



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Towards Google-like Search on Spoken Documents with Zero Resources

How to get something from nothing in a language that you've never heard of

Speech/EE Meeting

Language/CS Meeting

INTERSPEECH 2010



Towards Spoken Term Discovery At Scale With Zero Resources

Aren Jansen^{1,2}, Kenneth Church^{1,3}, Hynek Hermansky^{1,2}

¹Human Language Technology Center of Excellence,
²Department of Electrical and Computer Engineering, ³Department of Computer Science
Johns Hopkins University, Baltimore, Maryland

aren@jhu.edu, kenneth.church@jhu.edu, hynek@jhu.edu

Abstract

The spoken term discovery task takes speech as input and identifies terms of possible interest. The challenge is to perform this task efficiently on large amounts of speech with zero resources (no training data and no dictionaries), where we must fall back to more basic properties of language. We find that long (~ 1 s) repetitions tend to be contentful phrases (e.g. University of Pennsylvania) and propose an algorithm to search for these long repetitions without first recognizing the speech. To address efficiency concerns, we take advantage of (i) sparse feature representations and (ii) inherent low occurrence frequency of long content terms to achieve orders-of-magnitude speedup relative to the prior art. We frame our evaluation in the context of spoken document information retrieval, and demonstrate our method's competence at identifying repeated terms in conversational telephone speech.

Index Terms: spoken term discovery, zero resource speech recognition, dotplots

1. Introduction

The current stable of large vocabulary speech recognition systems have been developed on the assumption that large orthographically transcribed speech corpora are available for constructing detailed acoustic and language models. As the global

tain words or terms of possible interest. In this pursuit, we are faced with a fundamental computational complication: in n frames of continuous speech, there are $\binom{n}{2}$ possible intervals that can correspond to some word or term in the speech. Thus, we require a heuristic to sort out the potential relevance of these $O(n^2)$ possibilities. Derived from similar efforts in text information retrieval [2, 3], we root our search in the notion that interval length, repetition, and burstiness (inhomogeneity of occurrence frequency) are each strong relevance cues.

The lossiness of speech communication offers two additional cues that can be powerful indicators of interval relevance. First, contentful information must be conveyed clearly and, as a result, repeats of important terms are typically produced with relatively high fidelity. Thus, the more accurate the acoustic match, the more likely it was important. Second, speech communication often occurs over a noisy/lossy channel (e.g. telephone, loud restaurant), often prompting adjacent repeats of terms by the speaker if the listener missed something they deemed relevant from context. With these motivations, we reduce the problem of unsupervised spoken term discovery to a search for long, faithfully repeated intervals of speech.

Our solution to this problem is based on the graphical method of dotplots for comparing sequences, first applied to speech processing by Park and Glass [4] in the form of the segmental dynamic time warping (S-DTW) algorithm. The

NLP on Spoken Documents without ASR

Mark Dredze, Aren Jansen, Glen Coppersmith, Ken Church
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University

mdredze, aren, coppersmith, Kenneth.Church@jhu.edu

Abstract

There is considerable interest in interdisciplinary combinations of automatic speech recognition (ASR), machine learning, natural language processing, text classification and information retrieval. Many of these boxes, especially ASR, are often based on considerable linguistic resources. We would like to be able to process spoken documents with few (if any) resources. Moreover, connecting black boxes in series tends to multiply errors, especially when the key terms are out-of-vocabulary (OOV). The proposed alternative applies text processing directly to the speech without a dependency on ASR. The method finds long (~ 1 sec) repetitions in speech, and clusters them into pseudo-terms (roughly phrases). Document clustering and classification work surprisingly well on pseudo-terms: performance on a Switchboard task approaches a baseline using gold standard manual transcriptions.

1 Introduction

Can we do IR-like tasks without ASR? Information

This approach identifies long, faithfully repeated patterns in the acoustic signal. These acoustic repetitions often correspond to terms useful for information retrieval tasks. Critically, this method does not require a phonetically interpretable acoustic model or knowledge of the target language.

By analyzing a large untranscribed corpus of speech, this discovery procedure identifies a vast number of repeated regions that are subsequently grouped using a simple graph-based clustering method. We call the resulting groups pseudo-terms since they typically represent a single word or phrase spoken at multiple points throughout the corpus. Each pseudo-term takes the place of a word or phrase in bag of terms vector space model of a text document, allowing us to apply standard NLP algorithms. We show that despite the fully automated and noisy method by which the pseudo-terms are created, we can still successfully apply NLP algorithms with performance approaching that achieved with the gold standard manual transcription.

Natural language processing tools can play a key role in understanding text document collections. Given a large collection of text, NLP tools can clas-

Google

Google Search

I'm Feeling Lucky



oooooooooooo

Speak now

Cancel

New! Tour the Amazon in Google Maps with [Street View](#)

Google

images



Peek ahead at image results with new related search previews. [Learn more.](#)



Siri. Beta

Your wish is its command.

Siri on iPhone 4S lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so



Degree of difficulty

- Some queries are harder than others
 - More bits? (Long tail: long/infrequent)
 - Or less bits? (Big fat head: short/frequent)
- Query refinement/Coping Mechanisms:
 - If query doesn't work
 - Should you make it longer or shorter?
- Solitaire → Multi-Player Game:
 - Readers, Writers, Market Makers, etc.
 - Good Guys & Bad Guys: Advertisers & Spammers

Spoken Web Search is Easy

Compared to Dictation

Entropy of Search Logs

- How Big is the Web?
- How Hard is Search?
- With Personalization? With Backoff?

