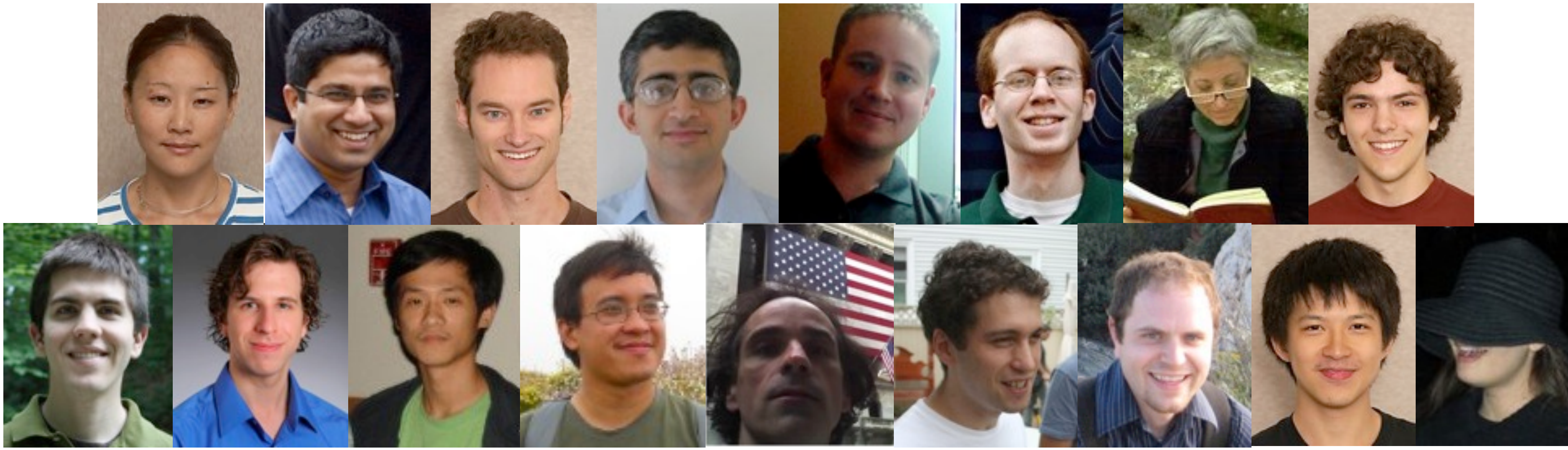
Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith

ACL 2011

Our goal:

Build a Twitter part-of-speech tagger in one day

- 17 researchers from Carnegie Mellon



non-standard spellings
(cf. Han & Baldwin 2011)

multi-word abbreviations

hashtags

Also: at-mentions, URLs, emoticons, symbols, typos, etc.

- Coarse treebank tags:

common noun	determiner
proper noun	preposition
pronoun	verb particle
verb	coordinating conjunction
adjective	numeral
adverb	interjection
punctuation	predeterminer / existential *there*

- Twitter-specific tags:

  hashtag

  at-mention

  URL / email address

  emoticon

  Twitter discourse marker

  other (multi-word abbreviations, symbols, garbage)

Hashtags

Twitter hashtags are sometimes used as ordinary words (35% of the time) and other times as topic markers

proper noun

```
Innovative , but traditional , too ! Another
fun one to watch on the #iPad ! http://bit.ly/
@user1 #utcd2 #utpol #tcot
```

hashtag

We only use “hashtag” for topic markers

Twitter Discourse Marker

Retweet construction:

RT @user1 : I never bought candy bars from those kids on my doorstep so I guess they're all in gangs now .

Twitter discourse marker

RT @user2 : LMBO ! This man filed an EMERGENCY Motion for Continuance on account of the Rangers game tonight ! ⟨⟨ Wow lmao

- Resulting tag set: 25 tags

- 17 researchers from Carnegie Mellon
- Each spent 2–20 hours annotating
- Annotators corrected output of Stanford tagger
- Two annotators corrected and standardized annotations from the original 17 annotators
- A third annotator tagged a sample of the tweets from scratch
  - Inter-annotator agreement: 92.2%
  - Cohen's κ: 0.914
- One annotator made a single final pass through the data, correcting errors and improving consistency

Experimental Setup

- 1,827 annotated tweets
  - 1,000 for training
  - 327 for development
  - 500 for testing (OOV rate: 30%)
- Systems:
  - Stanford tagger (retrained on our data)
  - Our own baseline CRF tagger
  - Our tagger augmented with Twitter-specific features

Phonetic Normalization Features

- One of several new feature types that proved helpful
- Metaphone algorithm (Philips, 1990) maps tokens to equivalence classes based on phonetics
- Examples:

