Distributional Identification of Non-Referential Pronouns

Shane Bergsma
University of Alberta

Dekang Lin
Google, Inc.

Randy Goebel
University of Alberta

ACL 2008

Non-Referential Pronouns

- No antecedent, e.g. "**It**'s raining men"
- Input: context of pronoun:

  `_ 's raining men`
- Output:

  Non-referential? Yes/No

Outline

1. Pronoun Resolution and Non-Referential Pronouns
2. Distributional Identification
3. Methodology
4. Experiments
5. Results

1) Pronoun Resolution

- Pronouns like ***he, she, his, they, them*** are used in place of nouns in text:
  - "***It*** was founded in 1940 by Dick and Mac McDonald in San Bernardino, California"
- Resolve pronoun ***It*** to antecedent (McDonald's)
  - For QA, Summarization, Translation, etc.

Non-Referential Pronouns

- Not all pronouns refer to preceding noun
- The pronoun ***it*** can refer to whole clauses or sentences, or simply be a grammatical placeholder (pleonastic):
  - "***It*** is important to eat your veggies."
  - "***It*** is raining cats and dogs… and men."

Prevalence

- ***it*** accounts for about 1% of tokens in text
- Between 25% and 50% of these are non-referential
- Non-referential pronouns are not just an English phenomenon:
  - "Wie geht ***es*** Ihnen?"
  - "S'il vous plaît."
  - Same in pro-drop languages

Previous Work

- Hirst (1981): detect non-referentials, "lest precious hours be lost in bootless searches for textual referents."
- Recent work: rule-based or machine learned systems based on manually-defined features of the pronoun's context

2) Distributional Identification

- Our hypothesis:
  - The contexts of non-referential pronouns are unique to these pronouns
  - Therefore in text these contexts mostly occur with the pronoun ***it***
- If we can estimate how often a context occurs with ***it***, we can estimate how likely it is to be non-referential.

Distributional Identification

- E.g.
  - "You can make ***it*** in advance."
  - "You can make ***it*** in Hollywood."
- How often do the contexts have fillers other than ***it***?

Distributional Identification

- Convert to a pattern:

  `“make _ in advance”`

- Collect counts from N-gram data:
  - “**`make it/them in advance`**”
    - 442 vs. 449 occurrences in N-gram data
  - “**`make it/them in Hollywood`**”
    - 3421 vs. 0 occurrences in N-gram data

Distributional Identification

- Convert to a pattern:

  "make _ in advance"

- Collect counts from N-gram Data:

  Referential

  - "**make *it/them* in advance**"
    - 442 vs. 449 occurrences in N-gram data
  - "**make *it/them* in Hollywood**"
    - 3421 vs. 0 occurrences in N-gram data

