Web-Scale N-Gram Models for Lexical Disambiguation

Shane Bergsma University of Alberta

Dekang Lin Google, Inc.

Randy Goebel University of Alberta

IJCAI 2009

N-grams for Disambiguation

- **Problem:** Choose a label for a word in text
  - **Noun** or **verb**? **Sense 1** or **Sense 2**?
- **Method:** Which *label* is most frequent in the word's (N-gram) *context*?
- Get counts from web-scale text
- Combine counts from multiple segments of context

Outline

1. Lexical Disambiguation
2. Gathering Web-Scale Counts
3. Combining Context Counts
4. Applications
   - **Preposition Selection**
   - **Context-Sensitive Spelling Correction**
   - **Non-Referential Pronoun Detection**

Lexical Disambiguation

- Choosing the correct meaning of a word from a set of candidates
- Input: a word in context

  "Bob ate a huge **bass** for dinner."
- Output: a label, e.g.:

or

Lexical Disambiguation

- Different meanings, same surface form:

  "Let me know **weather** you like it."

  "**weather**" or "**whether**"

- Also: Diacritic restoration, POS-tagging, etc. (Yarowsky 1994, Roth 1998)

Lexical Disambiguation

- Use corpus occurrences as unambiguous examples:
  - "know **weather** you" vs. "know **whether** you"
- Terminology:
  - "know _ you" : context pattern
  - "**weather**", "**whether**" : fillers
  - Get counts for fillers in context patterns, take highest-scoring as label

Non-word Labels

- Devise proxies for labels, get pattern counts (Mihalcea & Moldovan, 1999)
- "Bob ate a huge **bass** for dinner."

Sense	Proxies
	**tuna**, **salmon**, **pike**
	**guitar**, **drums**, **harmonica**

Web-Scale Data

- Where to get the counts?
  - More data = better data (Banko & Brill, 01)
  - Hmmm...

Search engine page-counts = *Awesome* corpus counts

Previous work

- Lapata & Keller 2005:
  - Query web with trigram of context:
  - "know weather you" : 1,370,000 pages
  - "know whether you" : 1,600,000 pages
- Correct one is higher, but ???

Previous work

- Lapata & Keller 2005:
  - Query web with trigram of context:
  - "know weather you" : 1,370,000 pages
  - "know whether you" : 1,600,000 pages
- Correct one is higher, but ???
  - July 6, 2009:
  - "know weather you" : 4,060 pages
  - "know whether you" : 2,530,000 pages

