

Coreference Resolution using Web-Scale Statistics

Based on

Shane Bergsma

Johns Hopkins University

Stony Brook University

Stony Brook, NY

Pronoun Resolution

From Wikipedia:

“The first European to explore the California coast was... **Juan Rodríguez Cabrillo**... landing on September 28, 1542, on the shores of San Diego Bay. **He** claimed California for Spain.”

- Task: **He** → “Juan Rodríguez Cabrillo”
- Query (e.g. Google Search):
“When did Cabrillo claim California for Spain?”

ويوضح محمد أن المردود المالي لعمله بات ضئيلا ولا يكفي لسد احتياجاته الأساسية

القدس كبرى مدن فلسطين [الوطن] وعاصمتها التاريخية تم تحريرها سنة 1108
القدس كبرى مدن الوطن وعاصمتها التاريخية تم تحريرها سنة

Pronouns but also others: that, الذي, The plaintiff, المتهم, Strauss, the president of
Brazil, شركة ابل ورئيسها ستيف جوبس استحقوا ثناء خاصا من الوزير



When did Cabrillo claim California for Spain?

Web [+ Show options...](#) Results 1 - 10 of about 156,000 for When did Cabrillo claim California for Spain?. (0.15 seconds)

[Juan Rodriguez Cabrillo](#)

Spain began looking northward with the intention of increasing her empire. ... for a voyage up the **California** coast under the flag of **Spain**. ... Because of adverse winds **Cabrillo** turned back, harboring at San Miguel Island, and **did** not ...

www.sandiegohistory.org/bio/cabrillo/cabrillo.htm - [Cached](#) - [Similar](#)

[Discoverers Web: Cabrillo](#)

In his days the **Cabrillo** expedition had no major impact. **Spain did** not make anything of its **claims** to **California** until the late 18th century, ...

www.win.tue.nl/~engels/discovery/cabrillo.html - [Cached](#) - [Similar](#)

[Juan Rodriguez Cabrillo](#)

19 Mar 2000 ... Meanwhile, in 1532, **Cabrillo** traveled to **Spain** where he met Beatriz ... As the **Cabrillo** family grew, so **did** his wealth and reputation as a ship builder. ... only a few hundred miles west of the coast of **California**. ...

www.nps.gov/cabr/juan.html - [Cached](#) - [Similar](#)

[Juan Rodríguez Cabrillo - Wikipedia, the free encyclopedia](#)

Cabrillo was then commissioned by the new Viceroy of New **Spain**, Antonio de Mendoza, ... The **Cabrillo** National Monument in San Diego, **California** ...

en.wikipedia.org/wiki/Juan_Rodríguez_Cabrillo - [Cached](#) - [Similar](#)

[Juan Rodríguez Cabrillo: Biography from Answers.com](#)

He was also instructed to discover and **claim** all new lands for **Spain** and, ... **Cabrillo's** ships sailed north, reaching the coast of southern **California**. ... sailed on to Northwest Cape beyond San Francisco Bay, which he **did** not find. ...

www.answers.com/topic/juan-rodriguez-cabrillo - [Cached](#) - [Similar](#)



متى ولد فون نيومان 

All

Images

Videos

News

Maps

More

Settings

Tools

About 896,000 results (0.64 seconds)

John von Neumann / Date of birth

December 28, 1903



People also search for



Alan Turing
June 23,
1912



Oskar
Morgenstern
January 24,
1902



Charles
Babbage
December
26, 1791

[Feedback](#)

جون فون نيومان - ويكيبيديا، الموسوعة الحرة

https://ar.wikipedia.org/wiki/جون_فون_نيومان [Translate this page](#)

إحداثيات: 40°20′55″N 74°35′32″W﻿ / ﻿40.348695°N 74.592251°W﻿ / 40.348695; -74.592251 ... جون فون نيومان ولد في بودابست
هنغاريا 28 ديسمبر 1903م، وهو الأخ الأكبر من بين ثلاثة أشقاء أظهر جون مواهبه وقدراته في اللغات والرياضيات، فقام والده باستجار ...
إسهاماته - حياته