tomarrow tommorow tomorr tomorrow tomorrowwww

hahaaha hahaha hahahah hahahahhaa hehehe hehehee

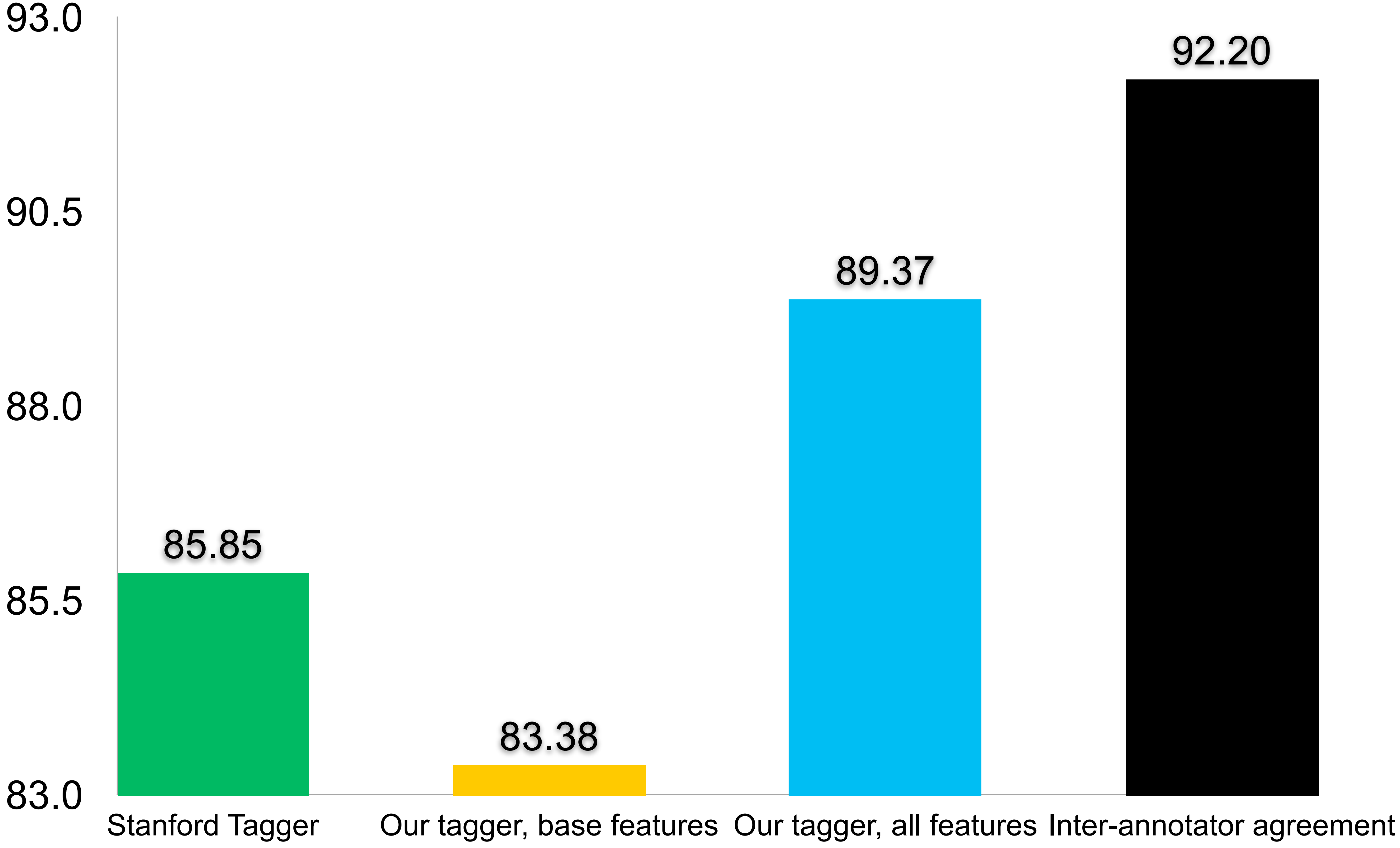
✓

thangs thanks thanksss thanx things thinks thnx

knew kno know knw n nah naw new no noo nooooooo now

✗

Results


93.0
90.5
88.0
85.5
83.0
85.85
83.38
89.37
92.20
Stanford Tagger
Our tagger, base features
Our tagger, all features
Inter-annotator agreement

Twitter POS Summary

- We developed a tag set, annotated data, designed features, and trained models
- Case study in rapidly porting a fundamental NLP task to a social media domain
- Tagger, tokenizer, and annotations are available:

  www.ark.cs.cmu.edu/TweetNLP/

Adapting NLP to social media: modeling strategies

1. Annotate and train on **appropriate data**
2. Add useful **features**
3. Modify the **learning algorithm**
4. Exploit **unlabeled data** *(semi-supervised learning)*

Recall-Oriented Learning for Named Entity Recognition in Wikipedia

Behrang Mohit

Rishav Bhowmick

Nathan Schneider

Kemal Oflazer

Noah A. Smith

ARK

lti

Carnegie Mellon

جامعة كارنيجي ميلون في قطر

Carnegie Mellon Qatar

الصندوق القطري لرعاية البحث العلمي

Qatar National Research Fund

WIKIPEDIA

English
The Free Encyclopedia
3 719 000+ articles

日本語
フリー百科事典
764 000+ 記事

Español
La enciclopedia libre
822 000+ artículos

Deutsch
Die freie Enzyklopädie
1 277 000+ Artikel

Français
L'encyclopédie libre
1 141 000+ articles

Русский
Свободная энциклопедия
759 000+ статей

Italiano
L'enciclopedia libera
834 000+ voci

Português
A enciclopédia livre
694 000+ artigos

Polski
Wolna encyklopedia
824 000+ haseł

中文
自由的百科全書
371 000+ 條目

search • suchen • rechercher • ricerca • szukaj • buscar • 検索 • поиск • zoeken • busca • sök • 搜尋 • cerca • søk • пошук • haku • tìm kiếm • hledání • keresés • ara • 찾기 • cari • căutare • بحث • جستجو • søg • serĉu • претрага • paieška • hľadať • חיפוש • cari • търсене • poišči • suk • bilatu • bilnga • traži

English →

100 000+

العربية • Български • Català • Česky • Dansk • Deutsch • English • Español • Esperanto • Euskara • فارسی • Français • 한국어 • Hrvatski • Bahasa Indonesia • Italiano • עברית • Lietuvių • Magyar • Bahasa Melayu • Nederlands • 日本語 • Norsk (bokmål) • Polski • Português • Русский • Română • Slovenčina • Slovenščina • Српски / Srpski • Suomi • Svenska • Türkçe • Українська • Tiếng Việt • Volapük • Winaray • 中文

10 000+

Afrikaans • Alemannisch • አማርኛ • Aragonés • Armãneashce • Asturianu • Kreyòl Ayisyen • Azərbaycan / آذربایجان دیلی • Беларуская (Акадэмічная • Тарашкевіца) • Bosanski • Brezhoneg • Чăваш • Cymraeg • Eesti • Ελληνικά • Frysk • Gaeilge • Galego • ગુજરાતી • Հայերեն • हिन्दी • Ido • Íslenska • Basa Jawa • ქართული • Kurdî / كوردی • Latina • Latviešu • Lëtzebuergesch • Lumbaart • Македонски • मराठी • नेपाल भाषा • नेपाली • Norsk (nynorsk) • Nnapulitano • Occitan • Piemontèis • Plattdüütsch • Қазақша • Ripoarisch • Runa Simi • شاہ مکھی پنجابی • Shqip • Sicilianu • Simple English •

http://www.wikipedia.org/

□□□□□
Català
Česky
Dansk
Deutsch
English
Español
Euskara
فارسی
Suomi
Français
עברית
Bahasa Indonesia
Italiano
ქართული
한국어
Nederlands
Norsk (nynorsk)
Norsk (bokmål)
Polski
پنجابی
Português
Русский
Simple English
Svenska
□□□□□□
ไทย
Türkçe
中文

- 1965年 メロン工業研究所を吸収合併。**カーネギーメロン大学**と改称
- 2005年 カーネギーメロン大学日本校 (Carnegie Mellon CyLab Japan) を設置

	理学部
	人文学部
	公共政策学部
	計算機科学部
	経済経営学部
研究科	工学研究科
	芸術学研究科
	理学研究科
	人文学研究科
	公共政策学研究科
	計算機科学研究科
	経済経営学研究科
ウェブサイト	カーネギーメロン大学公式サイト

表示

特色 [編集]

マサチューセッツ工科大学、カリフォルニア工科大学とともにアメリカ有数の名門工科大学の1つと評されており、計算機科学、公共政策学、経営学、音楽・映像分野を幅広くカバーしており、ブロードウェイにおいても有名な存在である。USNewsの2010年版大学ランキングでは総合23位の評価を得ている。