Qiaozhu Mei[†], Kenneth Church[‡]

[†] University of Illinois at Urbana-Champaign

[‡] Microsoft Research

Small

How ~~Big~~ is the Web?

5B? 20B? More? Less?

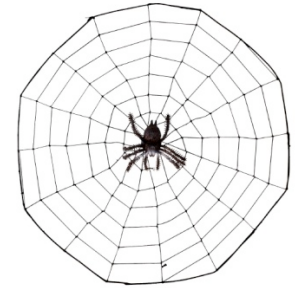
- What if a small cache of millions of pages
 - Could capture much of the value of billions?
- Could a **Big** bet on a cluster in the clouds
 - Turn into a big liability?
- Examples of Big Bets
 - Computer Centers & Clusters
 - Capital (Hardware)
 - Expense (Power)
 - Dev (Mapreduce, GFS, Big Table, etc.)
 - Sales & Marketing >> Production & Distribution
- Goal: Maximize Sales (Not Inventory)



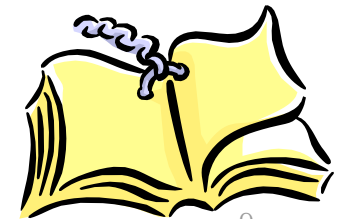
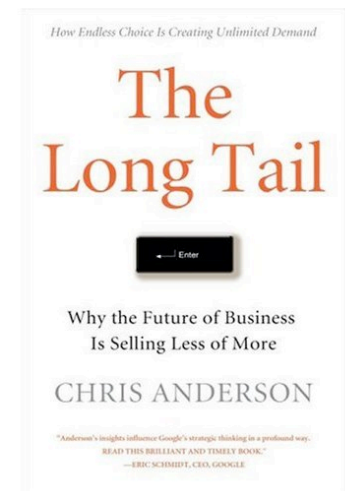
Millions (Not Billions)



Population Bound



- With all the talk about the Long Tail
 - You'd think that the Web was astronomical
 - Carl Sagan: Billions and Billions...
- Lower Distribution \$\$ → Sell Less of More
- But there are limits to this process
 - NetFlix: 55k movies (not even millions)
 - Amazon: 8M products
 - Vanity Searches: Infinite???
 - Personal Home Pages << Phone Book < Population
 - Business Home Pages << Yellow Pages < Population
- Millions, not Billions (until market saturates)

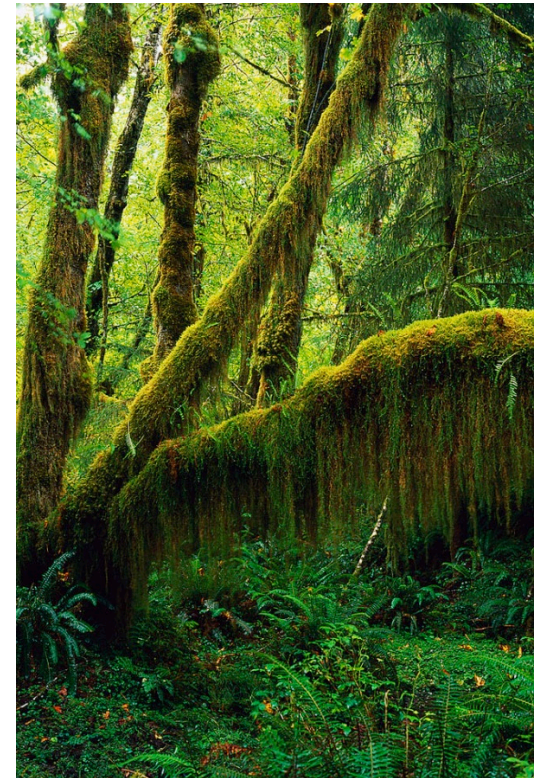


It Will Take Decades to Reach Population Bound

- Most people (and products)
 - don't have a web page (yet)
- Currently, I can find famous people
 - (and academics)
 - but not my neighbors
 - There aren't that many famous people
 - (and academics)...
 - Millions, not billions
 - (for the foreseeable future)

Equilibrium: Supply = Demand

- If there is a page on the web,
 - And no one sees it,
 - Did it make a sound?
- How big is the web?
 - Should we count “silent” pages
 - That don’t make a sound?
- How many products are there?
 - Do we count “silent” flops
 - That no one buys?



Demand Side Accounting

- Consumers have limited time
 - Telephone Usage: 1 hour per line per day
 - TV: 4 hours per day
 - Web: ??? hours per day
- Suppliers will post as many pages as consumers can consume (and no more)
- Size of Web: $O(\text{Consumers})$

How Big is the Web?

- Related questions come up in language
- How big is English?
 - Dictionary Marketing
 - Education (Testing of Vocabulary Size)
 - Psychology
 - Statistics
 - Linguistics
- Two Very Different Answers
 - Chomsky: language is infinite
 - Shannon: 1.25 bits per character

How many words do people know?

What is a word?
Person? Know?

Chomskian Argument: Web is Infinite

- One could write a malicious spider trap
 - <http://successor.aspx?x=0> →
 - <http://successor.aspx?x=1> →
 - <http://successor.aspx?x=2>
- Not just academic exercise
- Web is full of benign examples like
 - <http://calendar.duke.edu/>
 - Infinitely many months
 - Each month has a link to the next

How **Big** is the Web? 5B? 20B? More? Less?