John von Neumann

American-Hungarian mathematician

John von Neumann was an American mathematician and computer scientist. He made contributions to a number of fields, including mathematics, physics, economics, computing, and statistics. [Wikipedia](#)

Born: December 28, 1903
Hungary

Died: February 8, 1957
United States

Education: ETH Zurich,
University

Spouse: Klara Dan von
1938–1957), Mariette K
1930–1937)

Quotes

*There's no sense in being
you don't even know wh*

Uses for Pronoun Resolution

- Text summarization (e.g. Google News):
“**[Juan Rodríguez Cabrillo]** claimed California for Spain in 1542.”

Summary of a case:

رفع س دعوى إخلاء ضد ص
وحكم بسجن المتهم لمدة 5 سنوات

Uses for Pronoun Resolution

- Text summarization (e.g. Google News):
“**[Juan Rodríguez Cabrillo]** claimed California for Spain in 1542.”
- Machine translation (e.g. Google Translate):
“the dog/cat doesn’t like *its* food anymore”
“Der Hund/Die Katze mag *sein*/**sein[ihr]* Futter nicht mehr”
الكلب / القطة لم يعد/تعد يحب/تحب طعامه/ها



Or consider this:

قال الرجل أن الوزير قد استدعاه
قال الرجل أن الوزير قد أقالَه الرئيس

Outline

1. Introduction to Pronoun Resolution
2. Web-Scale Statistics for Pronoun Resolution:
Leveraging World Knowledge
 - A. Noun Gender (prevailant in Arabic, together with plural types, some detectable from verbs يحب/تحب)
 - B. Pattern Coreference
 - C. Non-Referential Pronouns

1) Introduction to Pronoun Resolution

Coreference Resolution:

- “**President Obama** says **he** plans to discuss **these issues** with **other leaders** at **the summit**.”
- What do we know about these *nouns*? What do they *refer* to? Which are *anaphora*?
- PhD topics: NP coreference, other-anaphora, detecting anaphora, resolving pronouns, etc.
- Related topics: database de-duping, citation matching, cross-document coreference, and many others

Anaphora: the use of a word referring back to a word used earlier in text or conversation, to avoid repetition, for example the pronouns *he, she, it,* and *they* and the verb *do* in “*I like it and so do they*”.

Pronoun Resolution

- Scope: third-person anaphoric pronouns, including reflexives:
 - *He, his, him, himself* هو، هما، هم؟ (masculine)
 - *She, her, herself* هي هن (feminine)
 - *It, its, itself* (neutral)
 - *They, their, them, themselves* هم، هن، هما (plural)
 - بالعربية هنالك الضمائر المتصلة: نحتاج لتحديد **ها** قبل التعامل معها «اشترى محمد لعمر وهدية واعطاه **اياها**»

Pronoun Resolution

“In 2004, Exxon Mobil paid *its* chairman, Lee Raymond, a total of \$38.1 million.”

في عام 2004 دفعت شركة اكسون-موبيل لرئيسها، لي رايموند،
ما مجموعه 38.1 مليون دولار

Question: “Who is the chairman of Exxon Mobil?”

Goal:

Resolve pronoun: “*its*” ها

To antecedent: “Exxon Mobil”

Get: “Exxon Mobil’s Chairman, Lee Raymond”

Traditional Pronoun Resolution

“In 2004, Exxon Mobil paid *its* Chairman Lee Raymond a total of \$38.1 million.”

Resolving a pronoun typically involves:

1. Parse text to determine **noun phrases**
2. Build list of preceding nouns as candidates
3. Filter candidates based on gender/number agreement, grammar violations, etc.
4. Select **most likely** candidate of remaining nouns based on measures of frequency, emphasis, etc.