計算機科学 (computer science) を筆頭に、ロボット工学 (robotics)、機械工学 (engineering)、理学 (the sciences)、ビジネス (business)、公共政策 (public policy)、美術 (fine arts) および人文学 (the humanities) などのスクールおよびカレッジを設置する。

特筆すべきは計算機科学で、全米で1位の評価を得ている（USNewsの2010年版大学院ランキング・計算機科学部門）[1]。コンピュータセキュリティ発信の中枢であるコンピューター緊急事態対策チーム (CERT) の統轄本部CERT Coordination Centerの運営も行っている[2]。Javaの生みの親であるジェームズ・ゴスリング、やLycosの創始者マイケル・モールディンの出身校として、またマイクロカーネルの代名詞でもあるMachを開発した大学という事などでも有名。

工学分野でも高い評価を得ている（2010年版USNews全米6位）[3]。なお、1992年〜2001年にかけて、日本人の金出武雄教授（現在・U.A.and Helen Whitaker記念全学教授）がロボット研究所の所長を務めていた。

公共政策大学院（ハインツ・カレッジ）は、情報セキュリティ分野（2007年版USNews全米1位）、公共政策管理分野（2007年版USNews全米4位）を中心に高い評価を得ており、多くの人材を国際機関、中央政府へと輩出している。

テッパー・スクール・オブ・ビジネス（旧Graduate School of Industrial Administration）も全米有数のビジネススクールとして高い評価を得ている（2010年版USNews全米16位）[4]。

ノーベル賞受賞学者は、ハーバート・サイモン教授、エドワード・プレスコット教授を始め、現・前教授および卒業生より13名輩出している。関係者の受賞したその他の著名な賞と人数は、チューリング賞が9名、エミー賞が7名、アカデミー賞が3名、トニー賞が4名である。

卒業生にモダンアートのアンディー・ウォーホルやアカデミー賞受賞俳優のホリー・ハンター、新しくはTVシリーズ「ER緊急救命室」のミン・ナ、そしてTVシリーズ「HEROES」のサイラーでブレイクし、2009年度版「STAR TREK」Mr.スポックに大抜擢されたザカリー・クイントなどがいる。

2005年にアジアの情報セキュリティ教育研究拠点を目指し、兵庫県と共同で同県神戸市にカーネギーメロン大学日本校 (Carnegie Mellon CyLab Japan) を設置した。2007年に東京工科大学と片柳コンピュータ科学賞を設立[1]。同年大阪府大阪市に「エンターテイメントテクノロジーセンター」の設置が決定し、2008年から同センターが稼働している[2]。

キャンパス [編集]

メインキャンパスは103エーカー (0.4 km²) あり、ピッツバーグ中心地より約3マイル（5 km）離れた近郊に位置する。西側はピッツバーグ大学と隣接している。

カーネギーメロン大学のパノラマ

http://ja.wikipedia.org/wiki/カーネギーメロン大学

ברוכים הבאים לוויקיפדיה!

ויקיפדיה היא מיזם רב לשוני לחיבור אנציקלופדיה שיתופית, חופשית ומהימנה, שכולם יכולים לערוך. כעת יש בוויקיפדיה העברית 123,029 ערכים.

מתמטיקה	פיזיקה	כימיה
מדעי החלל	טכנולוגיה	ביולוגיה
רפואה	גאוגרפיה	מדעי החברה
כלכלה	מדע המדינה	היסטוריה
דת	יהדות	ישראל
מוזיקה	ספרות	אישים
ספורט	אמנות	מדינות העולם

http://he.wikipedia.org/wiki/עמוד_ראשי

Named Entity Recognition

In the 20th century, the study of mathematical logic provided the essential breakthrough that made artificial intelligence seem plausible. The foundations had been set by such works as Boole's *The Laws of Thought* and Frege's *Begriffsschrift*. Building on Frege's system, Russell and Whitehead presented a formal treatment of the foundations of mathematics in their masterpiece, the *Principia Mathematica* in 1913. Inspired by Russell's success, David Hilbert challenged mathematicians of the 1920s and 30s to answer this fundamental question: "can all of mathematical reasoning be formalized?"[15] His question was answered by Gödel's incompleteness proof, Turing's machine and Church's Lambda calculus.[15][22] Their answer was surprising in two ways. First, they proved that there were, in fact, limits to what mathematical logic could accomplish.

http://en.wikipedia.org/wiki/History_of_artificial_intelligence

Named Entity Recognition

Muammar Gaddafi tunnetaan eräistä erikoisuuksistaan. Hän asuu ja ottaa vastaan vieraansa beduiiniteltassa. Vierailevat valtiovieraat joutuvat kiipeämään Yhdysvaltain pommitusten jättämien hänen entisen palatsinsa raunioiden yli, jotka on jätetty mielenosoituksellisesti raivaamatta.[7] Gaddafi asuu teltassa myös ulkomailla vieraillessaan, jolloin hänen telttansa pystytetään yleensä isännän presidentinpalatsin tms. läheisyyteen, esim. Pariisissa Hôtel Marignyn pihamaalle[8], Moskovassa Kremliin ja Roomassa Pamphilin puistoon[9]. Hänellä on myös pelkästään naisista koostuva henkivartiokaartti[10][11].

http://fi.wikipedia.org/wiki/Gaddafi

Named Entity Recognition

اسس المجال الحديث لبحوث الذكاء الاصطناعي في مؤتمر في حرم كليه دارتموث في صيف عام 1956.[11] أصبح هؤلاء الحضور قادة بحوث الذكاء الاصطناعي لعدة عقود، وخاصة جون مكارثي ومارفن مينسكاي، ألين نويل وهربرت سيمون الذي اسس مختبرات للذكاء الاصطناعي في معهد ماساتشوستس للتكنولوجيا (MIT) وجامعة كارنيجي ميلون (CMU) وستانفورد.هم وتلاميذهم كتبوا برامج أدهشت معظم الناس.[47] كان الحاسب الآلي يحل مسائل في الجبر ويثبت النظريات المنطقيه ويتحدث الإنجليزيه.[12] بحلول منتصف الستينات أصبحت تلك البحوث تمول بسخاء من وزاره الدفاع الأمريكيه.[56] و هؤلاء الباحثون قاموا بالتوقعات الآتية:

http://ar.wikipedia.org/wiki/ذكاء_اصطناعي [Artificial Intelligence]