Entropy (H)

- More (Chomsky)
 - <http://successor?x=0>
- Less (Shannon)

MSN Search Log
1 month

Query	21.1
URL	22.1
IP	22.1

Comp Ctr (\$\$\$\$) →
Walk in the Park (\$)

More Practical
Answer

Cluster in Cloud →
Desktop → Flash

Millions
(not Billions)



Dec 2009

Entropy (H)

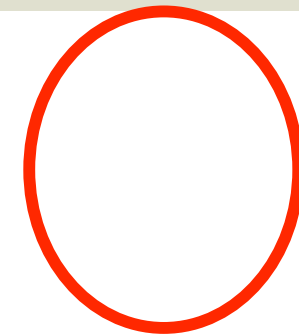
- $H(X) = - \sum_{x \in X} p(x) \log p(x)$
 - Size of search space; difficulty of a task
- $H = 20 \rightarrow$ 1 million items distributed uniformly
- Powerful tool for sizing challenges and opportunities
 - How hard is search?
 - How much does personalization help?

How Hard Is Search? Millions, not Billions

- Traditional Search
 - $H(\text{URL} \mid \text{Query})$
 - 2.8 (= 23.9 – 21.1)
- Personalized Search
 - $H(\text{URL} \mid \text{Query}, \textit{IP})$
 - 1.2 (= 27.2 – 26.0)

Entropy (H)

Query	21.1
URL	22.1
IP	22.1



Personalization
cuts H in Half!



Difficulty of Queries

- Easy queries (low $H(\text{URL} | Q)$):
 - google, yahoo, myspace, ebay, ...
- Hard queries (high $H(\text{URL} | Q)$):
 - dictionary, yellow pages, movies,
 - “what is may day?”

How Hard are Query Suggestions?

The Wild Thing? C* Rice → Condoleezza Rice

- Traditional Suggestions

- $H(\text{Query})$
- 21 bits

- Personalized

- $H(\text{Query} \mid \underline{IP})$
- 5 bits (= 26 – 21)



Personalization
cuts H in Half!

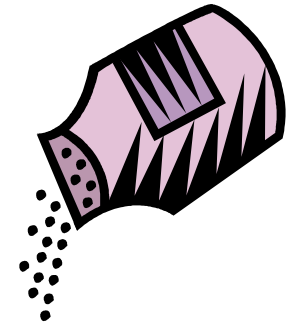
Twice

Entropy (H)

Query	21.1
URL	22.1
IP	22.1
All But IP	23.9
All But URL	26.0
All But Query	27.1
All Three	27.2


Personalization with Backoff

- Ambiguous query: MSG
 - Madison Square Garden
 - Monosodium Glutamate
- Disambiguate based on user's prior clicks
- When we don't have data
 - Backoff to classes of users
- Proof of Concept:
 - Classes defined by IP addresses
- Better:
 - Market Segmentation (Demographics)
 - Collaborative Filtering (Other users who click like me)



Conclusions: Millions (not Billions)

- How Big is the Web?
 - Upper bound: $O(\text{Population})$
 - Not Billions
 - Not Infinite
- Shannon \gg Chomsky
 - How hard is search?
 - Query Suggestions?
 - Personalization?
- Cluster in Cloud (\$\$\$\$) \rightarrow Walk-in-the-Park (\$)
- Goal: Maximize Sales (Not Inventory)



Entropy is a great
hammer

Search Companies have massive resources (logs)

- “Unfair” advantage: logs
- Logs are the crown jewels
- Search companies care so much about data collection that...
 - Toolbar: business case is all about data collection
- The \$64B question:
 - What (if anything) can we do without resources?

toolbar



About 406,000,000 results (0.13 seconds)

Ads for **toolbar**

[Why these ads?](#)

[Install Google **Toolbar** | **toolbar.google.com**](#)

[toolbar.google.com/](#)

Search instantly with Google from any site. Download Today!

[The New Bing™ **Toolbar** - Simple. Powerful. Beautiful.](#)

[www.bingtoolbar.com/](#)

Check Out Other Features Now!

[Google **Toolbar**](#)

[toolbar.google.com/](#)

Take the best of Google everywhere on the web. With a fresh look and new features, Google **Toolbar** is faster, sleeker, and more personalized than ever before.

↳ [Install Google Toolbar](#) - [Google Toolbar](#) - [Toolbar Help](#) - [Features](#)

[Yahoo! **Toolbar** - Stay in touch with your world.](#)

[toolbar.yahoo.com/](#)

With Yahoo! **Toolbar**, you're always in touch with your world. Preview the latest news, emails, weather, and more, right in your **toolbar**! Yahoo! **Toolbar** 2.4 ...

[Toolbar - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Toolbar](#)

In a graphical user interface, on a computer monitor, a **toolbar** is a GUI widget on which

Why Google Toolbar?

There's lots to see and do online. Google Toolbar is designed to help you find what you're looking for quickly and discover new things along the way.



Search with Google anywhere

Google Toolbar lets you search Google from anywhere on the web. Start typing your search and you'll see suggestions for what you might be looking for.



Share your web

Google+ in Toolbar makes it easy to share interesting things from across the web and keep up with the people you care about.



Find things faster

Google Toolbar makes it easy for you to find exactly what you're looking for on any page. Highlight search terms, find specific words, even jump to relevant sections of the page with three easy-to-use tools.



Browse the whole web

Language shouldn't be a barrier to exploring the web. With Google Toolbar, visit a page written in a foreign language and Toolbar will automatically offer to translate it for you.