Linguistic Knowledge

- a) “John never saw the car. He arrived late.”
“John never saw the car. It arrived late.”

Linguistic Knowledge

a) “John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

Knowledge = Noun Gender

Linguistic Knowledge

a) “John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

b) “John needs his friend.”

“John needs his support.”

Linguistic Knowledge

a) “John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

b) “John needs his friend.”

“John needs his support.”

**Knowledge =
You don't need your
own support**

Linguistic Knowledge

a) “John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

b) “John needs his friend.”

“John needs his support.”

c) “You can make it in advance.”

“You can make it in Hollywood.”

Linguistic Knowledge

a) “John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

b) “John needs his friend.”

“John needs his support.”

Knowledge = “make it” is an idiom

c) “You can make it in advance.”

“You can make it in Hollywood.”

Linguistic Knowledge

a) “John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

b) “John needs his friend.”

“John needs his support.”

c) “You can make it in advance.”

“You can make it in Hollywood.”

Knowledge via Web-Scale Counts

- Use an Internet search engine to harness all the linguistic data on the web



“Britney Spears”

64,100,000 pages

“Britany Spears”

434,000 pages

Deep Knowledge via Web-Scale Counts

- Positive or negative sentiment, connotation (Turney '02)

– E.g. “unethical”

unethical excellent

916,000 pages

unethical poor

1,600,000 pages

A) Noun Gender Knowledge

“John never saw the car. He arrived late.”

“John never saw the car. It arrived late.”

Learning Noun Gender

- **Input:**
 - English noun (*John, car, Google*)
- **Output:**
 - Gender class: masculine (*he*), feminine (*she*), neutral (*it*), or plural (*they*)
- **Result:** improved pronoun resolution

If we had labeled data...

- “John never saw the car. He arrived late.”
John.countMale++;
- “John loves his Honda.”
John.countMale++;
- “Was that John? Has he lost weight?”
John.countMale++;
- John: 100% male

Web-Mining Gender

- Count number of pages returned, e.g:
“John * himself” “John * herself” “John * itself”

Web-Mining Gender

- Count number of pages returned, e.g:
“John * himself” “John * herself” “John * itself”
- Other patterns:
“John * his” “John * her” “John * its”
“John * he” “John * she” “John * it”
...

WordNet vs. Statistical Gender

Noun	WordNet: Masculine acceptable?	Corpus Reflexives: P(Masculine)
John	OK	99.7%
Company	OK	0% (93% neutral)
Computer	OK	0% (99.2% neutral)

Supervised Machine Learning

- $\mathbf{x} = (x_1, x_2, \dots, x_n)$: Features (log-counts)
- $\mathbf{w} = (w_1, w_2, \dots, w_n)$: Learned weights
- Decision function:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$$

- Set the weights using a small number of labeled examples (SVM)

Gender Implementation

- Get counts using Google API
- Tested:
 - Gender assignment accuracy
 - within end-to-end pronoun resolution system
(not as common as you might think)

