Information Extraction from Voicemail Transcripts

Martin Jansche and Steven P. Abney

EMNLP 2002

Email vs. Voicemail

```
Date: Wed, 3 Jul 2002 17:41:36 -0400 (EDT)
From: Sam Cauler <sc@erbra.in>
To: Pat Lissner <pl@itu.de>
Subject: Re: now what
```

---

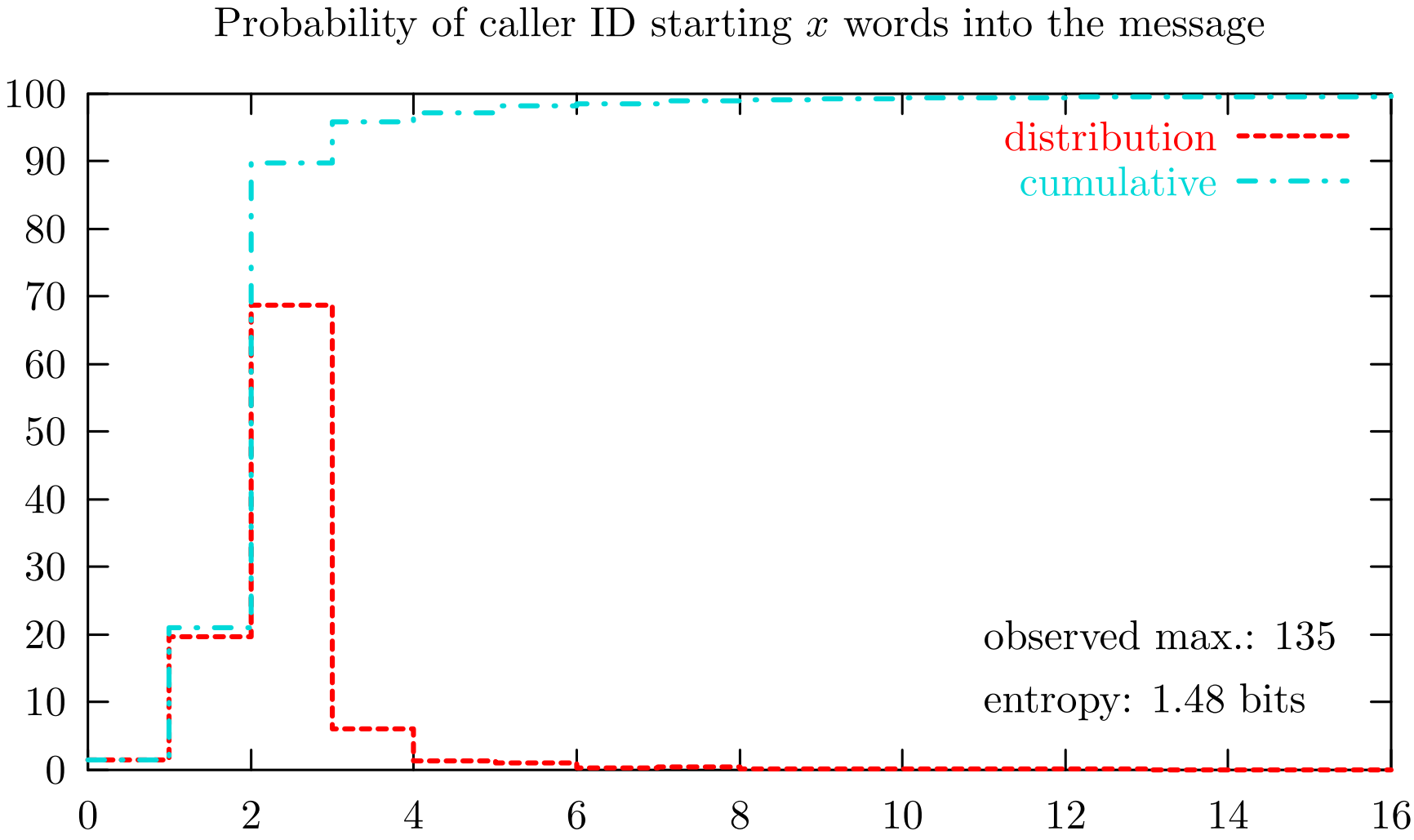
hi Pat this is Sam Cauler I just wanted to . . . so if you could give me a call at one two three four five when you get the message I'd like to chat about it hope things are well with you talk to you soon