Google Toolbar is not available for this browser

Requires Internet Explorer 6+
Windows XP SP3/Vista/7+

Permitted Uses

By agreeing to this EULA, your permitted uses and restrictions with respect to the Software are as follows:

- a. Generally: You may use the FreeCause Toolbar to enhance your Internet surfing experience by being provided with relevant references displayed in your browser.
 - b. Search the Internet: During installation, the Software may change the default search engine in your browser's search box, if such a search box exists in your browser. This alteration allows FreeCause to track your searches.
 - c. eCommerce: When visiting a participating merchant's website (via direct type-in, clicking on a PPC ad or an organic search result) with which FreeCause has a business relationship, the Software will prompt you that: (1) this is a participating merchant; and (2) depending on the configuration of the Software with respect to such merchant, the Software will either set cookies on your computer through an affiliate network or it will give you the option of earning contributions/rewards/points etc. on an opt-in basis. (Please see paragraph below for additional information about our use of cookies). In the latter case, if you then decide not to opt in, you will not be entitled to earn any contributions/rewards/points, etc. in connection with your activity. If the cookie is automatically placed on your computer by an affiliate network, or if you do opt-in as set forth above, then you or the cause or organization for which you are earning contributions/rewards/points, etc. will be entitled to earnings or other benefits upon the successful completion of a sales transaction on a commissionable item. Please note, however, that if cookies or other similar tracking devices from other referral partners are already present on your computer, the Software will not overwrite those cookies, but it may still prompt you when you have arrived at a participating merchant's website. In lay terms, if you had previously visited another referral partner's website at some time within a merchant-determined cookie period, that referral partner would gain the commission from being associated with a participating merchant that you now visit. In that case the commission/rewards/points would belong exclusively to the other referral partner. Our use of cookies is more specifically described in our Toolbar Privacy Policy. See Privacy Policy section below.
 - d. Enhanced Experience: When visiting certain sites, the FreeCause Toolbar may add to the page you are visiting certain html code to enhance your experience by adding deals/coupons to links and text, as well as a notification system that will keep you connected with the brand associated with the FreeCause Toolbar that you have installed.
 - e. Updates. When installed on your computer, the Software periodically communicates with our servers. We may update the Software on your computer when a new version is released or when new features are added. These updates occur automatically. We also reserve the right to add features or functions to the Software. Notwithstanding the above, we have
-



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Towards Google-like Search on Spoken Documents with Zero Resources

How to get something from nothing in a language that you've never heard of

Speech/EE Meeting

Language/CS Meeting

INTERSPEECH 2010



Towards Spoken Term Discovery At Scale With Zero Resources

Aren Jansen^{1,2}, Kenneth Church^{1,3}, Hynek Hermansky^{1,2}

¹Human Language Technology Center of Excellence,
²Department of Electrical and Computer Engineering, ³Department of Computer Science
Johns Hopkins University, Baltimore, Maryland

aren@jhu.edu, kenneth.church@jhu.edu, hynek@jhu.edu

Abstract

The spoken term discovery task takes speech as input and identifies terms of possible interest. The challenge is to perform this task efficiently on large amounts of speech with zero resources (no training data and no dictionaries), where we must fall back to more basic properties of language. We find that long (~ 1 s) repetitions tend to be contentful phrases (e.g. University of Pennsylvania) and propose an algorithm to search for these long repetitions without first recognizing the speech. To address efficiency concerns, we take advantage of (i) sparse feature representations and (ii) inherent low occurrence frequency of long content terms to achieve orders-of-magnitude speedup relative to the prior art. We frame our evaluation in the context of spoken document information retrieval, and demonstrate our method's competence at identifying repeated terms in conversational telephone speech.

Index Terms: spoken term discovery, zero resource speech recognition, dotplots

1. Introduction

The current stable of large vocabulary speech recognition systems have been developed on the assumption that large orthographically transcribed speech corpora are available for constructing detailed acoustic and language models. As the global

tain words or terms of possible interest. In this pursuit, we are faced with a fundamental computational complication: in n frames of continuous speech, there are $\binom{n}{2}$ possible intervals that can correspond to some word or term in the speech. Thus, we require a heuristic to sort out the potential relevance of these $O(n^2)$ possibilities. Derived from similar efforts in text information retrieval [2, 3], we root our search in the notion that interval length, repetition, and burstiness (inhomogeneity of occurrence frequency) are each strong relevance cues.

The lossiness of speech communication offers two additional cues that can be powerful indicators of interval relevance. First, contentful information must be conveyed clearly and, as a result, repeats of important terms are typically produced with relatively high fidelity. Thus, the more accurate the acoustic match, the more likely it was important. Second, speech communication often occurs over a noisy/lossy channel (e.g. telephone, loud restaurant), often prompting adjacent repeats of terms by the speaker if the listener missed something they deemed relevant from context. With these motivations, we reduce the problem of unsupervised spoken term discovery to a search for long, faithfully repeated intervals of speech.

Our solution to this problem is based on the graphical method of dotplots for comparing sequences, first applied to speech processing by Park and Glass [4] in the form of the segmental dynamic time warping (S-DTW) algorithm. The

NLP on Spoken Documents without ASR

Mark Dredze, Aren Jansen, Glen Coppersmith, Ken Church
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University

mdredze, aren, coppersmith, Kenneth.Church@jhu.edu

Abstract

There is considerable interest in interdisciplinary combinations of automatic speech recognition (ASR), machine learning, natural language processing, text classification and information retrieval. Many of these boxes, especially ASR, are often based on considerable linguistic resources. We would like to be able to process spoken documents with few (if any) resources. Moreover, connecting black boxes in series tends to multiply errors, especially when the key terms are out-of-vocabulary (OOV). The proposed alternative applies text processing directly to the speech without a dependency on ASR. The method finds long (~ 1 sec) repetitions in speech, and clusters them into pseudo-terms (roughly phrases). Document clustering and classification work surprisingly well on pseudo-terms: performance on a Switchboard task approaches a baseline using gold standard manual transcriptions.