Gender Results

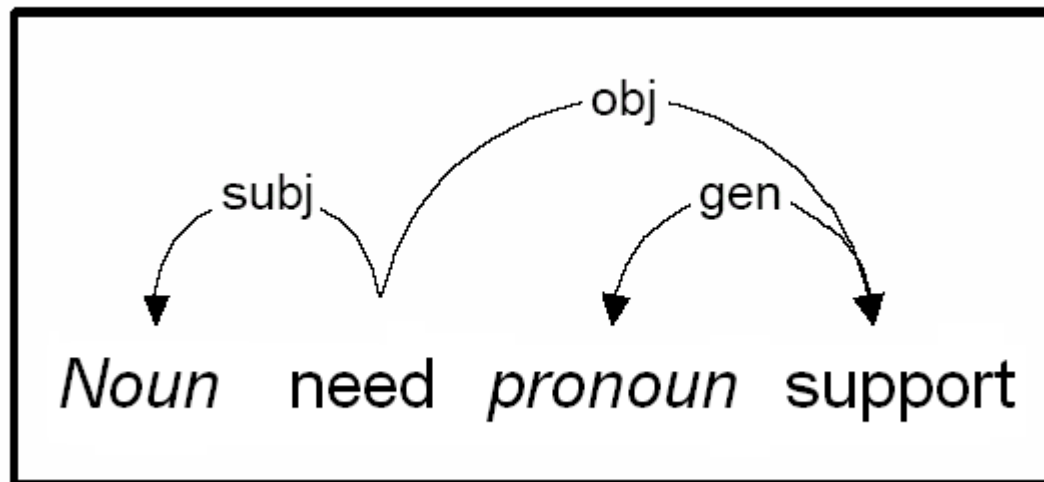
- Gender classification F-Score
 - **85.4%** using a corpus
 - **90.4%** using the web
 - **92.2%** using both
 - **88.8%** for humans
- Pronoun resolution accuracy:
 - From **63.2%** to **73.3%** when adding learned gender information

B) Pattern Coreference

- “John needs **his** friend.”
- “John needs **his** support.”
- “John offered **his** support.”

Dependency Path/Pattern

- Sequence of dependency links between terminal entities in a typed dependency tree



- Short form: Noun needs pronoun's support

Learning Pattern Coreference

- Input:

$p = \text{“}\underline{\textit{Noun}} \text{ needs } \underline{\textit{pronoun's}} \text{ support”}$

- Output:

- Are *Noun* and *pronoun* coreferent in p ?

Learning Pattern Coreference

- If we had labeled data...
 - Count how often noun and pronoun co-refer in each labeled *p*
- Instead, learn from unambiguous examples:

“We need your support”

» not coreferent

Pronoun Class Agreement

<u>I</u> need <u>his</u> support	Disagree
<u>They</u> need <u>my</u> support	Disagree
<u>We</u> need <u>her</u> support	Disagree
<u>He</u> needs <u>his</u> support	Agree
<u>I</u> need <u>your</u> support	Disagree

Agreement = 20% → *non-coreferent*

Learned Patterns

- Example non-coreferent patterns:
 - “**John** married **his** mother.”
 - “**Sue** wrote **her** obituary.”
- Example coreferent patterns:
 - “**Google** says **it** intends to...”
 - “**The revolutionaries** consolidated **their** power.”

Using Pattern Coreference

- Use it directly as a feature in your system
- Enhancing a semantic compatibility model
- Use it to bootstrap probabilistic noun gender/number information:
 - “*The newspaper* says *it* intends to...”
 - Assume coreference, count as instance of “*newspaper*” being neutral

Results

- Adding path coreference:
 - From 1.3% to 1.8% improvement on three datasets (significant, $p=0.05$)
- Share large corpus of noun gender collected using the patterns
 - Useful for many people (e.g., used in NLP course at Stanford, 2011 CoNLL shared task)
 - Described in Jurafsky & Martin 2nd Edition

C) Non-Referential Pronouns

E.g. the word “*it*” in English

- “You can make **it** in advance.”
 - referential (50-75%)
- “You can make **it** in Hollywood.”
 - non-referential (25-50%)

Non-Referential Pronouns

- [Hirst, 1981]: detect non-referential pronouns, “lest precious hours be lost in bootless searches for textual referents.”
- Most existing pronoun/coreference systems just ignore the problem
- A common ambiguity:
 - “*it*” comprises 1% of English tokens

Non-Referential Pronouns

- Not just an English phenomenon:
 - “Wie geht *es* Ihnen?” (German)
 - “S’*il* vous plaît.” (French)
- Non-ref pronouns also in pro-drop languages:
 - “<> Es importante.” (Spanish, referential)
 - “<> Es importante que ...” (Spanish, non-referential)

Non-Ref Detection as Classification

- Input:

s = “You can make *it* in advance”

- Output:

Is *it* a non-referential pronoun in s ?