Distributional Identification

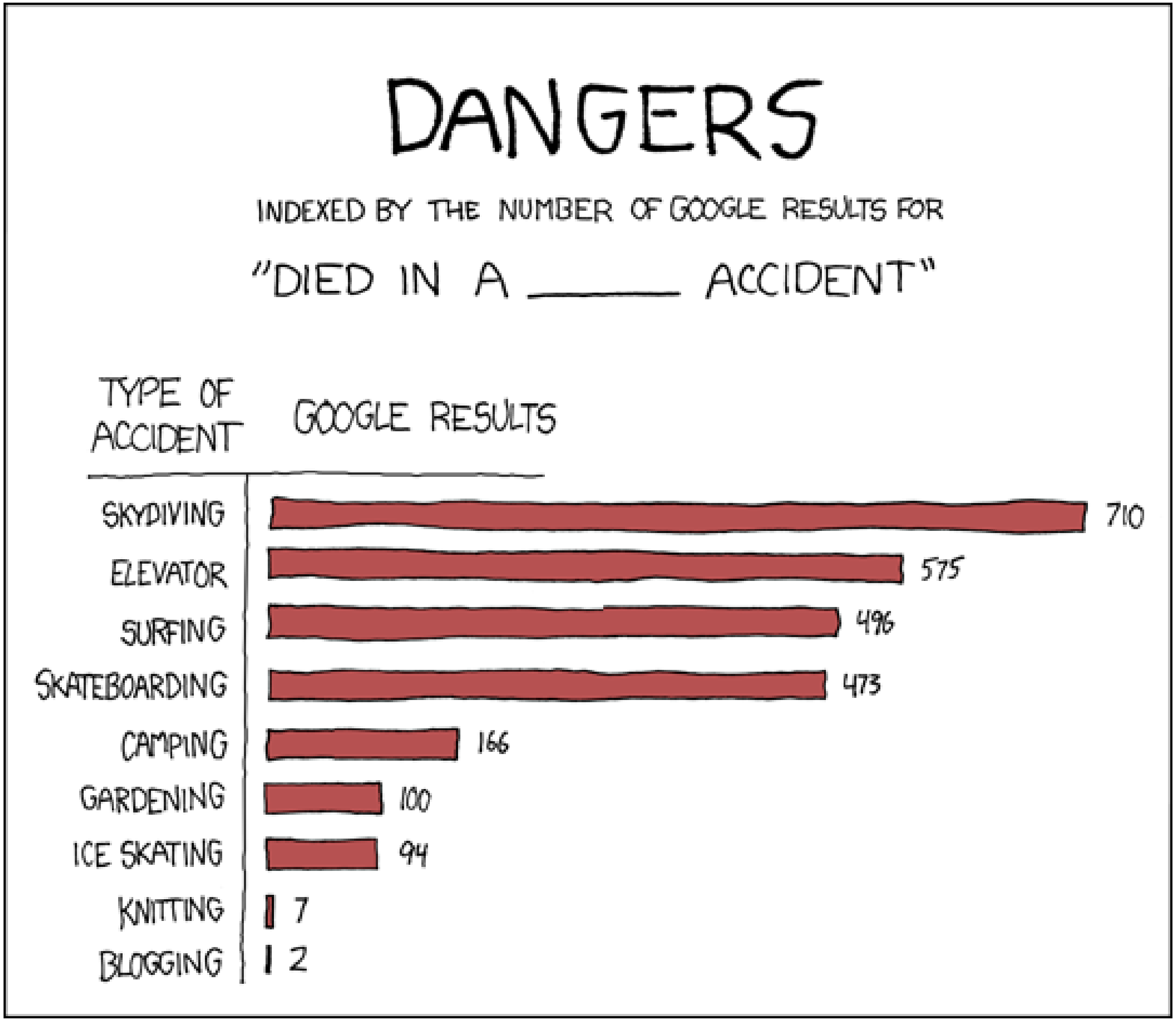
- Convert to a pattern:

  "make _ in advance"

- Collect counts from N-gram Data:
  - "**make *it/them* in advance**"
    - 442 vs. 449 occurrences in N-gram data
  - "**make *it/them* in Hollywood**"
    - 3421 vs. 0 occurrences in N-gram data

Referential

Non-Referential



From:
xkcd.com

3) Methodology

- How much context to include in a pattern?
- There are many possible patterns that span the pronoun – which do we use?
- How do we use the distribution for classification?

Methodology

- How much context?
  - N-gram corpus has up to 5-grams
- E.g.

... said here Thursday that ***it*** is unnecessary to continue ...