1 Introduction

Can we do IR-like tasks without ASR? Information

This approach identifies long, faithfully repeated patterns in the acoustic signal. These acoustic repetitions often correspond to terms useful for information retrieval tasks. Critically, this method does not require a phonetically interpretable acoustic model or knowledge of the target language.

By analyzing a large untranscribed corpus of speech, this discovery procedure identifies a vast number of repeated regions that are subsequently grouped using a simple graph-based clustering method. We call the resulting groups pseudo-terms since they typically represent a single word or phrase spoken at multiple points throughout the corpus. Each pseudo-term takes the place of a word or phrase in bag of terms vector space model of a text document, allowing us to apply standard NLP algorithms. We show that despite the fully automated and noisy method by which the pseudo-terms are created, we can still successfully apply NLP algorithms with performance approaching that achieved with the gold standard manual transcription.

Natural language processing tools can play a key role in understanding text document collections. Given a large collection of text, NLP tools can clas-

Definitions

- Towards:
 - Not there yet
- Zero Resources:
 - No nothing (no knowledge of language/domain)
 - The next crisis will be where we are least prepared
 - No training data, no dictionaries, no models, no linguistics
- Motivate a New Task: Spoken Term Discovery
 - Spoken Term Detection (Word Spotting): Standard Task
 - Find instances of spoken phrase in spoken document
 - Input: spoken phrase + spoken document
 - Spoken Term Discovery: Non-Standard task
 - Input: spoken document (without spoken phrase)
 - Output: spoken phrases (interesting intervals in document)



The next crisis will be where we are least prepared



[EMNLP 2011 SIXTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION](#)

Featured Translation Task: Translating Haitian Creole Emergency SMS messages

July 30 - 31, 2011
Edinburgh, UK

[\[HOME\]](#) | [\[TRANSLATION TASK\]](#) | [\[FEATURED TRANSLATION TASK\]](#) | [\[SYSTEM COMBINATION TASK\]](#) | [\[EVALUATION TASK\]](#)
[\[BASELINE SYSTEM\]](#) | [\[BASELINE SYSTEM 2\]](#)
[\[SCHEDULE\]](#) | [\[PAPERS\]](#) | [\[AUTHORS\]](#)

The featured translation task of [WMT11](#) is to translate Haitian Creole SMS messages into English. These text messages (SMS) were sent by people in Haiti in the aftermath of the January 2010 earthquake. The messages were sent to an emergency response service and information service called "Mission 4636". They were originally written in Haitian Creole, and were translated into English by a group of volunteers during the disaster response so that first responders (many of whom did not speak Haitian Creole) could understand and act on them. Simultaneously, volunteers were making maps of Haiti and helping to pinpoint the locations described in the messages. More than 30,000 messages were sent to the 4636 number. First responders used the volunteer created translations and maps, and were able to act on the vast majority of requests for help.

Secretary of State Clinton described one success of the Mission 4636 program: "The technology community has set up interactive maps to help us identify needs and target resources. And on Monday, a seven-year-old girl and two women were pulled from the rubble of a collapsed supermarket by an American search-and-rescue team after they sent a text message calling for help." Ushahidi@Tufts described another: "The World Food Program delivered food to an informal camp of 2500 people, having yet to receive food or water, in Diquini to a location that 4636 had identified for them."



RECIPE FOR REVOLUTION



- 2 Cups of a long-standing leader
- 1 Cup of an aggravated population
- 1 Cup of police brutality
- ½ teaspoon of video
- Bake in YouTube for 2-3 months
 - Flambé with Facebook and sprinkle on some Twitter

Egypt's Internet Shut Down, According To Reports

The Huffington Post · Craig Kanalley  

First Posted: 01/27/11 06:33 PM ET · Updated: 05/25/11 07:30 PM ET



React > [Important](#) [Fascinating](#) [Typical](#) [Scary](#) [Outrageous](#) [Amazing](#) [Infuriating](#) [Beautiful](#)

Follow > [+ Egypt](#), [+ Egypt Protests](#), [+ Internet](#), [Egypt Internet Down](#), [Egypt Internet Outage](#), [Egypt Firewall](#), [Internet Down In Egypt](#), [Cnn Egypt Internet](#), [Egypt Internet](#), [Egypt Internet Blackout](#), [Internet Egypt](#), [World News](#)

SHARE THIS STORY

 Recommend  10,482 people recommend this.

3,832

2,063

97

1

 share  tweet  email  +1



Get World Alerts

Sign Up

Reports are emerging that Internet has gone down in Cairo and throughout Egypt, only hours before [the largest planned protests yet](#).

According to [a report from The Arabist](#), "Egypt has shut off the internet."

Multiple Internet Service Providers are affected according to the report, which states:

I just received a call from a friend in Cairo (I won't say who it is now because he's a prominent activist) telling me neither his DSL nor his USB internet service is working. I've just checked with two other friends in different parts of Cairo and their internet is not working either.