Method: train a supervised classifier to make this decision on the basis of some features

[Evans, 2001, Boyd et al. 2005, Müller 2006]

A Machine Learning Approach

$$h(\underline{x}) = \underline{w} \cdot \underline{x} \quad (\text{predict non-ref if } h(\underline{x}) > 0)$$

- Typical ‘lexical’ features: binary indicators of context:

$\underline{x} = (\text{previous-word}=\textit{make}, \text{next-word}=\textit{in}, \text{previous-two-words}=\textit{can+make}, \dots)$

- Use training data to learn good values for the weights, $\underline{w}$
 - Classifier learns, e.g., to give negative weight to PPs immediately preceding ‘*it*’ (e.g. *... from it*)

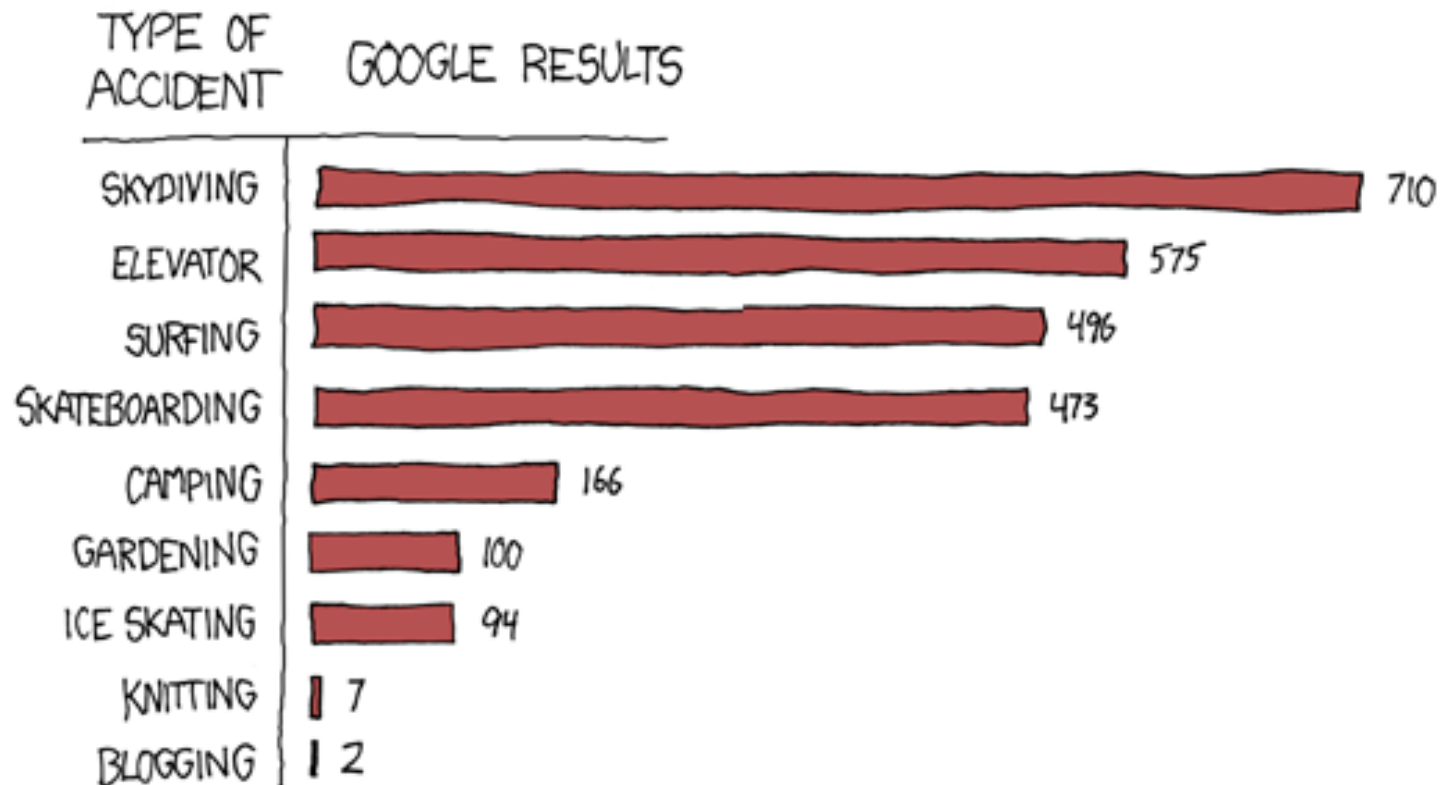
Better: Features from the Web

[Bergsma, Lin, Goebel, ACL 2008, IJCAI 2009]

- Convert sentence to a context pattern:
“make _____ in advance”
- Collect counts from the web:
 - “make *it/them* in advance”
 - 442 vs. 449 occurrences in Google N-gram Data
 - “make *it/them* in Hollywood”
 - 3421 vs. 0 occurrences in Google N-gram Data

DANGERS

INDEXED BY THE NUMBER OF GOOGLE RESULTS FOR
"DIED IN A _____ ACCIDENT"



From:
xkcd.com

Applying the Web Counts

- How wide should the patterns span?