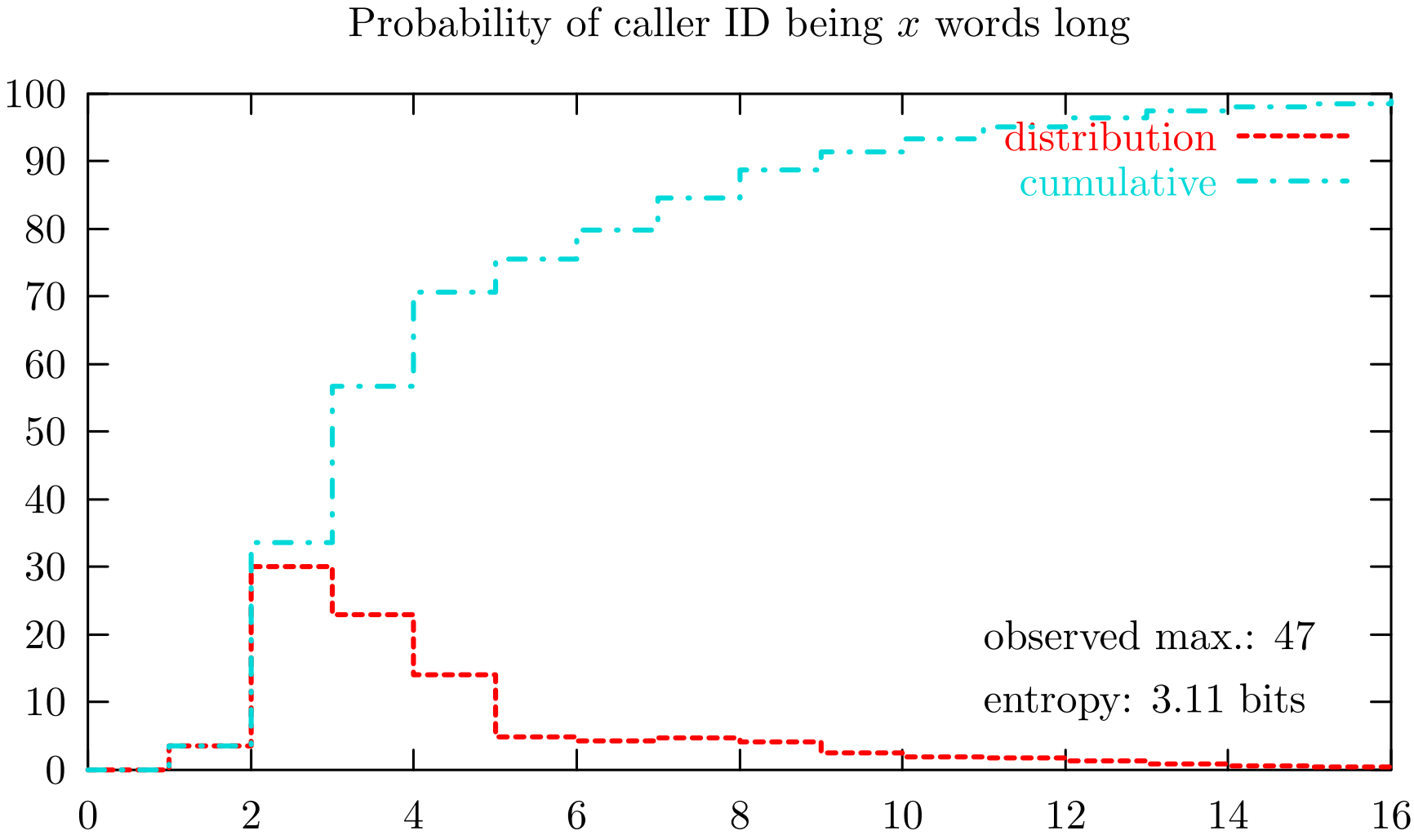
Background and Goals

- SCANMail (Hirschberg et al. 2001): use speech technology to aid browsing, indexing, search, retrieval, etc. of (corporate) voicemail.
- Want to know who called and how to reach them.
- Extract information from voicemail transcripts. Ultimately needs to work with ASR transcripts.
- Comparison with Huang, Zweig & Padmanabhan (ACL 2001, henceforth HZP).

