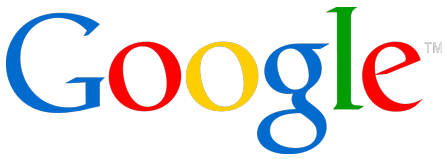

Voice Search at Google

Kyoto University, 4/2/2012

Mike Schuster
Research Scientist

Agenda

- Speech at Google: Some Applications
  - Voice Search
    - Search by voice using your phone
    - Use dictation for E-Mail, Chat etc.
    - Use on maps for directions
    - Use with Chrome!
  - Transcription & Audio Indexing
    - transcribe (and search) spoken content (YouTube!)
  - Combine with translation

Speech recognition: Looking back

Google

- Speech recognition 30 years ago
  - Speaker-dependent
  - Isolated phones or words (very small vocabulary)
  - Clean speech
  - Not real-time
- Speech recognition today, after 1000s of man-years of work
  - Speaker-independent
  - Large vocabulary (1M for VoiceSearch)
  - Continuous speech
  - Noisy-telephone channels
  - Real-time
- Biggest reason for improvements?
  - (More) learning from data
  - Faster machines!

Voice Search

Mobile search made easy

Voice Search

Google

- Speech-enabled Google Search
  - Free-form queries just like the search box on google.com
  - Smartphone clients
    - Android, iPhone/iPod/iPad, Blackberry, Nokia s60
  - Languages as of 3/2012
    - 2008: US English
    - 2009: UK English, Australian English, Indian English, Mandarin, Japanese
    - 2010: French, Italian, German, Spanish, Korean, Taiwanese, Dutch, Czech, Russian, Polish, Turkish, Brazilian Portuguese, South African English, Zulu
    - 2011: Cantonese, various Spanish & Arabic dialects, Hebrew, Pig Latin! - and lots of improvements
- Examples
  - “Used Honda Civic craigslist Los Angeles”, “German beer”
  - “高の原から国際会館” , " ドイツのビール”
  - ” 부산 비빔밥” , “ 독일 맥주”
  - “Romantische Strasse”, “Deutsches Bier”
- Videos (click on keywords to launch video)
  - English, Voice Actions, Chrome
  - Korean, Korean viral marketing

Technical challenges before launch (server)

Google

- Acoustic Model
  - For English used initially GOOG-411 data
  - In-house data collection to have a test set to measure performance
  - Good enough first model took months (year?) of experiments
  - Other languages much faster as development process was streamlined, now a few weeks per language including data collection
- Language Model
  - Assumption: People say what they type (not true but close enough)
  - Let's use anonymized google.com queries to estimate LM
  - 700M unique words in queries (in English), years of queries total
  - Took months of effort to get a good enough LM (low perplexity)
- Dictionary & Pronunciations
  - 1M words vocabulary and pronunciations
  - 95% of prons automatically generated, difficult

More details (server)

Google

- Acoustic Model
  - HMM system, decision trees, 3-10k states, up to 300k Gaussians, LDA, STC, ML & MMI training
  - 1000s of hours of acoustic data used for training (English, Japanese, ...)
  - One training run takes days on 1000s of machines
- Language Model
  - 3-grams to 6-grams, entropy-pruned Katz back-off models
  - Massive amounts of data in, massive LM out (before pruning 100GB?)
  - Lots of junk, misspelled words, porn (even after filtering)
  - Vocabulary/textnorm improvements requiring iterative approach
- Search
  - Time-synchronous FST-based beam search
- Dictionary & Pronunciations
  - Lots of high-frequency exceptions (mp3, metropcs, 007, numbers, ...)
  - Depending on the language can be quite time-consuming

Even more details (server)

Google

- Segmentation
  - Mandarin, Cantonese, Japanese
    - no marked word boundaries and no spaces
  - Korean
    - few spaces compared to other languages
  - However, people use spaces in search queries to separate keywords, but not always consistently
    - “ncis 시즌 6”, “ncis 시즌 6”
    - ” 京都駅 ラーメン”
    - ASCII words are always separated by spaces
- WordPieceModel Segmenter
  - Wrote segmenter that maximizes LM likelihood while finding the best word definitions
    - Language independent
    - No OOVs
  - Separate statistical algorithm to predict where to put spaces as part of decoding
  - (as described in “Japanese & Korean Voice Search”, ICASSP 2012)

And more details (server)