Named Entity Recognition

In the 20th century, the study of mathematical logic provided the essential breakthrough that made artificial intelligence seem plausible. The foundations had been set by such works as Boole's *The Laws of Thought* and Frege's *Begriffsschrift*. Building on Frege's system, Russell and Whitehead presented a formal treatment of the foundations of mathematics in their masterpiece, the *Principia Mathematica* in 1913. Inspired by Russell's success, David Hilbert challenged mathematicians of the 1920s and 30s to answer this fundamental question: "can all of mathematical reasoning be formalized?"[15] His question was answered by Gödel's incompleteness proof, Turing's machine and Church's Lambda calculus.[15][22] Their answer was surprising in two ways. First, they proved that there were, in fact, limits to what mathematical logic could accomplish.

http://en.wikipedia.org/wiki/History_of_artificial_intelligence

Beyond traditional NE categories

- NER work has traditionally focused on the **news** domain and a small number of categories, namely **PERSON**, **ORGANIZATION**, **LOCATION** (POL)
    - these are important, but not usually sufficient to cover important names for other domains
    - one solution: Develop a fine-grained taxonomy— domain-specific (Settles, 2004; Yao et al., 2003) or general-purpose (Sekine et al., 2002; Weischedel & Brunstein, 2005; Grouin et al., 2011). Doesn't scale well to many domains, non-expert annotators.
    - **our approach:** Annotators invent new categories on an article-specific basis. Simple yet flexible.

Arabic Wikipedia Data

- Downloaded a full snapshot of ar.wikipedia.org (>100K articles)
- **Dev+test data:** 28 articles manually selected and grouped into 4 domains for annotation
  - history, science, sports, technology
  - >1,000 words; cross-linked to an English, German, and Chinese article; subjectively deemed high-quality

Annotation

- 2 CMU-Q undergraduates (native Arabic speakers) marked entities in:
    - the 3 canonical NE classes: **PERSON**, **ORGANIZATION**, **LOCATION** (POL)
    - up to 3 salient categories specific to the article
    - a generic **MISCELLANEOUS** category
- Proportion of non-POL entities varies widely by domain: 6% for history, 83% for technology
- High inter-annotator agreement on a held-out article (see TR for details)
- Will be publicly released

Annotation

article titles (in English)

	History	Science	Sports	Technology
dev	Damascus	Atom	Raúl Gonzáles	Linux
	Imam Hussein Shrine	Nuclear power	Real Madrid	Solaris
test	Crusades	Enrico Fermi	2004 Summer Olympics	Computer
	Islamic Golden Age	Light	Christiano Ronaldo	Computer Software
	Islamic History	Periodic Table	Football	Internet
	Ibn Tolun Mosque	Physics	Portugal football team	Richard Stallman
	Ummaya Mosque	Muhammad al-Razi	FIFA World Cup	X Window System

Claudio Filippone (PER) كلوديو فيلبون

Spanish League (CHAMPIONSHIPS) الدوري الاسباني

nuclear radiation (GENERIC-MISC) الاشعاع النووي

Linux (SOFTWARE) لينكس

proton (PARTICLE) بروتون

Real Zaragoza (ORG) ريال سرقسطة

example NEs in conventional & article-specific categories

From annotation to modeling

- Next, we report on experiments on **detecting entity mentions** (boundaries) in this data
    - We show that standard supervised learning is plagued by low out-of-domain recall
    - Two techniques are proposed to mitigate the domain gap: a recall-oriented learning bias and semi-supervised learning

Supervised learning

labeled training data

ACE, ANER:

200K words, 16K entities

test data

Arabic Wikipedia:

50K words, 4K entities

20 articles: history, science, sports, technology

Model

- **Structured perceptron** with features based on prior work in Arabic NER (Benajiba et al., 2008; Abdul-Hamid & Darwish, 2010)
  - Local context (neighboring words)
  - Shallow morphology: character n-grams
  - Morphology: normalized spelling, POS, aspect/case/gender/number/person/definiteness from MADA tool (Habash & Rambow, 2005; Roth et al., 2008)
  - Presence of diacritics
  - Projected English capitalization (using a bilingual lexicon induced heuristically from article titles)

Decoding

features

tag	word=whrbrt	length=6	char0=w	prev=stmwn	pos=noun	...	**total**
B	1.53	-8.54	12.90	-0.24	-0.05	...	10.88
I	-4.15	-25.09	-4.89	1.67	0.66	...	16.42
O	-9.00	45.12	11.12	-12.01	19.45	...	-3.50