said here Thursday that _

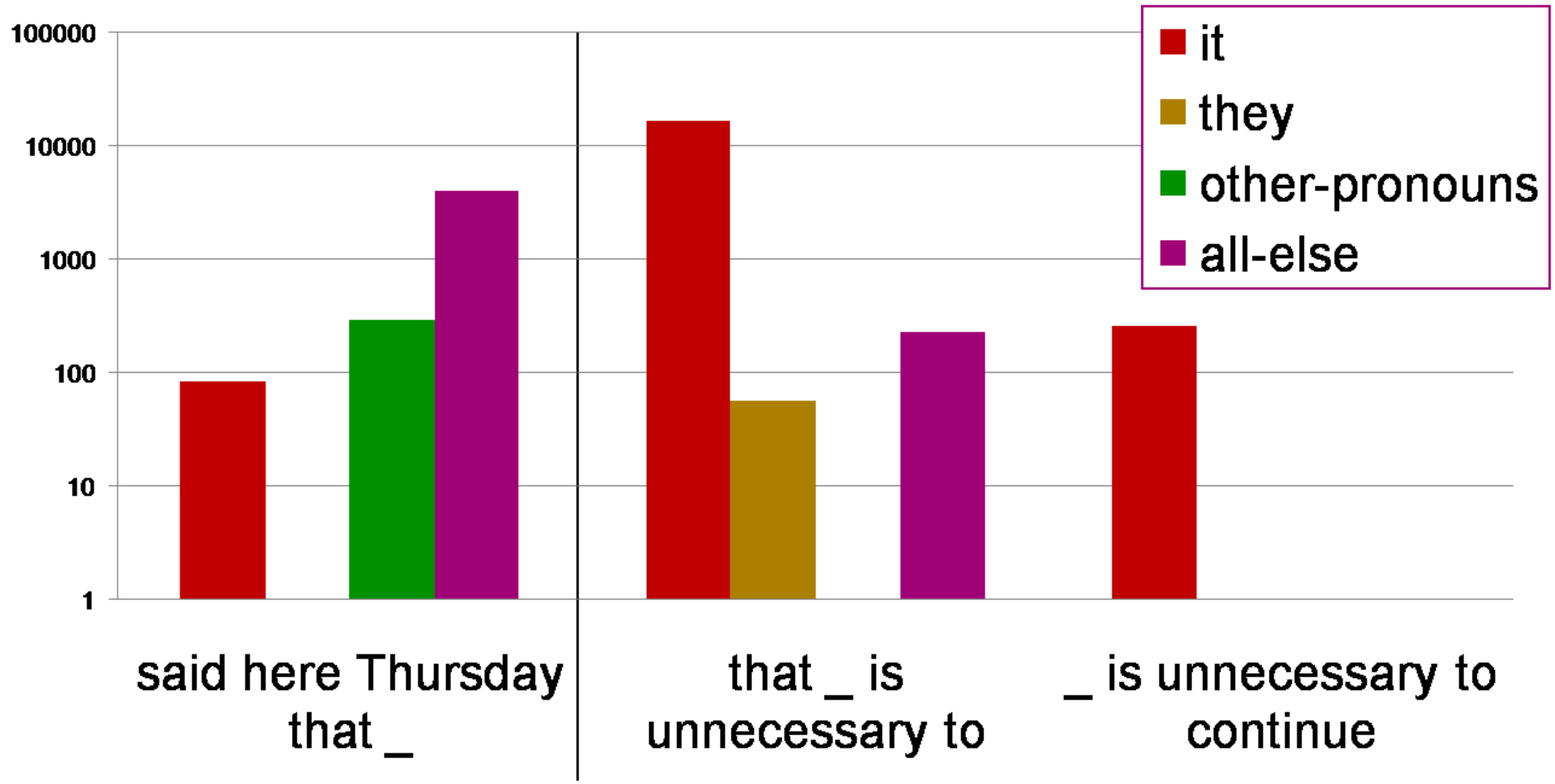
here Thursday that _ is

Thursday that _ is unnecessary

that _ is unnecessary to

_ is unnecessary to continue

Filler Counts by Type/Position



Combining Counts

- Use supervised machine learning to optimally combine count information
- Define non-referential as positive class, referential as negative
- Use ***`MaxEnt`*** to learn classifier
- Features: **`log(#counts)`**
  - indexed by pattern position and filler type

Labelled Data

- Data indicates, for every instance of ***it***, whether it refers to preceding noun or not
- Extracted this information directly from MUC-7 coreference corpus
- Also annotated data from other sources
  - Good agreement among three annotators (κ between 0.8 and 0.9)

Labelled Data

Data Set	Number of *It*	% Non-Referential
Europarl	279	50.9
Science News	1020	32.6
WSJ	709	25.1
MUC	129	31.8
Train	1069	33.2
Test	1067	31.7

4) Experiments

- Evaluation Criteria:
  - Precision:

    Percent of our system's positive non-referential decisions which are truly non-referential
  - Recall:

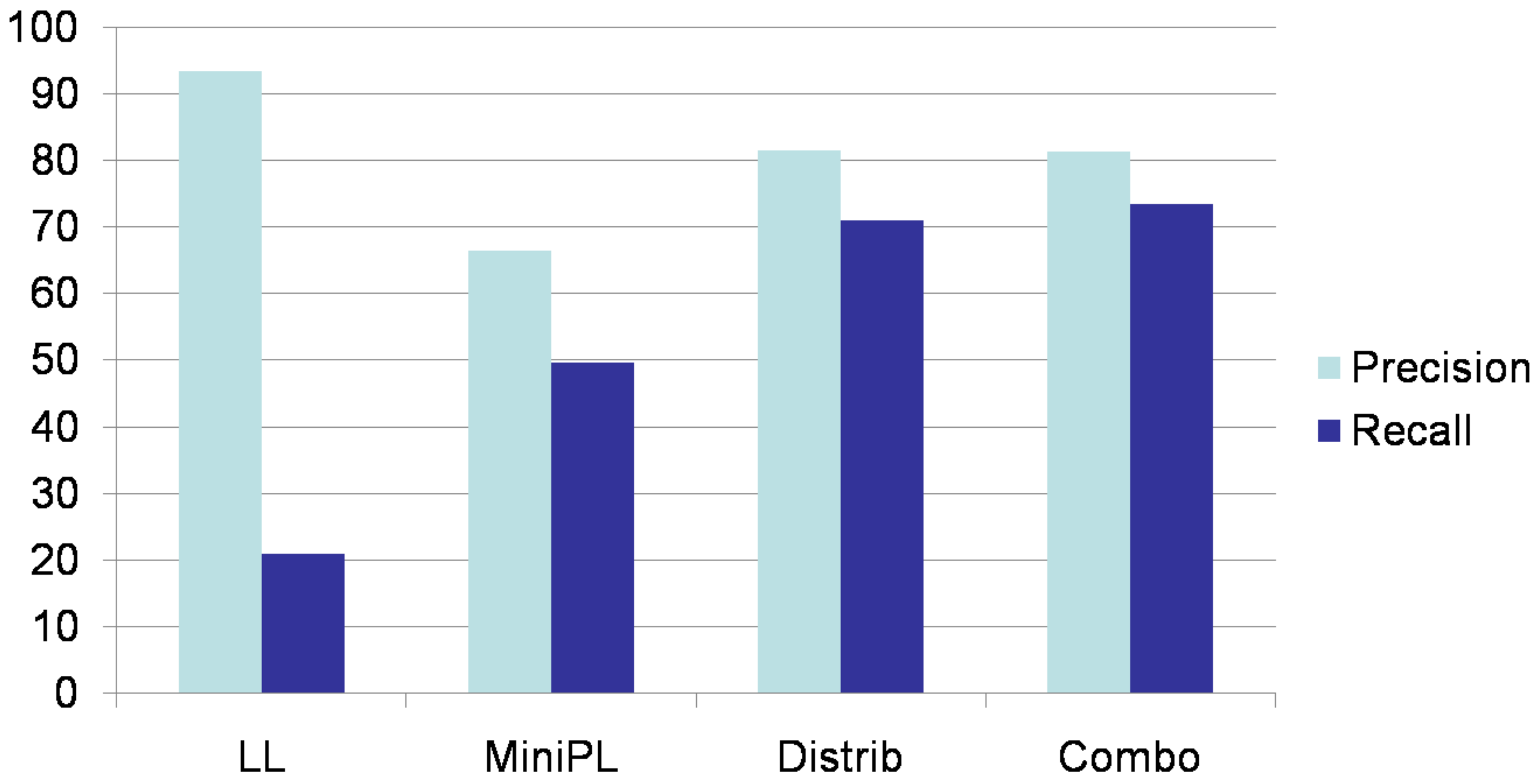
    Percent of true non-referentials that our system labels as positive

Experiments

- Comparison Approaches:

LL	Lappin & Leass's syntactic, rule-based non-referential pronoun detector
MiniPL	Extended version of LL
Distrib	Our distributional approach
Combo	Distrib + extra features for LL and MiniPL decisions