Google

- Encoding & writing systems
  - Google uses Unicode everywhere
  - Korean queries have 2 (3) different writing systems mixed, Japan has 4!
    - Korea: 한글 , ASCII, (漢字)
      - “ 자바 api”
    - Japan: 漢字 , ひらがな , カタカナ , ASCII
      - “ 京都 安い テレビ sony bravia”, “ 価格 .com”,
  - Some words are written in multiple ways depending on user or context (difficult for transcription & scoring!)
    - Korea
      - " 자바 api” or “java api”
      - " 스포츠” , or “sport”
    - Japan
      - “nintendo” or “ ニンテンドー” or " 任天堂” or “ にんてんどう”
  - Many Japanese Kanji have multiple pronunciations and depend heavily on context
    - " 一人 " (hitori) 、 " 一人前” (ichininmae)
    - " 二十歳 " -> ? (nijussai, hatachi)
    - “AKB 48”, “ 貴社の記者が汽車で帰社した”
  - Every language has its unique problems, lots depends on nasty details

Text Normalization and Denormalization

Google

- Convert typed query to spoken for training (several ways possible):
  - xbox 360 8mb
  - “x box three sixty eight megabytes” (or 'three hundred sixty'?)
- And back at decoding time:
  - “x box three sixty eight megabytes”
  - xbox 360 8mb (not 'xbox 368 mb!)
- Example of local listings ambiguity
  - dr smith on st mark st @ acorn dr
  - “doctor smith on saint mark street at acorn drive”
- Spoken URLs (de)compounding
  - “cancer centers of america dot com”
  - cancercentersofamerica.com
- Current question
  - spoken system + rules for display OR
  - written system without (or less) rules

Model in spoken or written domain?

Google

- Written
  - These 17 NBA players had their best year in 2011.
- Spoken
  - these seventeen n b a players had their best year in twenty eleven period
- Output, display & scoring
  - Clearly written domain desired
- Transcriptions
  - Spoken sounds easy but is not for most transcribers (ambiguous)
  - AM database transcriptions often in spoken domain
- LM data
  - Lots available in written domain, almost none in spoken
- Dictionary
  - Easier in written domain, but more pronunciations (example: “ms”)
- We currently use both both depending on language
  - Not ideal but performance & development history is what drives this

Infrastructure

Google

- Hardware
  - Using 1000s of machines in data centers across the world for development
- Software
  - Using lots of existing Google infrastructure (MapReduce, BigTable, etc.)
  - Everything developed in house (all speech-related SW by our group)
    - Writing/optimizing all training/server/client software took years
    - Mostly C++, but also Java & Python
  - As many people involved, everything changes constantly
- Process
  - Training/evaluation process optimized to process large amounts of data quickly
  - Process constantly optimized with everybody's input
- Complexity is a problem
  - Hard to avoid, speech was complex everywhere I have been (University, ATR, Nuance, NTT, Google...)

Quality

Google

- How to measure quality for Voice Search?
  - Did user get relevant webpage? Hard to measure!
  - What is displayed on the screen as “recognized results” should make sense but more important is the final web result
    - Example: “slumdog millionaire”
    - Speech Recognition Result: “small dog millionaire” (wrong!)
    - Web Result: relevant webpage is displayed regardless of speech rec errors
  - Approximation: WebScore measure
    - Feed the hyp and ref to google.com and compare the URLs
    - Is in most cases a good measure for “relevant webpage found”
    - Fixes some ambiguous queries like “google” vs “구글” or “クーグル”
- Launch criteria (accuracy & speed)
  - WebScore > 50% measured on offline test set, dictation usable
    - Now for many languages much higher
    - These numbers don't make sense for anybody but us
    - Word error rates depend on word definition → once you get used to a certain range for a language WebScore metric less important, WER sufficient
    - Dictation: no WebScore
  - Speed acceptable for user (wait time average < ~2 sec)

Challenges after launch

Google

- 99.9999% Up-time goal
  - Used to have 24/7 rotation-based pager duty, now more traffic → we are getting help from SREs (site reliability engineers)
  - Production system very complex spanning multiple continents
- Quality
  - No simple real-time measure yet
    - Good to look at online statistics (click-through-rate, traffic etc.)
    - WER is still the best overall (don't compare across languages!)
  - LM needs improvement, retraining (freshness an issue)
  - AM needs retraining, using latest acoustic data from application
    - Critical to eliminate mismatch of channel, user distribution, spoken content
  - Dictionary
    - Pronunciations fixed/added/removed
- Speed
  - Minimize streaming client-server transmission times and data loss
    - often out of our control (carrier network issues)
  - Optimize UI & data flow for improved usability
  - Optimize server parameters for latency

Other interesting facts

Google

- Global usage
  - Has increased by factor of six since a year ago
  - Voice search traffic: #1 (US), #2, #3 (Japan and Korea)
- Usage patterns
  - Voice Search usage is highest on evenings and weekends and on wireless
    - Probably at home?
  - TV/radio shows mentioning Voice Search result in big traffic spikes
- Human transcriptions
  - Recognizer often better than human transcribers
    - Humans often don't know the specific environment of the speaker and therefore often don't know how to spell specific place names, product names etc.
    - We now train large parts of the system without human transcriptions
- Audio input quality
  - Almost every phone we have seen had initially problems with the audio input
  - Google has guidelines how to avoid these problems and can help identify them -> contact us if you work on a new phone!