The news of the Internet outage came minutes after the Associated Press published [a video of an Egyptian protestor being shot](#).

CNN reporter Ben Wedeman confirmed Internet is down in Cairo and [writes](#), "No internet, no SMS, what is next? Mobile phones and land lines? So much for stability. #Jan25 #Egypt"

Egyptians Connecting To The Internet Via Modem, Fax, Ham Radio

The Huffington Post | First Posted: 01/29/11 03:44 PM ET | Updated: 05/25/11 07:30 PM ET



Follow > [+ Egypt Riots](#), [+ Egypt](#), [+ Egypt Protests](#), [+ Omar Suleiman](#), [Egypt Internet Down](#), [Egypt Internet Outage](#), [Egypt Internet](#), [Egypt Internet Blackout](#), [Egypt Internet Cutoff](#), [Internet Egypt](#), [Slidepollajax](#), [Technology News](#)

*****CLICK HERE FOR COMPLETE EGYPT COVERAGE*****

Despite Egyptian authorities [shutting down access to the internet](#), protesters in Cairo have been able to get online by some creative methods. Check out the slideshow to see how they're doing it.

SHARE THIS STORY

Like 1,520 people like this.

512

400

14

0

share

tweet

email

+1

Find a picture, click the participate button, add a title and upload your picture

Add a Slide

Jan25 Voices

◀ 19 of 44 ▶

🔍 ZOOM

SHARE THIS SLIDE



@Jan25voices

Jan25 Voices

@bastlynn In USA and want to help? Look for feeds in telephone contact with Egy. Ask them to contact us. [#Jan25](#) [#Egypt](#) [#Jan28](#) [#Jan29](#)

Rate This Slide

Rank #16 | Average: 4.6

1 2 3 4 5 6 7 8 9 10

Vote

Current Top 5 Slides



Choose your Top 5 Slides

T-Mobile



Unlimited data plans from

The Zero Resource Case

Zero Resources

No training data, no existing models, and no knowledge of linguistic structure for target language

Zero Resource Term Discovery

Find repeated terms of possible interest in a speech stream without using any models or training data for the target language

How well can you do?



Our zero resource system discovers:

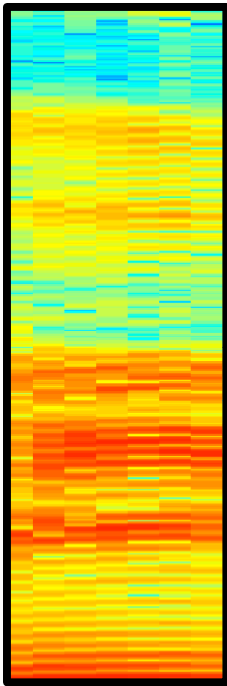
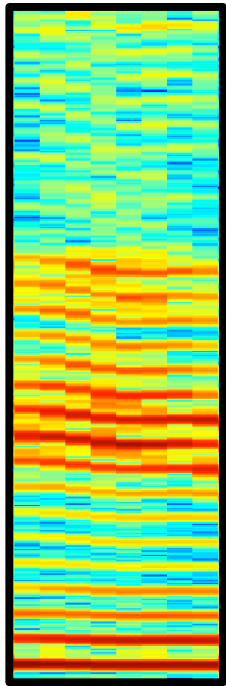


