

---

---

# New Tools for Web-Scale N-grams

**Dekang Lin, Kenneth Church, Heng Ji,  
Satoshi Sekine, David Yarowsky, Shane Bergsma,  
Kailash Patil, Emily Pitler, Rachel Lathbury,  
Vikram Rao, Kapil Dalwani, Sushant Narsale**

**Presented by: Shane Bergsma,  
University of Alberta  
LREC 2010**



---

---

# The Team

<b>Member</b>	<b>Affiliation</b>	<b>Member</b>	<b>Affiliation</b>
Dekang Lin	Google	Ken Church	JHU
Heng Ji	CUNY	Satoshi Sekine	NYU
David Yarowsky	JHU	Shane Bergsma	Univ. of Alberta
Kailash Patil	JHU	Emily Pitler	UPenn
Rachel Lathbury	Univ. of Virginia	Vikram Rao	Cornell
Kapil Dalwani	JHU	Sushant Narsale	JHU



---

---

# Goals

- Investigate the use of web-scale N-grams
- Create tools for the NLP community:
  - Better tools for big data
  - Flexible, efficient ways to collect counts from web-scale text
- Apply big data to big problems



---

---

# Search Engines vs. N-grams

- Search Engines
  - Too slow for *millions* of queries
- Web-Scale N-gram Corpus:
  - Compressed version of text on web
  - N words in sequence + their count on web:

**Workshop at ACL**                      **367**

**Workshop at COLING**                      **53**

**Workshop at LREC**                      **156**

...



---

---

# N-grams For Lexical Knowledge

- Animate Nouns:
  - **divorcee** is animate, **divorce** is not
- Simple patterns: “**NP who**” vs. “**NP which**”

...

recent conversation	which 10	
recent debate	which 10	
recent divorcee	who 60	
recent meeting	which 232	who 13
recent opinion poll	which 24	

...

---

---

# N-gram Data

- Google N-gram Version 1:
  - 1 trillion token corpus (Brants & Franz, 2006)
- Google N-gram Version 2: with POS tags
  - De-duped, converted digits to '0', URLs and e-mail addresses to '<URL>' and '<EMAIL>'
  - Today: focus on tools for Google V2



---

---

# N-gram Data

- N-grams in Wikipedia
  - by Satoshi Sekine at NYU
- Inverted-Index Tools:
  - Part-of-speech, chunk, and named-entity N-gram matching in Wikipedia
  - **Sekine & Dalwani, LREC 2010:**
    - **Today, 18:20-19:40, P34: Knowledge Discovery**

---

---

# Google N-grams Version 2

- POS Tags:

<b>flies</b>	<b>1643568</b>	<b>NNS 611646 VBZ 1031922</b>
<b>caught the flies ,</b>	<b>11</b>	<b>VBD DT NNS , 11</b>
<b>plane flies really well</b>	<b>10</b>	<b>NN VBZ RB RB 10</b>

- Organization

- 1000 files, 500 MB each, roughly 500 GB total
- Index → given a query, seek to a position in a file





---

---

# Tool Design

- Typical usage: Retrieve all the N-grams containing the word ***cheetah***
- Typical N-gram Data:
  - ...
  - cheetah eats grass**
  - cheetah is an animal**
  - ...
  - faster than a cheetah**
  - ...



---

---

# Rotated N-grams

**faster than a cheetah →**

**faster than a cheetah**

**than a cheetah >< faster**

**a cheetah >< than faster**

**cheetah >< a than faster**

- Sort rotated N-grams: all the N-grams containing ***cheetah*** are now sequential

---

---

# *cheetah* N-grams

<b>cheetah &gt;&lt; a by attacked</b>	<b>13</b>	<b>VBN IN DT NN 13</b>
<b>cheetah &gt;&lt; captive-born</b>	<b>12</b>	<b>JJ NN 12</b>
<b>cheetah &gt;&lt; endangered the save</b>	<b>12</b>	<b>VB DT JJ NN 12</b>
<b>cheetah &gt;&lt; missing a rescue</b>	<b>21</b>	<b>VB DT JJ NN 21</b>
<b>cheetah &gt;&lt; stuffed</b>	<b>69</b>	<b>VBD NN 8 VBN NN 61</b>
<b>cheetah attacks</b>	<b>26</b>	<b>NN NNS 22 NN VBZ 4</b>
<b>cheetah breeding</b>	<b>248</b>	<b>NN NN 55 NN VBG 193</b>
<b>cheetah chasing a gazelle</b>	<b>12</b>	<b>NN VBG DT NN 12</b>
<b>cheetah enclosure</b>	<b>100</b>	<b>NN NN 100</b>
<b>cheetah fur</b>	<b>109</b>	<b>NN NN 109</b>
<b>cheetah habitat</b>	<b>131</b>	<b>NN NN 131</b>

...



---

---

# Patterns

`(word-seq ([A-Z] [A-Z]* 0000 Workshop))`

- Apply to all N-grams that contain “Workshop”

# Patterns

(word-seq ([A-Z] [A-Z]\* 0000 Workshop))

ACL	524	AAAI	229	INEX	83	SIGMM	45
OOPSLA	475	AAMAS	189	UML	68	IJCAR	45
CHI	452	CLEF	167	ECDL	67	AOSD	41
ECOOP	384	NIPS	159	ICAPS	66	GECCO	40
SIGIR	346	EACL	157	ICDM	58	IROS	39
ACM	291	NAACL	151	JSAI	55	PRICAI	37
ICSE	273	ESSLLI	151	SIGCOMM	53	GONG	37
IJCAI	261	COLING	128	FNCA	53	CVPR	36
LREC	245	CSCW	116	KDD	50	AIPS	34
ECAI	244	ITS	102	VR	47	ETAPS	33
IEEE	243	WWW	89	IPDPS	47	LICS	32
SIGPLAN	230	ICML	89	VLDB	46	ISWC	31



---

---

# Applications of Patterns

- Lexical Property: **Countability**
- The noun ***water*** is not countable:
  - much water, some water, etc. → good
  - many waters, a water → bad
- “some water”      169,017
- “a water”            1,048,362            ???

---

---

# Applications of Patterns

a water {supply, bath, bottle, system, tank, treatment, molecule, tower, shortage, filter, balloon, buffalo, fountain, pipe...}



---

---

# Patterns – using POS tags

- Composite patterns:  
(seq (word = a)  
      (word = water)  
      (tag ~ [^N].\*))

doesn't match:

a water bottle

a water tank





---

---

# Commands

- Commands:
  - Process returned N-grams
  - Count things, print things
- Modes:
  - batch processing*: collect information for *all* NPs
  - vs.
  - sequential*: get counts for one NP at a time

---

---

# Availability

- Data: Google V2 coming soon
- Code:
  - <http://code.google.com/p/ngramtools/>
  - For matching raw text AND N-grams
- Applications:
  - Ji & Lin, Gender & Number for Mention Detection, PACLIC 2009
  - Bergsma, Pitler, & Lin, Web-scale N-grams in Supervised Classifiers, ACL 2010



---

---

# Thanks

- **Center for Language & Speech Processing, Johns Hopkins University**
- **IBM/Google Academic Cloud Computing Initiative**
- **Workshop Sponsors:**
  - **NSF, Google Research, DARPA**