 - We can use all that Google N-gram Data allows:

You can make _
can make _ in
make _ in advance
_ in advance .

- Five 5-grams, four 4-grams, three 3-grams and two bigrams
- What fillers to use? (e.g. *it, they/them, any NP?*)

Web Count Features

“it”:

log-cnt(“You can make it in”) 5-grams

log-cnt(“can make it in advance”)

log-cnt(“make it in advance .”)

...

log-cnt(“You can make it”) 4-grams

log-cnt(“can make it in”)

...

...

“them”:

log-cnt(“You can make them in”) 5-grams

...

...

A Machine Learning Approach Revisited

$$h(\underline{\mathbf{x}}) = \underline{\mathbf{w}} \cdot \underline{\mathbf{x}} \quad (\text{predict non-ref if } h(\underline{\mathbf{x}}) > 0)$$

- Typical features: binary indicators of context:
 $\underline{\mathbf{x}} = (\text{previous-word}=\textit{make}, \text{next-word}=\textit{in}, \text{previous-two-words}=\textit{can+make}, \dots)$
- New features: real-valued counts in web text:
 $\underline{\mathbf{x}} = (\text{log-cnt}(\textit{make it in advance}), \text{log-cnt}(\textit{make them in advance}), \text{log-cnt}(\textit{make * in advance}), \dots)$
- Key conclusion: classifiers with web features are robust on new domains! [Bergsma, Pitler, Lin, ACL 2010]