Longer is Easier

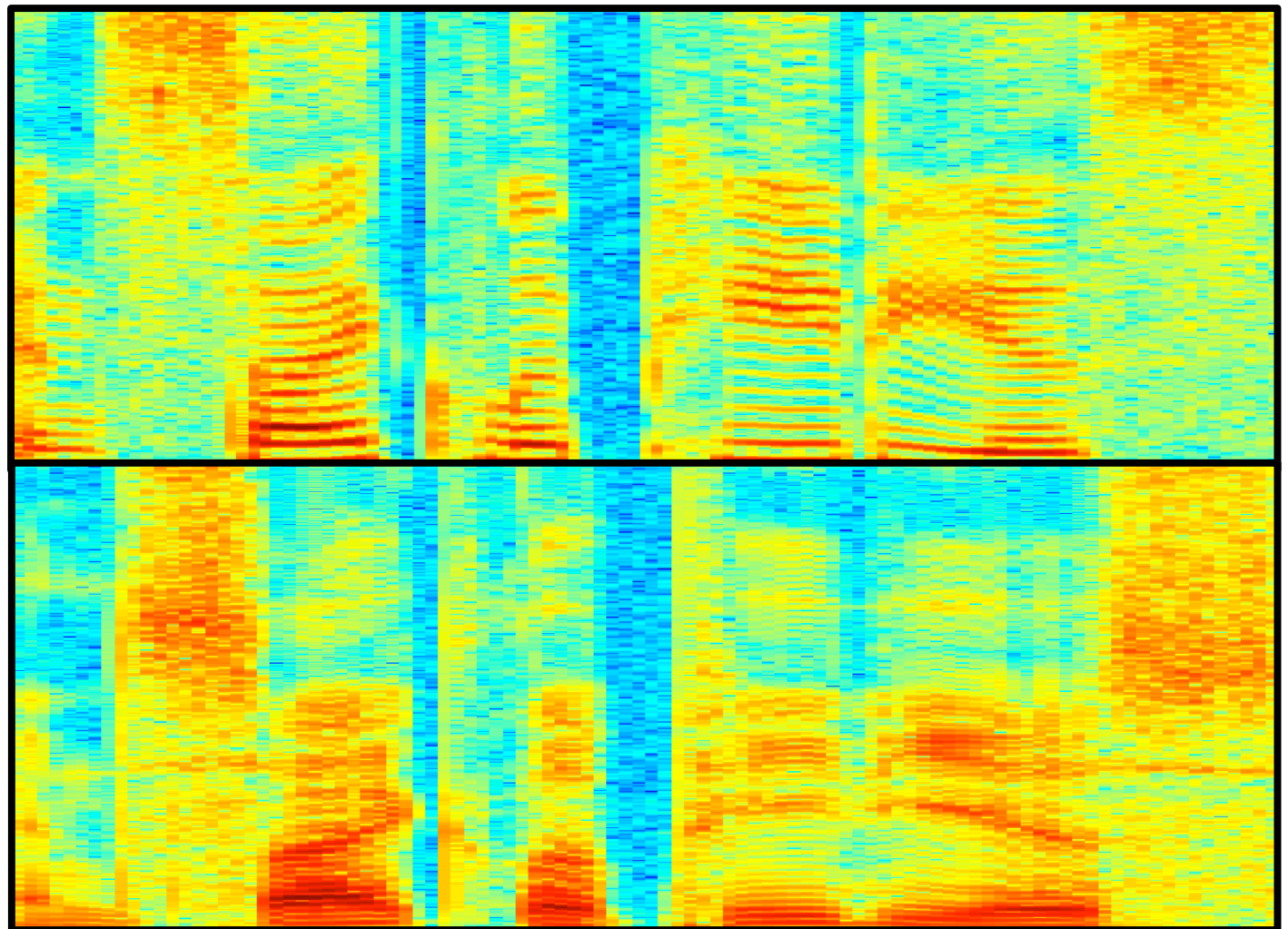
/iy/

female

encyclopedias



male



What makes an interval of speech interesting?



- **Cues from text processing:**
 - Long (~ 1 sec such as “University of Pennsylvania”)
 - Repeated
 - Bursty (tf * IDF)
 - tf: lots of repetitions *within* documents
 - IDF: with relatively few repetitions *across* documents
- **Unique to speech processing:**
 - Given-New:
 - First mention is articulated more carefully than subsequent
 - Dialog between two parties (A & B): multi-player game
 - A: utters an important phrase
 - B: what?
 - A: repeats the important phrase

The Answer to All Questions

Kenneth Church
((almost) former) president

Overview: Taxonomy of Questions

- Easy Questions:
 - How do I improve my performance?
- Google:
 - Q&A: The answer to all questions is a URL
- Siri:
 - Will you marry me?
 - I love you
 - Hello
- Harder questions:
 - What does a professor do?
- Deeper questions (Linguistics & Philosophy):
 - What is the meaning of a lexical item?

Q&A: The Answer to All Questions

- What is the highest point on Earth?
- When was Bill Gates Born?

Siri: “Real World” Q&A

(Not all questions can be answered with a URL)

- Siri
 - Will you marry me?
 - I love you
 - Hello



“ What's the meaning of life ”

I can't answer that now, but give me some time to write a very long play in which nothing happens.



Spoken Web Queries

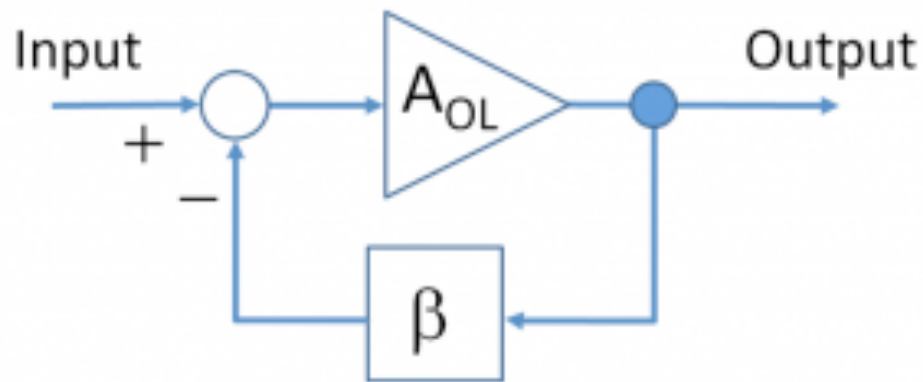
- You can ask any question you want
 - As long as the answer is a URL
 - Or a joke
 - Or an action on the phone (open app, call mom)
- Easy (high freq/low entropy)
 - Wikipedia
 - Google
- Harder (OOVs)
 - Wingardium Leviosa → When Guardian love Llosa
 - Non-speech: Water running → Or (universal acceptor)

Coping Strategies

- When users don't get satisfaction
 - They try various coping strategies
 - Which are rarely effective
- Examples:
 - Repeat the query
 - Over-articulation (+ screaming & clipping)
 - Spell mode

Pseudo-Truth (Self-training)

- Train on ASR Output
 - Negative Feedback Loop (with $\beta = 1$?)
- Obviously, pseudo-truth \neq real truth
 - Especially for universal acceptors
 - If ASR output is “or” (universal acceptor)
 - Pseudo-truth is wrong



Pseudo-Truth & Repetition

- Average WER is about 15%
 - But average case is not typical case
 - Usually right:
 - Easy queries (Wikipedia)
 - Usually wrong:
 - Hard queries (OOVs) and Universal acceptors (“or”)
- Pseudo-truth is more credible when
 - Lots of queries (& clicks) across users (speaker indep)
- Pseudo-truth is less credible when
 - A single user repeats previous query (speaker dependent)

