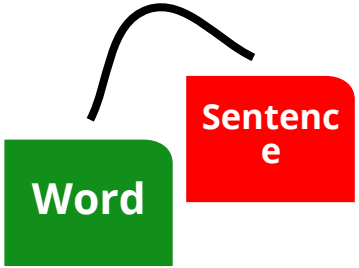
Language Modeling

Language Modeling (LM) Technology

State-of-the-art technology: n-gram models

- Consider large corpus of typed sentences
- Count occurrences of word sequences (n-grams)
- Estimate P(W_i | W_i-1, W_i-2, …)
- Decode utterance with high-order n-grams; back-off to lower-order ngram when high-order n-gram not in model (Katz smoothing)



Language Modeling

Built from Google Search logs

> 200B word tokens

> 10M word vocabulary

Built from Voice Search logs

Learning natural spoken input

Built at scale

1 - 12B n-grams

Built usings 100s years of CPU

Large-Scale Language Modeling

Language model built in a distributed fashion

Text is divided into ‘shards’ (batches) and n-gram counts found for each shard on different machines

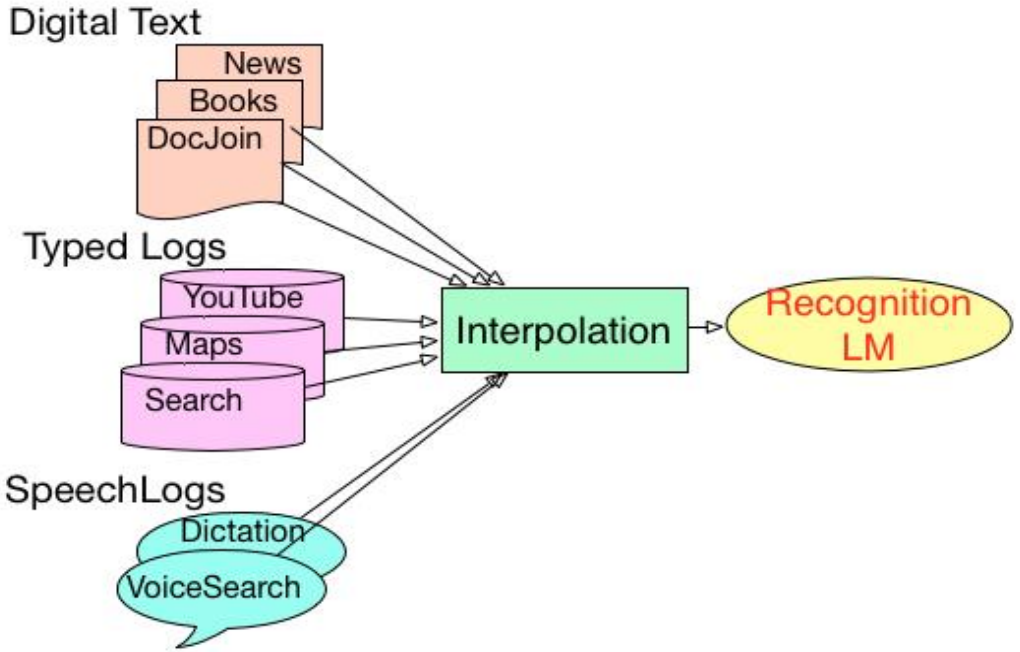
Counts are shuffled to merge n-grams across different shards and bring n-grams with the similar contexts to the same machine to facilitate model estimation

Resulting model can be (entropy) pruned to fit on a single machine or served in a distributed fashion.

Language Modeling: Google-Style

Use a variety of data sources, billions of sentences.

Include recognition results from previous user interactions with the system.

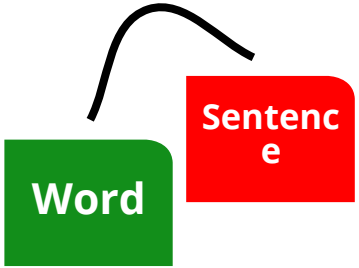


Unsupervised Training

- Speech logs are the most helpful training material for speech LMs
  - best matched to what users will speak next.
- Can't have them transcribed (too much data) → use unsupervised training.

Challenge: Unsupervised Training can lead to feedback loops that boost the probabilities of undesired items in the models:

- misspellings, e.g. "probly" or "prolly" for "probably"
- random words hypothesized over noise
  - Korean: "keu-a"
  - British English: "kdkdkdkdkdkdkdkd"
  - Dutch: "fuck"



Language Model - Challenges

Automatic Capitalization

weather in Scarsdale New York

Automatic / Spoken punctuation

how old is Barack Obama?