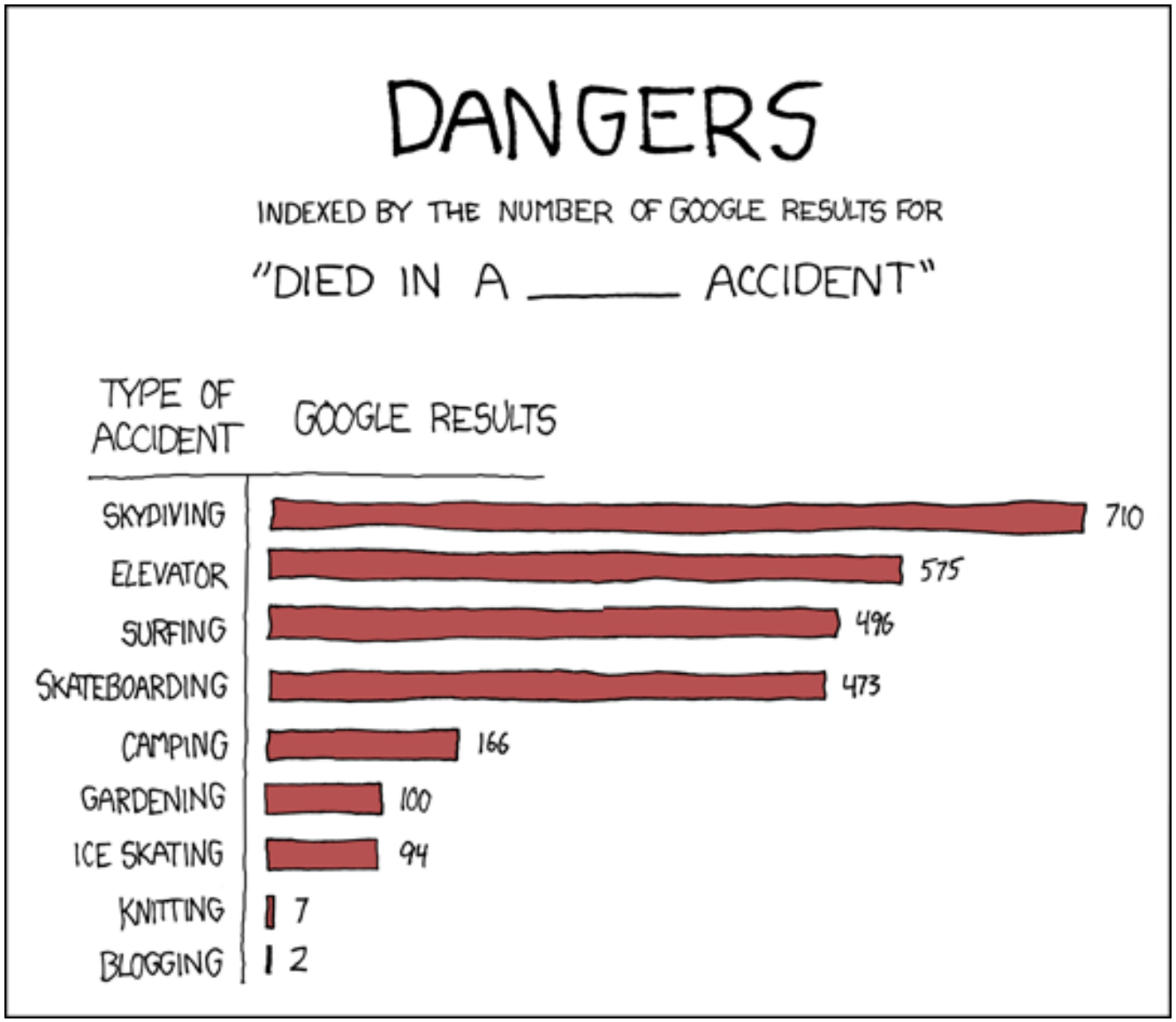
Google N-gram Data

- 2006: Google releases web-scale N-gram corpus
- From 1 trillion words of online English text
  - Doesn't fit on your hard drive
- *1*-grams to *5*-grams with > 40 counts
- A compressed version of the whole web
  - Approximately 24 GB gzipped
  - Does fit on your hard drive

Web vs. N-Gram Corpus

- For training a preposition selection system, needed 267 million unique counts.
  - Using Google API with 1000 query/day limit, that would have taken over 732 years
- Search-engine counts are extremely inefficient

How much context to include?



From: xkcd.com

Multiple Patterns

- Many contexts span the confusable word:

```
Let me know _
    me know _ you
        know _ you like
               _ you like it
```

  - Five 5-grams, four 4-grams, three 3-grams and two 2-grams span the confusable word
- Like a LM

Combining Counts

- ***SuperLM:***
  - Use supervised machine learning to combine counts (Bergsma et al., ACL 2008)
  - Features:

`log(Count(`*`context-pattern`*`{filler}))`

    - indexed by pattern position, length, filler, class
  - learns the association of fillers and classes
  - exploits most predictive fillers, positions

Example

- "... to choose **among/between** the three candidates..."
- Predicting: is it **among**?

Feature	Weight
log(C("to choose among"))	+1
log(C("to choose between"))	-1
log(C("among the three"))	+3
log(C("between the three"))	-3

Other Approaches

- ***Trigram:***
  - Compare trigram counts of fillers, take highest as label
- ***SumLM:***
  - Sum the log-frequencies across all context patterns for each filler, take highest as label

Applications

1) Preposition Selection

"Study in California **at** UCLA."

– Fillers: 34 prepositions: "at, by, from, in, on..."

System	Accuracy
Baseline	**20.9%**
Trigram	**58.8%**
SumLM	**73.7%**
SuperLM	**75.4%**

Applications

1) Preposition Selection

"Study in California **at** UCLA."

– Fillers: 34 prepositions: "at, by, from, in, on..."

System	Accuracy
Baseline	20.9%
Trigram	58.8%
SumLM	73.7%
SuperLM	75.4%

Applications

1) Preposition Selection

"Study in California **at** UCLA."

– Fillers: 34 prepositions: "at, by, from, in, on..."

System	Accuracy
Baseline	20.9%
Trigram	58.8%
SumLM	73.7%
SuperLM	75.4%

SumLM from ***MIN*** to ***MAX***

	MAX			
MIN	**2**	**3**	**4**	**5**
2	**50.2%**	**63.8%**	**70.4%**	**72.6%**
3		**66.8%**	**72.1%**	**73.7%**
4			**69.3%**	**70.6%**
5				**57.8%**

SumLM from ***MIN*** to ***MAX***

	MAX			
MIN	**2**	**3**	**4**	**5**
2	**50.2%**	**63.8%**	**70.4%**	**72.6%**
3		**66.8%**	**72.1%**	**73.7%**
4			**69.3%**	**70.6%**
5				**57.8%**

SumLM from ***MIN*** to ***MAX***

	MAX			
MIN	**2**	**3**	**4**	**5**
2	**50.2%**	**63.8%**	**70.4%**	**72.6%**
3		**66.8%**	**72.1%**	**73.7%**
4			**69.3%**	**70.6%**
5				**57.8%**

Applications

2) Context-Sensitive Spelling Correction

– Fillers: among/between, amount/number, cite/sight/site, peace/piece, raise/rise.

System	Accuracy (Avg.)
Baseline	66.9%
Trigram	88.4%
SumLM	94.8%
SuperLM	95.7%

Applications

3) Non-referential Pronoun Detection

"**it** is hungry." vs. "**it** is important to eat."

– Fillers: "it", "he/she/they/etc.", ".*" (proxies)

System	Accuracy
Baseline	59.4%
Trigram	74.3%
SumLM	79.8%
SuperLM	82.4%

Conclusion

- Web-scale N-gram counts for many tasks
- Use as much context as possible, combine in intelligent ways
- Get state-of-the-art performance
- Johns Hopkins Summer Workshop 2009:
  - *Unsupervised Acquisition of Lexical Knowledge from N-grams*
  - Google N-grams Version 2: with POS-tags!

Thanks!

ALBERTA INGENUITY

NSERC CRSNG

iCORE