...	وهربرت	سيمون	الذي	اسس	...
	whrbrt	symwn	Alzy	Ass	

Decoding

I

B

O

…	وهربرت	سيمون	الذي	اسس	…
	whrbrt	symwn	Alzy	Ass	

Decoding

→ **I**

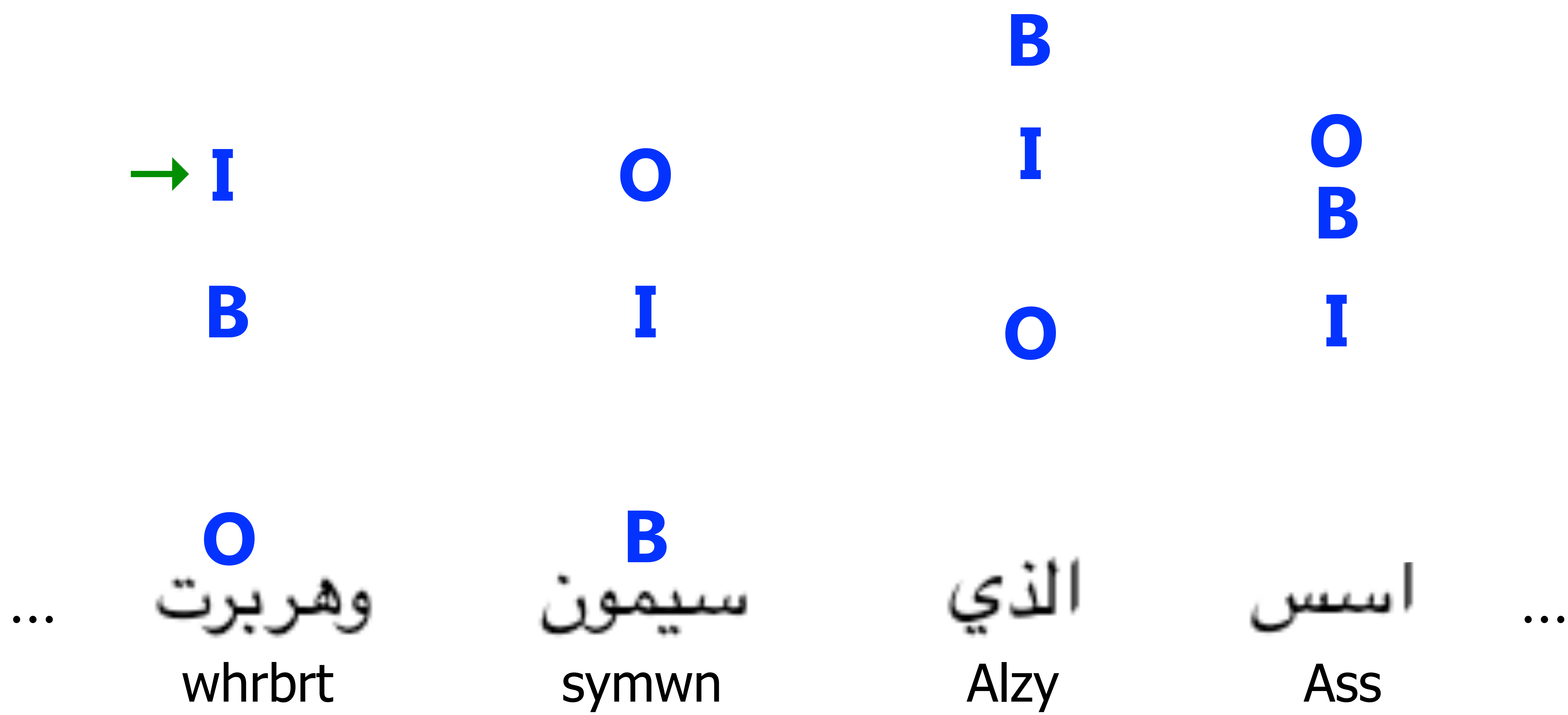
B

O

… وهربرت سيمون الذي اسس …

whrbrt symwn Alzy Ass

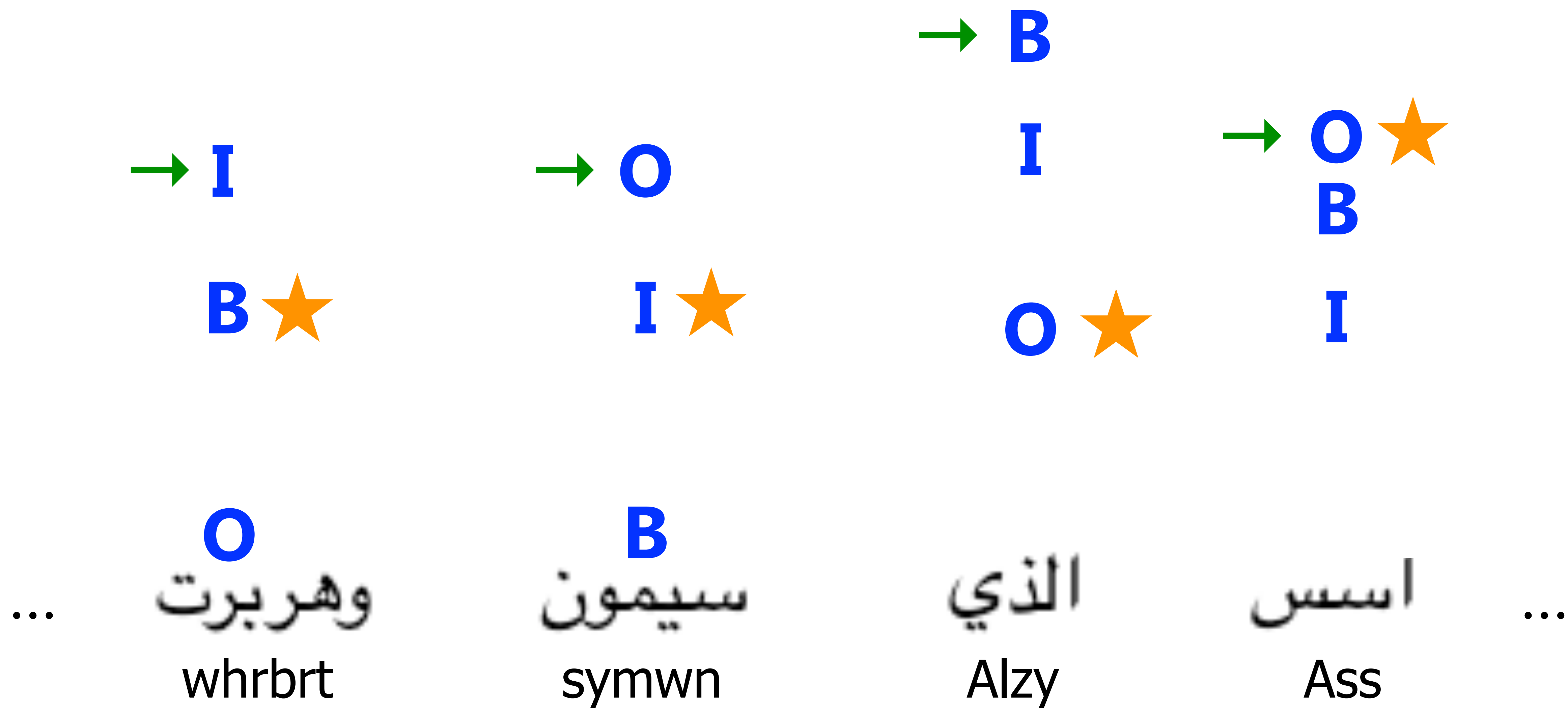
Decoding



Decoding

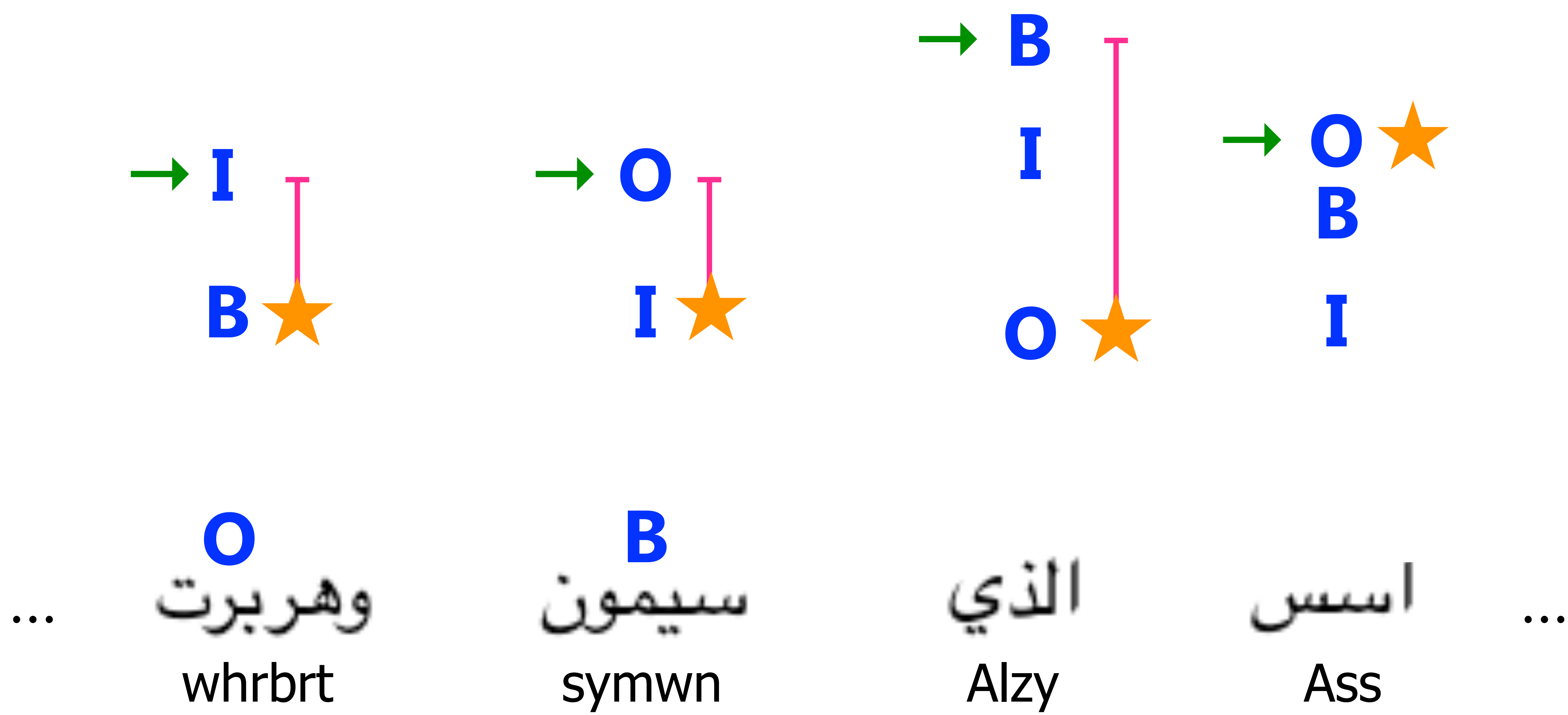
	→ I	→ O	→ B	→ O	
	B	I	I	B	
	O	B	O	I	
…	وهربرت	سيمون	الذي	اسس	…