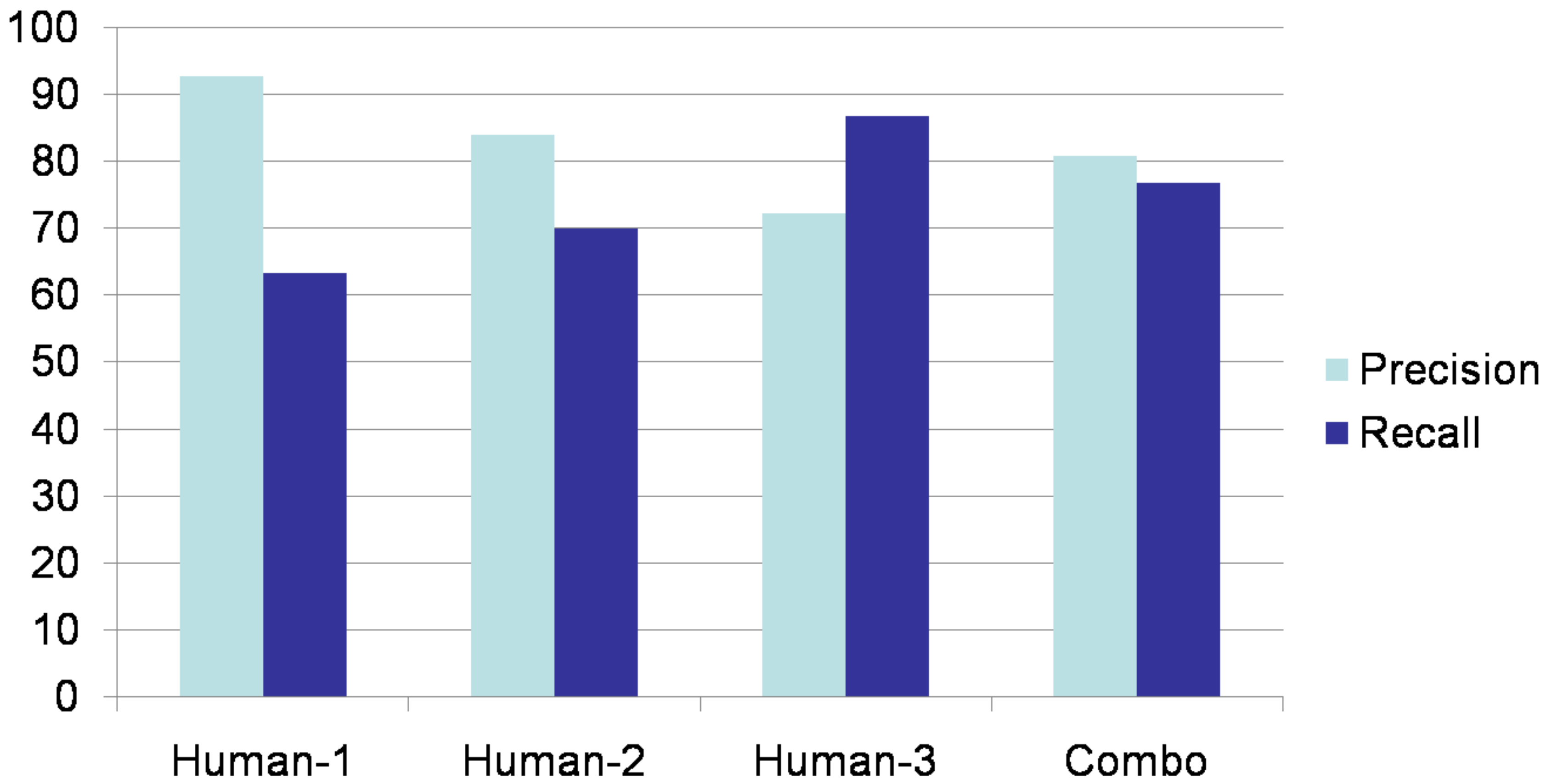
Train/Test Performance (%)



Analysis

- To our knowledge, best results on this task
- Decision ultimately based only on 9-token context window around pronoun
  - no other info from previous discourse
- Question: How well can humans do with the same information?

Human Performance



Error Analysis

- “***It*** takes an astounding amount ...

Error Analysis

- “***It*** takes an astounding amount ... of time to compare very long DNA sequences with each other.”

Error Analysis

- "***It*** takes an astounding amount ... of time to compare very long DNA sequences with each other."
- "... the machine he had installed ***it*** on."

Error Analysis

- "***It*** takes an astounding amount ... of time to compare very long DNA sequences with each other."
- "... the machine he had installed ***it*** on."
  - Need information about preceding candidate nouns. Non-referential detection should ultimately be done jointly with pronoun resolution (cf. Denis and Baldridge (2007))

Conclusion

- Presented simple, distributional technique to classify non-referential pronouns
- Simple to implement, potentially portable to other languages
- State-of-the-art results
- Experimental data is available
- Lots more details and ideas in the paper

ALBERTA INGENUITY

NSERC CRSNG

iCORE

Additional Slides

1. Pro-drop languages
2. Combining Counts
3. N-gram Data

Pro-drop Languages

- Omitted pronouns (zero pronouns) are often non-referential:
  - "Es importante que nosotros…"

    (it is important that we…)
  - "¡Esta lloviendo [hombres]!"
- Need to know which zero pronouns are referential before resolving them:
  - "Es importante para mí."

Combining Counts

- Intuitively, classifier should assign:
  - positive weights to the counts of ***it*** fillers
  - negative weights to the other counts
  - higher weights to counts from more predictive context positions

N-gram Data

- N-gram counts:
  - Google Web 1T 5-gram Corpus Version 1.1
  - Generated from roughly 1 trillion tokens of online text
  - Only includes tokens occurring more than 200 times, n-grams occurring more than 40 times