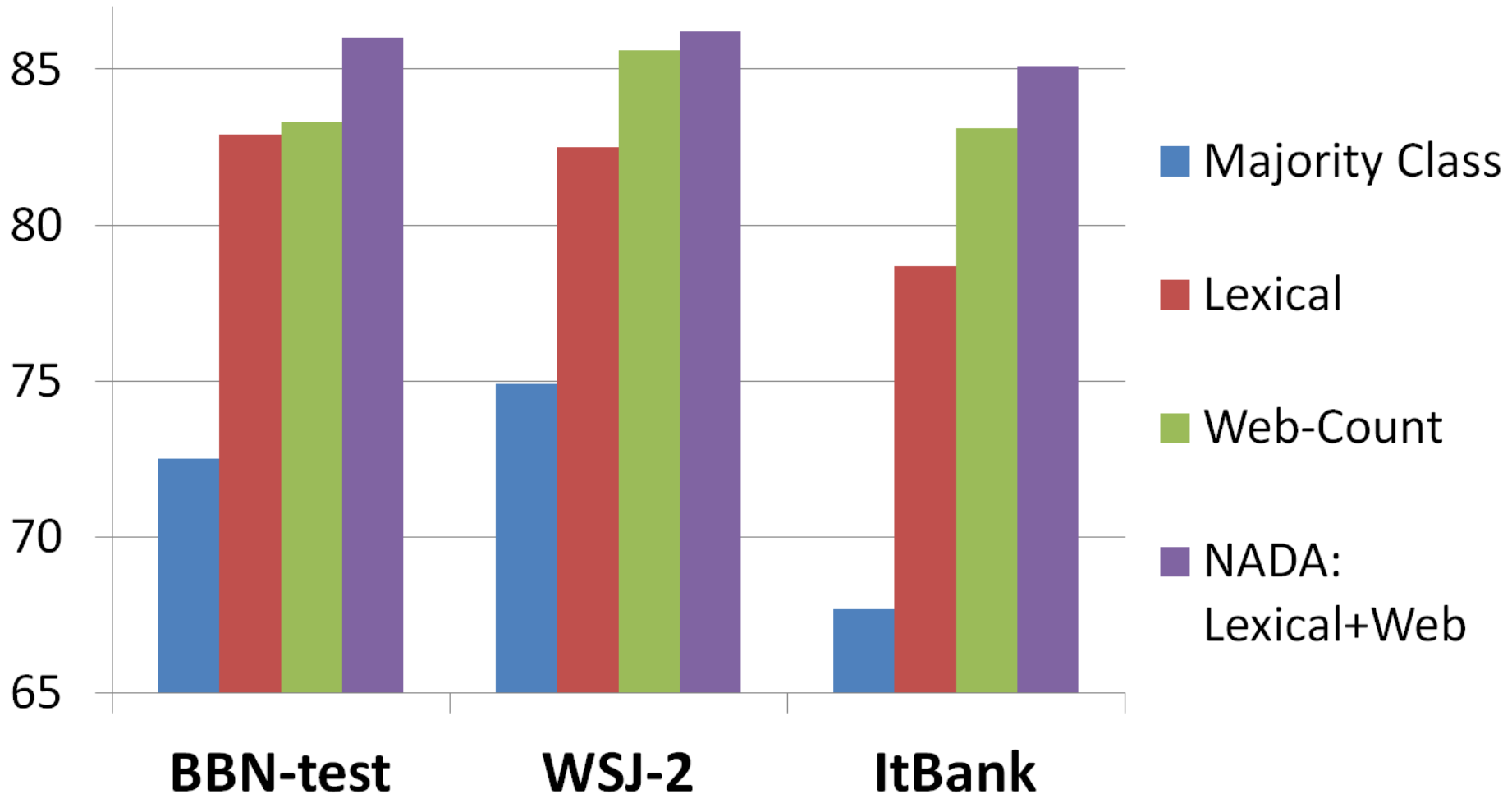
NADA [Bergsma & Yarowsky, DAARC 2011]

- Non-Anaphoric Detection Algorithm:
 - a system for identifying non-referential pronouns
- <http://code.google.com/p/nada-nonref-pronoun-detector/>
- Works on raw sentences; no parsing/tagging of input needed
- Classifies 'it' in up to 20,000 sentences/second
- It works well when used out-of-domain
 - Because it's got those *Web count* features