	whrbrt	symwn	Alzy	Ass	

Learning



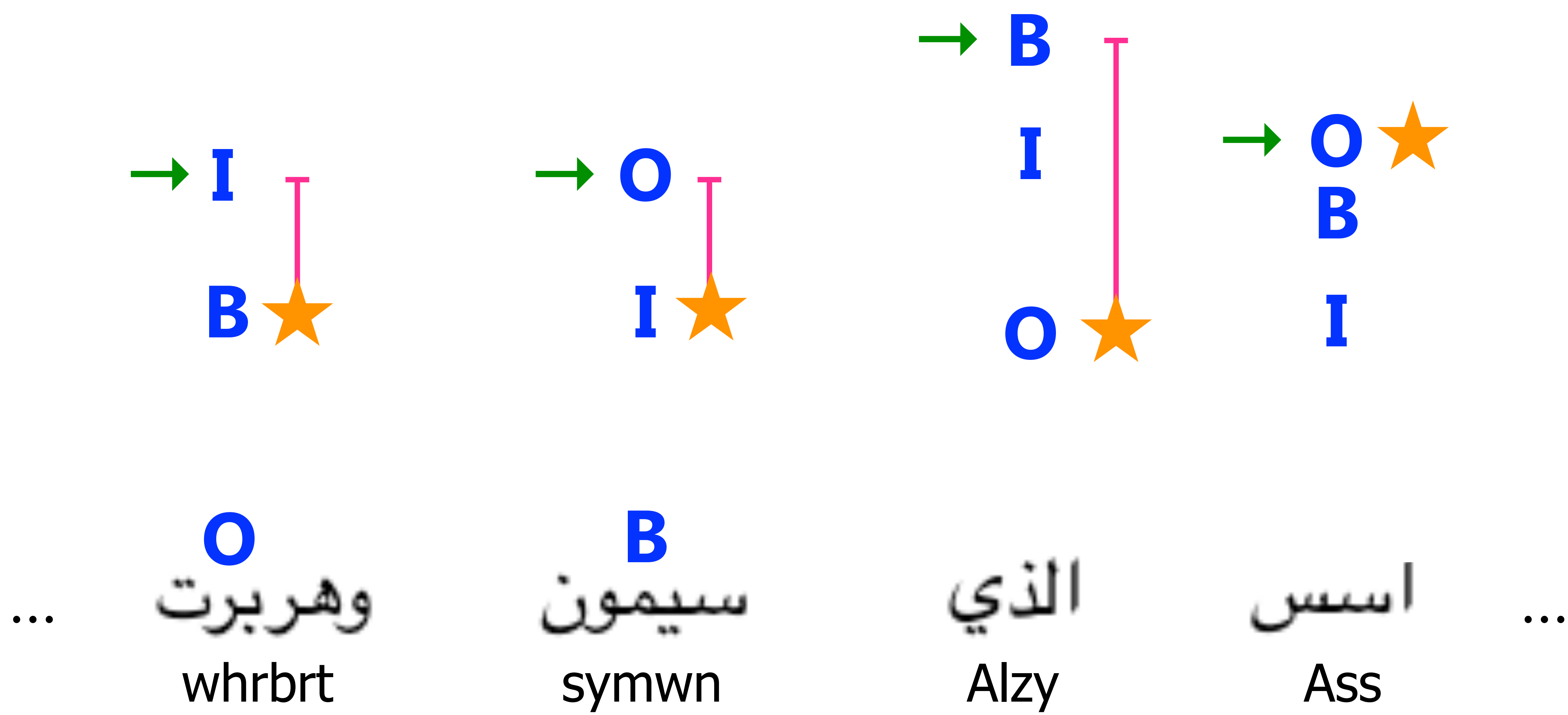
Learning

objective: update weights so as to minimize the loss (summed over all training data points)



Learning

objective: update weights so as to minimize the loss (summed over all training data points)



First-order model allows us to encode features over tag bigrams. **O I** sequence is forbidden.