Zero Resource Apps

- Detect repetitions (with DTW)
 - If user repeats the previous query,
 - Pseudo-truth is probably wrong (speaker dependent)
 - If lots of users issue the same query,
 - Pseudo-truth is probably right (speaker independent)
- Need to distinguish “Google” from “Or”
 - Both are common,
 - But most instances of “google” will match one another
 - Unlike “or” (universal acceptor)
- Probably can’t use ASR output to detect repetitions
 - Wingardium Leviosa → lots of diff (wrong) ASR outputs
- Probably can’t use ASR output to distinguish
 - “google” from “or”
- DTW >> ASR output
 - for detecting repetition & improving pseudo-truth

Spoken Web Search \neq Dictation

- Spoken Web Search:
 - Queries are short
 - Big fat head & Long tail (OOVs)
 - Large Cohorts:
 - Lots of “Google,” “or” & “wingardium leviosa”
 - OOVs are not one-offs:
 - Lots of instances of “wingardium leviosa”
 - Small samples are good enough for
 - Labeling (via Mechanical Turk?)
 - Calibration (via Mechanical Turk?)

Spelling Correction Products: Microsoft Office ≠ Bing

- Microsoft Office
 - A dictionary of general vocabulary is essential
- Bing (Web Search)
 - A dictionary is essentially useless
- General vocabulary is more important in documents than web queries
- Pre-Web Survey: Kukich (1992)
 - Cucerzan and Brill (2004) propose an iterative process that is more appropriate for web queries.

Spelling Correction: Bing ≠ Office

Google

Search

wingardium leviosa

wingardium leviosa

wingard & company inc

wingard

wingaria

Web

Images

Maps

Videos

News

Shopping

More

Baltimore, MD

Change location

Show search tools

[List of spells in Harry Potter - Wikipedia](http://en.wikipedia.org/wiki/List_of_spells_in_Harry_Potter)
en.wikipedia.org/wiki/List_of_spells_in_Harry_Potter
Spells in Harry Potter occur in the fictional universe created by author J. K. Rowling. Magic spells are

[Urban Dictionary: Wingardium Leviosa](http://www.urbandictionary.com/define.php?term=Wingardium_Leviosa)
www.urbandictionary.com/define.php?term=Wingardium_Leviosa
Nov 18, 2010 – A fictional spell from the Harry Potter series that causes objects to hover in midair; The Levitation Spell

[Hover Charm - Harry Potter Wikia](http://harrypotter.wikia.com/wiki/Hover_Charm)
harrypotter.wikia.com/wiki/Hover_Charm
The Hover Charm (**Wingardium Leviosa**) is taught to first years at Hogwarts School of Witchcraft and Wizardry.

[Wingardium Leviosa - YouTube](http://www.youtube.com/watch?v=nAQBzjE-kvI)
www.youtube.com/watch?v=nAQBzjE-kvI
Apr 12, 2006 - 3:39
Ron and Hermione attract... Adorable...

[Harry Potter Magic Spell - Wingardium Leviosa - YouTube](http://www.youtube.com/watch?v=nAQBzjE-kvI)
www.youtube.com/watch?v=nAQBzjE-kvI
Apr 16, 2009 - 2 min - Uploaded by MarkArcana

Spoken Web Search ≠ Dictation

- Spoken Web Search:
 - Queries are short
 - Big fat head & Long tail (OOVs)
 - Large Cohorts:
 - Lots of “Google,” “or” & “wingardium leviosa”
 - OOVs are not one-offs:
 - Lots of instances of “wingardium leviosa”
 - Small samples are good enough for
 - Labeling (via Mechanical Turk?)
 - Calibration (via Mechanical Turk?)

albert einstein	4834
albert einstien	525
albert einstine	149
albert einsten	27
albert einsteins	25
albert einstain	11
albert einstin	10
albert eintein	9
albeart einstein	6

Cucerzan and Brill (2004)

- Context is helpful
- Easier to correct
 - MWEs (names)
 - than words

Misspelled query:

anol scwartegger

First iteration:

arnold schwartnegger

Second iteration:

arnold schwarznegger

Third iteration:

arnold schwarzenegger

Fourth iteration:

no further correction

ASR without a Dictionary

- More like did-you-mean
 - Than spelling correction for Microsoft Office
- Use Zero Resource Methods to cluster OOVs
 - Label/Calibrate Small Samples with Mechanical Turk
- Wisdom of the crowds
 - Crowds: smarter than all the lexicographers in the world
- How many words/MWEs do people know?
 - 20k? 400k? 1M? 13M?
 - Is there a bound or does V (vocab) increase with N (corpus)?
 - Is V a measure of intelligence or just experience?
 - Is Google smarter than we are?
 - Or just more experienced?

Conclusions:

Opportunities for Zero Resource Methods

- Bing \neq Office
 - Spoken Web Queries \neq Dictation
- Pseudo-Truth \neq Truth ($\beta \ll 1$)
 - Use Zero-Resource Methods to distinguish “Google” from “Or” (tie β s)
 - Google: same audio
 - Or: different audio
- Assign confidence to Pseudo-Truth
 - More credible: Repetition *Across* Users (with clicks)
 - Less credible: Repetition *Within* Session (without clicks)
 - (Ineffective) Coping Strategy:
 - Ask the same question again (and again)
- OOVs: Like Did-you-mean spelling correction
 - Use zero-resource methods to find lots of instances of same thing
 - Label/Calibrate small samples with Mechanical Turk