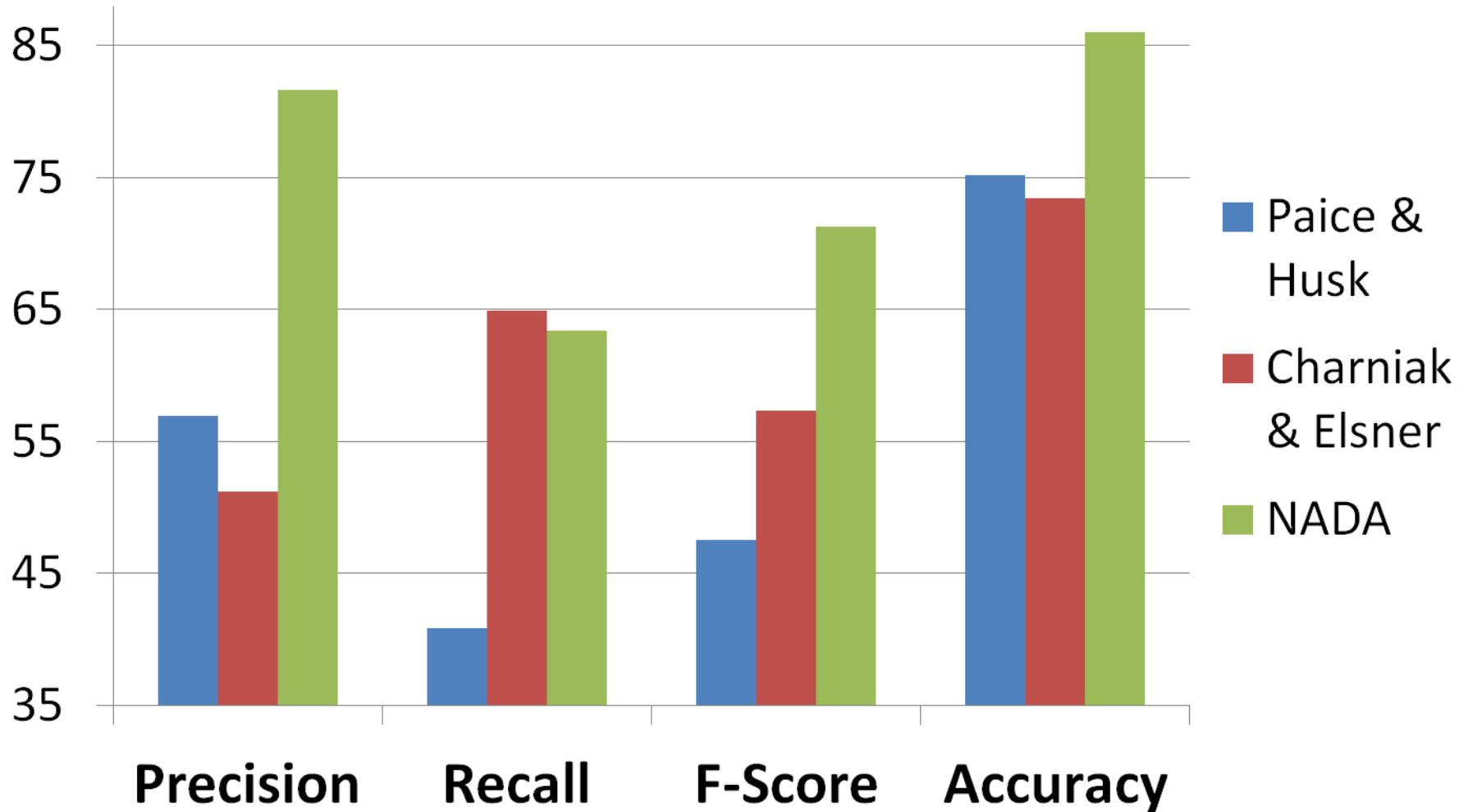
Using web counts works great... but is it practical?

All N-grams in the Google N-gram corpus	93 GB
Extract N-grams of length-4 only	33 GB
Extract N-grams containing <i>it, they, them only</i>	500 MB
Lower-case, truncate tokens to four characters, replace special tokens (e.g. named entities, pronouns, digits) with symbols, etc.	189 MB
Encode tokens (6 bytes) and values (2 bytes), store only changes from previous line	44 MB
gzip resulting file	33 MB

Accuracy with Different Features



NADA versus Other Systems



Conclusion

- Web-derived knowledge is very useful for pronoun resolution:
 - Gender
 - Pattern coreference
 - Non-referential pronoun patterns
- Useful to share data and software that leverage web-scale knowledge
 - always the best way to have an impact on the field

Thanks