Supervised learning results

TRAIN

Arabic news

TEST

same domain

cross-domain

	P	R	F
fold 1	70.43	63.08	66.55
fold 2	87.48	81.13	84.18
fold 3	65.09	51.13	57.27
average	74.33	65.11	69.33

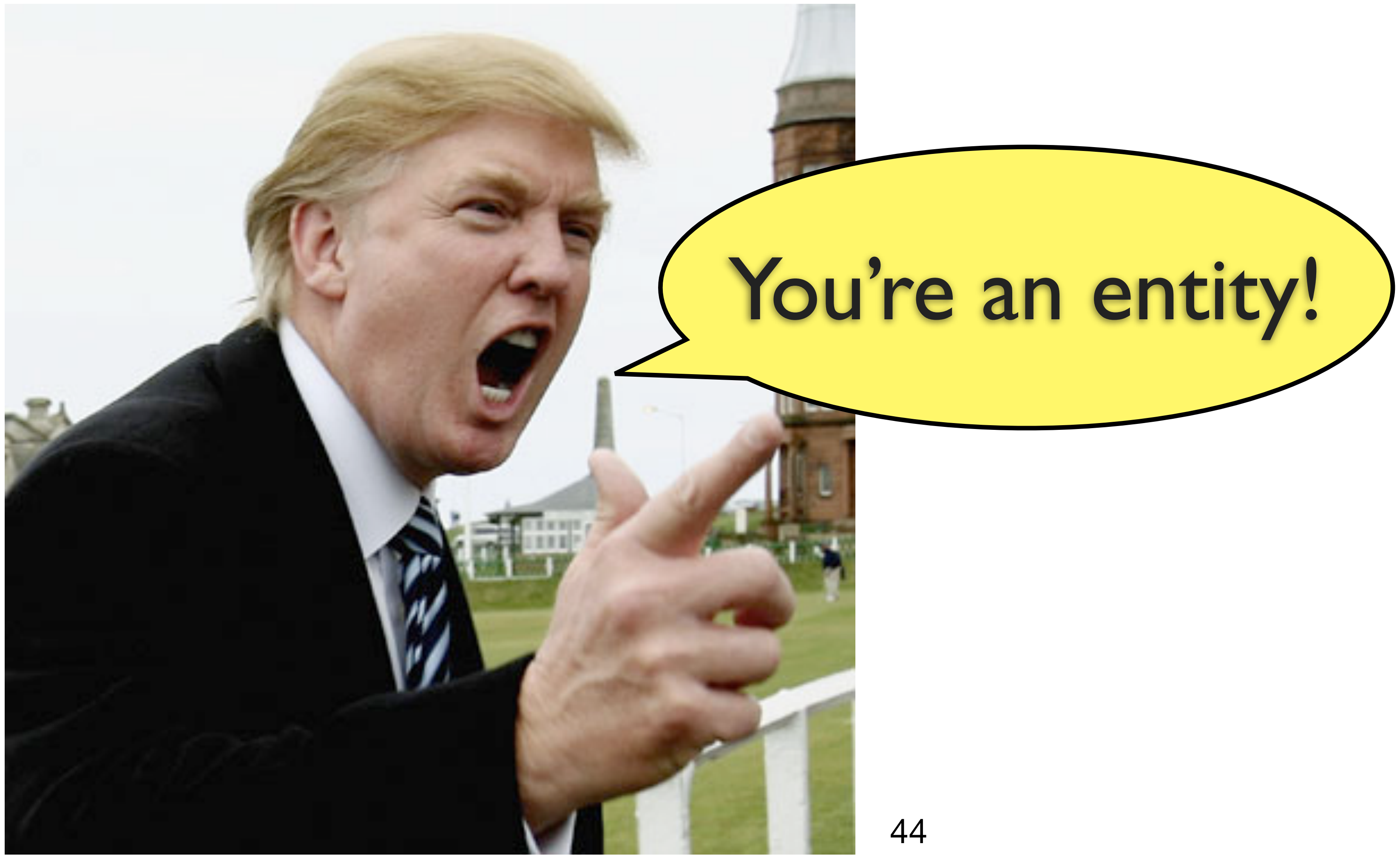
	P	R	F
technology	60.42	20.26	30.35
science	64.96	25.73	36.86
history	63.09	35.58	45.50
sports	71.66	59.94	65.28
overall	66.30	35.91	46.59

on par with state of the art
(Abdul-Hamid & Darwish, 2010)

☹

Recall-oriented learning

- Problem: The model is too hesitant to propose new entities in the new domain.
- Idea: Bias the model so it learns to be arrogant about proposing entities.



Precision-recall tradeoff

- The precision-recall tradeoff sometimes matters for applications (e.g., whether output will be filtered by a user).
  - Known techniques to impose such a bias in structured prediction.
- We propose that biasing the learner with one of these techniques is appropriate for domain adaptation.

Recall-oriented learning results

supervised	P	R	F
regular	66.3	35.9	46.6
tweaking: oracle	*66.2*	*39.0*	*49.1*

- "Tweaking" the model after supervised learning—namely, tuning the weight of the "**O**" feature, effectively thresholding on confidence (Minkov et al., 2006)
  - ~3 point improvement if we cheat and use the test data to choose the best weight