Caller Phrases: Data

- Used manually transcribed and annotated voicemail corpus with approx. 10,000 messages.
- Split 4:1 into development and evaluation sets.
- 8120 messages in training data
- 7686 non-empty (95%)
- 7065 messages have a caller phrase (92% of the non-empty messages)


Probability of caller ID starting x words into the message
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
distribution
cumulative
observed max.: 135
entropy: 1.48 bits


Probability of caller ID being x words long
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
distribution
cumulative
observed max.: 47
entropy: 3.11 bits

Caller Phrases: Approaches

- HZP: tagger based on log-linear models with unigram, bigram and other lexical features.
- Tried to replicate this using Michael Collins' named entity tagger. Similar to (Ratnaparkhi 1996).
- JA: predict caller phrase start and length with classifiers. Feature engineering ensures that we don't rely too much on knowledge of names, to reduce effect of expected recognition errors.

Caller Phrases: Evaluation (1)

Best HZP tagger on IBM dataset vs. Collins' tagger on AT&T dataset (manual transcriptions).

	P	R	F
HZP	89	80	84
Collins	83	78	81

Caller Phrases: Evaluation (2)

F-measure of HZP model ME2-U-f1 (unigram lexical features and number dictionary features) vs. classifier-based extractor described earlier.

	manual xscrpt	ASR xscrpt
HZP	84	19
HZP containment		52
JA containment	71	70

Names are Problematic

- *Frank Ianna* transcribed as *Frank I N A* by ASR.
- *Mehryar (Mohri)* transcribed as *Mary uh*, *Mario*, *Mauri*, etc. by human labelers.
- *John Siskus from Nest* is really *Jon Fiscus from NIST*.

Phone Numbers: Data

- 8120 training messages, 7686 (95%) non-empty
- 5303 phone numbers mentioned (0.7 phone numbers per non-empty message):
  - 4472 (84%) phone numbers are spoken numbers
  - 679 (13%) phone numbers are spoken numbers possibly including *area*, *code*, or *extension*
  - Remaining 152 (3%) made up of corrections, fragments, and questionable markup

Phone Numbers: Approaches

- HZP rules: hand-crafted rules.
- HZP log-linear: tagger based on log-linear models, used with IBM data.
- Again, Collins' tagger based on log-linear models, used with AT&T data.
- Digits (baseline): find all maximal substrings consisting of spoken digit sequences (0 through 9), keep those of length 4, 7, or 10.

JA extract

- Transduce word sequences to digit strings, e.g., *three hundred fourteen ninety nine* to 300-1499.
- Want to get high recall, so try to extract all numbers. Ratio of extracted entities to actual entities approx. 3.2 : 1.
- Huang et al. 2001 report that recall was highest when using hand-written extraction rules.
- But writing high-recall high-precision rules is hard.

JA extract + prune

- Same transducer as before.
- Prune away numbers with less than three digits.
- Adds one false negative on the test set (there was no change on the heldout set), ratio of extracted to actual entities is cut in half, and precision doubles.

JA extract + classify

- Same transducer as before.
- Let a classifier label the extracted numbers to determine whether they are phone numbers.
- Decision is made based on contextual features and the length of the transduced digit string. All other approaches only see the word sequence.

Phone Numbers: Evaluation (1)

	P	R	F
HZP rules	81	83	82
HZP log-linear	90	83	86
Collins	88	93	91
Digits	78	70	74
JA extract	30	96	45
JA extract + prune	59	96	73
JA extract + classify	94	94	94

Phone Numbers: Evaluation (2)

F-measure compared on manual vs. ASR transcripts.

	manual xscrpt	ASR xscrpt
HZP	86	54
HZP containment		82
JA	94	95

Conclusions

- Position relevant for extracting caller phrases. Small inventory of lexical features suffices.
- Length of phone numbers is important. Don't count words, count digits.
- Two-phase approach for phone numbers – transducer with high recall (easy to write by hand), followed by classifier – beats all other approaches, including the previous state of the art.

Acknowledgments

Thanks to Michiel Bacchiani, Michael Collins, Julia Hirschberg, and the SCANMail group at AT&T Labs.