What's next besides quality improvements?

Google

- More languages
  - Google has presence in many countries, voice search will come, usage will increase
- More user-visible features
  - Contact search (already launched in US)
    - Name recognition difficult
  - Speech input for every text box, dictation
  - Navigation (already launched in US)
  - Other voice actions
  - Define interface that can be used by outside apps
    - Voice Actions
    - Android intent/recognizer API, HTTP interface?
  - Combine with translation → launched by translate team
- More phones/clients
  - Voice Search is strong on Smartphones (Android, iPhone)
  - Many more Android phones/devices coming
  - Many will have voice search on home screen

What's next besides quality improvements?

Google

- More channels
  - Google recently launched voice search on desktop within Chrome browser
    - currently only for English
- More invisible quality features
  - Personalization (launched in US)
    - Uses user-specific speech to adapt etc. to improve quality for each user separately
    - Difficult as per-user statistics hard to keep track of, privacy needs to be guaranteed
  - Dictation (in G-mail etc.)
    - Big quality improvements
  - Better integration within Android
    - Some features hard to discover, hard to use because of user interface
    - Very big factor in adoption of using speech on the phone
  - Make more reliable connection
- Dialog
  - Too difficult for now, likely later than earlier

What's the secret?

Google

- No secrets!
  - Google Speech has published quite a few papers on Voice Search & Dictation
  - all combined describe quite well what we did
- But if I had to pick three things...
  - Experience
    - Speech is not so complicated if you know what you are doing
    - With a small number of experienced people you can do a lot
  - Diligence
    - Voice search didn't work very well in the beginning (70% error???)
    - Constantly improving the system is a must (AM & LM retraining, dictionary fixes, new techniques)
  - Simplicity
    - Complexity is the biggest enemy
    - Many new “features” not worth it in the longer run
    - Avoid rules, learn from data
    - Measure and improve

Google Audio Indexing/Transcription

Large-scale transcription and indexing

YouTube Transcription

Google

- Attach captions to all English YouTube videos where it is useful
  - Launched November 2009
  - Some great examples:
    - Speech by Eric Schmidt
    - AutoCaps Demo Video
    - Steve Jobs on CNBC
- Hard problem
  - Lots of noise, music, home made videos, talk shows, ...
  - Lots of videos not English! (language detection also difficult)
  - Doesn't need to be real-time as off-line recognition
  - Not always perfect, but very useful to users as is
- Obviously useful for many languages
  - In languages other than English even harder, launched for Japanese recently
  - Even more useful with translation attached

Voice Mail Transcription

Google

- Only in the US lots of people use voice mail
  - Speech group built a system that automatically transcribes voice mail and sends the transcript to you in E-Mail
  - Part of “Google Voice” in the US
  - Quite useful despite imperfect recognition!
- Very hard problem
  - Extremely noisy channels (mostly cell phone)
  - People don't speak clearly
  - Lots of personal words used (individual's names etc., often only easy to understand for the person getting the message)
  - Doesn't need to be real-time as off-line recognition
  - Not perfect, quality needs to improve

Summary

Google

- Speech recognition for mobile devices
  - Long term predictions same as last year and the year before...
  - Has had many ups and downs in the past (remember year 2000?)
    - Speech is back!
  - Is not going to replace the keyboard completely, it's just an additional input method
  - Will become faster, more accurate and more reliable (better networks!)
  - Is not going away as it's just too useful (typical examples)
    - “ウォールナット 一枚板 京都” → good website
    - “USJ 入場券” → good website
    - “京都市左京区浄土寺真如町” → Map
    - Total time from start of speech to useful result: **5 sec!**
    - Try to beat that with any other input method...
- Speech recognition in combination with translation
  - Will help to bridge language barriers
- Overall on the technology
  - Lots is obvious now but it was a long way to get there
  - We want the phone to understand you like your brother

Q&A

Google

Thank you! Questions, please...

schuster@google.com