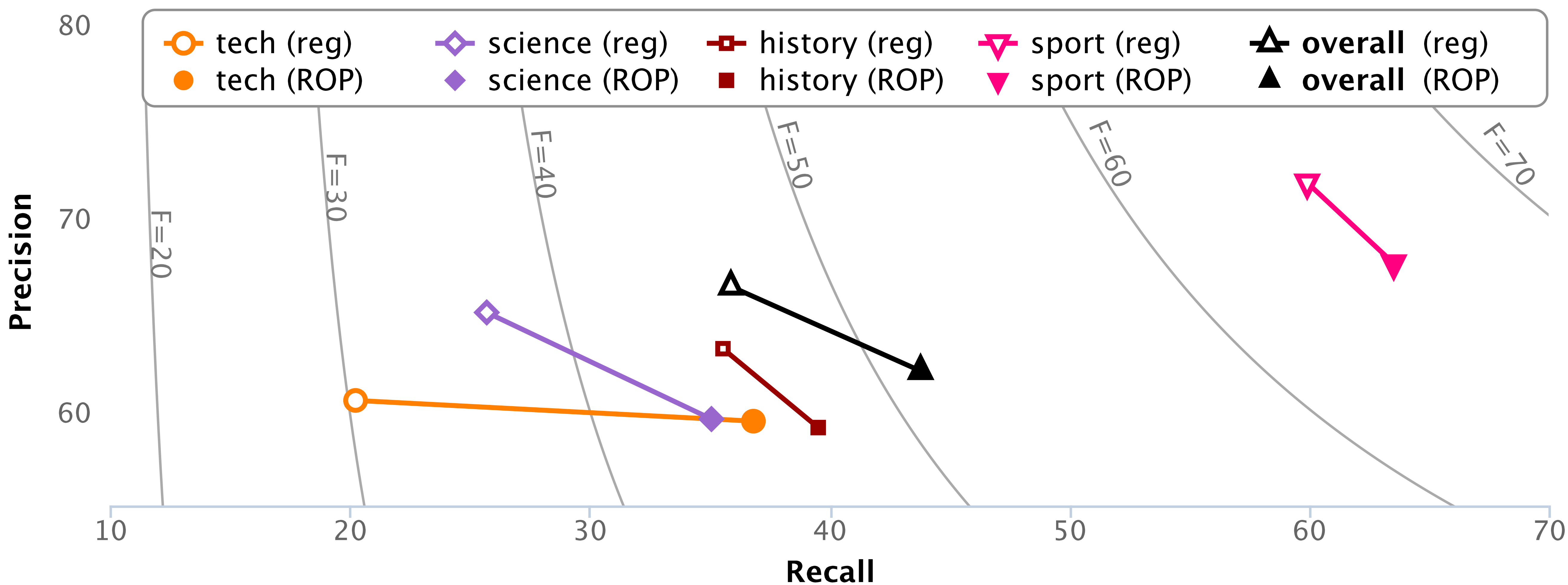
Recall-oriented learning results

supervised	P	R	F
regular	66.3	35.9	46.6
tweaking: oracle	*66.2*	*39.0*	*49.1*
cost function	61.9	43.8	51.33

- Cost-augmented decoding (Crammer et al., 2006; Gimpel & Smith, 2010), which (unlike tweaking) affects *all* features *during* learning

Recall-oriented learning results

supervised	P	R	F
regular	66.3	35.9	46.6
tweaking: oracle	*66.2*	*39.0*	*49.1*
cost function	61.9	43.8	51.33



Semi-supervised learning

labeled training data

test data

unlabeled data, same domain as test

Self-training

- Simple procedure:
  1. **supervised learning** on training data
  2. use learned model to **predict** labels for large amounts of target-domain data
  3. **retrain**, treating those predictions as gold-standard labels
  4. go back to step 2 and repeat *(optional)*

Gaddafi, **ruler of Libya**

Gaddafi **(1942—)**

______ (<num>-

Simon **(1916–2001)**

Self-training results

supervised	self-training	P	R	F
regular	—	66.3	35.9	46.59
recall-oriented	—	61.9	43.8	51.33
regular	**regular**	66.7	35.6	46.41
recall-oriented	**regular**	61.8	43.0	50.75

☹

Why does self-training hurt?

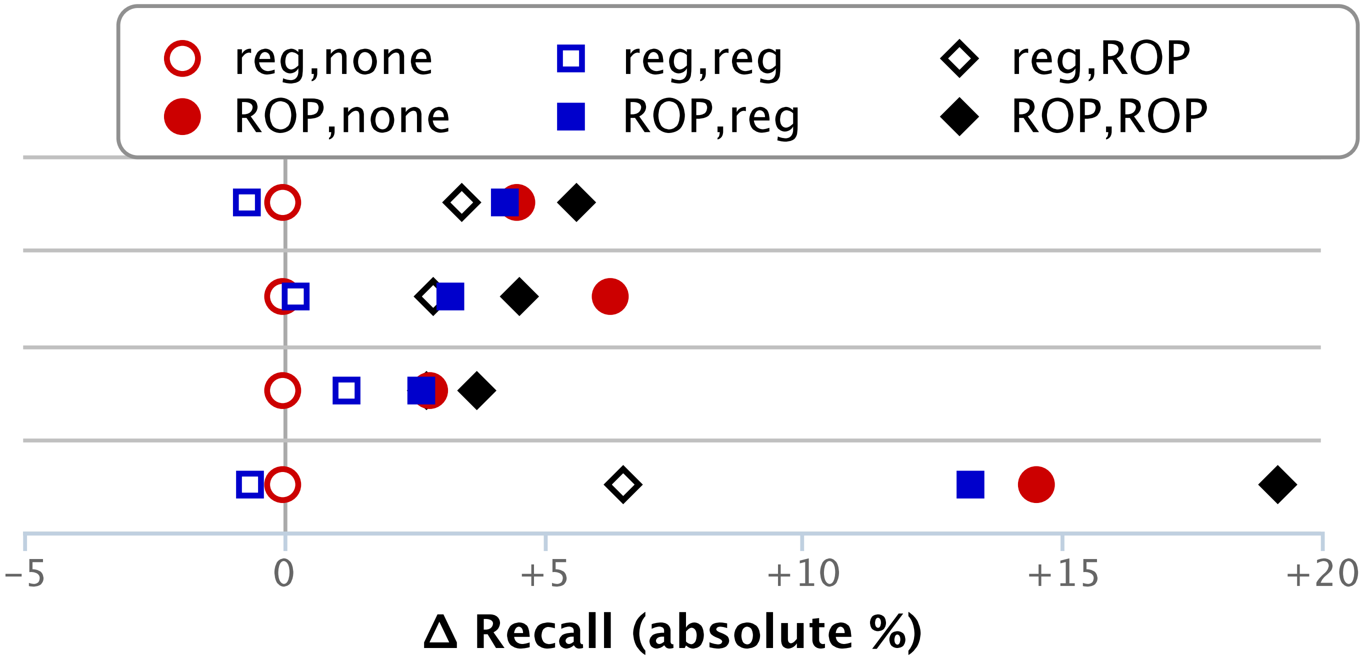
- The initial labeling phase of self-training will still miss a lot of entities, so training on those labels effectively teaches the final model to prefer “**O**”!

Self-training results

supervised	self-training	P	R	F
regular	—	66.3	35.9	46.59
recall-oriented	—	61.9	43.8	51.33
regular	regular	66.7	35.6	46.41
recall-oriented	regular	61.8	43.0	50.75
regular	**recall-oriented**	59.2	40.3	47.97
recall-oriented	**recall-oriented**	59.5	46.0	51.88

Class breakdown

- If we look at where the recall-oriented bias makes a difference in recall, it is mainly the non-POL entities (most room for improvement).



	entities	**words**	**baseline recall**
PER	1081	1743	49.95
ORG	286	637	23.92
LOC	1019	1413	61.43
MIS	1395	2176	9.30
overall	3781	5969	35.91

Wikipedia NER Conclusions

- Wikipedia poses a number of challenges for NLP, a chief one being domain diversity
- Many different types of entities are important to non-news domains, and annotation should reflect this
- A recall-oriented bias in supervised and semi-supervised learning results in models that generalize better to new domains
- More details: http://tinyurl.com/ar-ner-tr

Future work

- Modeling the various **entity categories**, including domain-specific ones
- Entity **coreference** and **resolution** (cf. Florian et al. 2004; Cucerzan 2007; Ratinov et al. 2011)
- Further leveraging the **structure** of Wikipedia text, including page structure, hyperlinks, categories, and multilingual correspondences
- NLP tools that work at **scale** and in **real time**

Thanks for listening!

- Questions?