Verbalization

2012 - twenty twelve, two thousand twelve, two zero one two

Contextual & Personal

Location, Dialog state, Application, Contacts ….

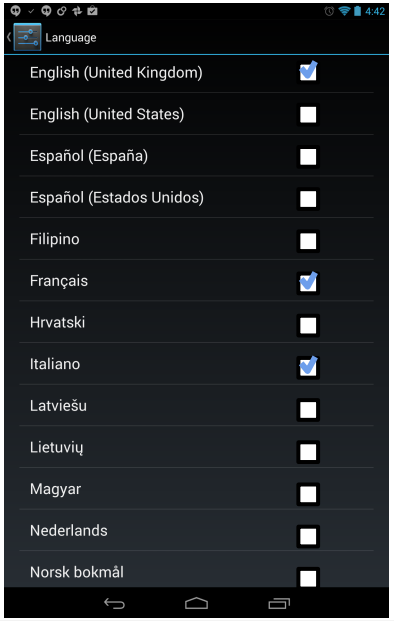
Modeling i18n Linguistic Phenomena

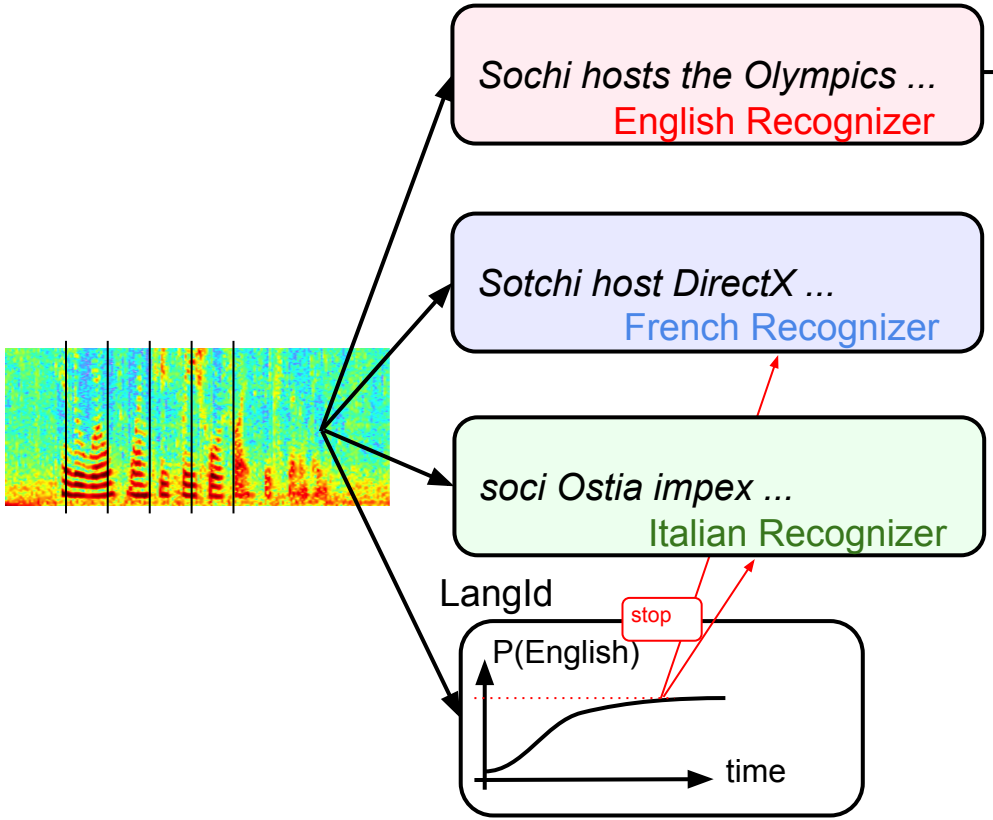
- Inflections in Russian
  позвони Джону (“call John”, verb requires dative)
  набери Джона (“dial John”, verb requires accusative)

- Liaisons, hyphens, contractions in French
  recognized keskispass for qu’est-ce qui se passe (~ “watshapning”)

- Suffixes and white spaces in Korean
  서울시_장애인_복지 (“benefits for the disabled in Seoul”), was rendered as
  서울_시장_애인_복지 (“benefits for Seoul mayor's girlfriend”)
- ...

Multilingual Recognition





Language Identification

40 Filterbank e.

↓

Frame Stacker (-20, +5)

↓

VAD Segmenter

↓

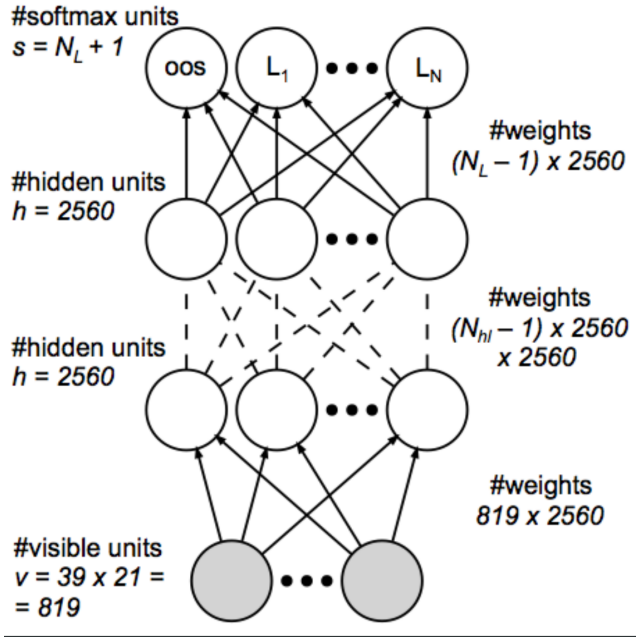
Mean / Std Norm.

↓

DNN

↓

Log Posterior Average



- Topology:
  - **8 hidden lyr**: 2560 nodes.
  - SoftMax lyr: 34 lang classes.
  - Rectified linear units.
- Training:
  - No regularization.
  - Asynchronous Stochastic Gradient Descent.
  - 400K steps (5d in 200 workers).

J. Gonzalez-Dominguez, et al. “A Real-Time End-to-End Multilingual Speech Recognition Architecture”. *IEEE Journal of Selected Topics in Signal Processing.* To appear.

Data Operations

Collection, Annotation, Transcription

What type of data do we collect

Speech systems rely on tons of data to develop and evaluate its models. When that data requires human intervention, Data Operations gets involved.

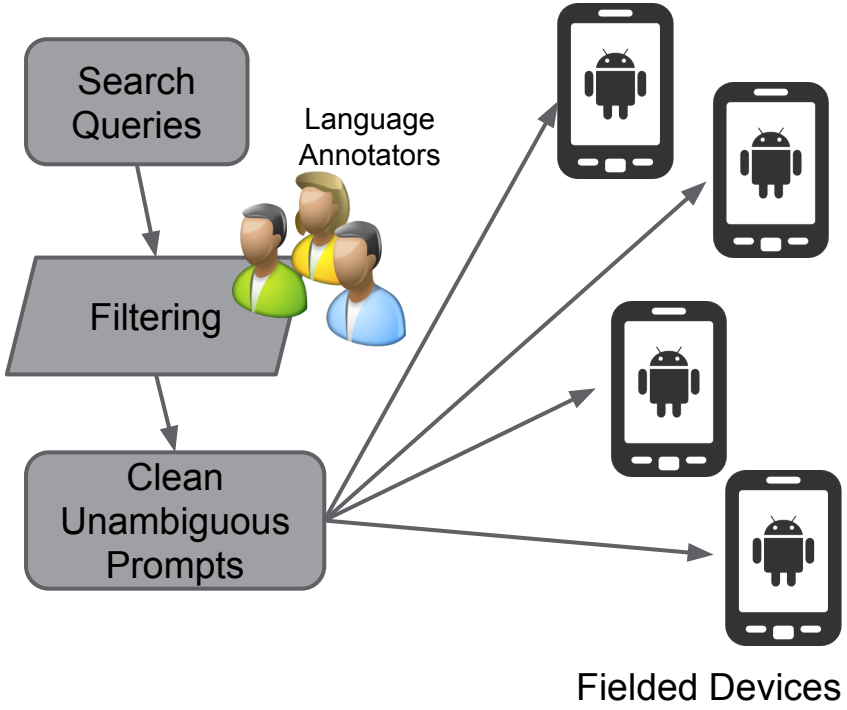
3 kinds of data:

Audio	• to bootstrap AMs for Voice Search, Hotwords, devices, cars
Transcription	• test sets for ASR evaluation • supervised training sets for ASR development
Linguistic	• phonetic lexicons for ASR and TTS • ASR error analysis • annotations about accents, voice actions, etc. • text norm grammars

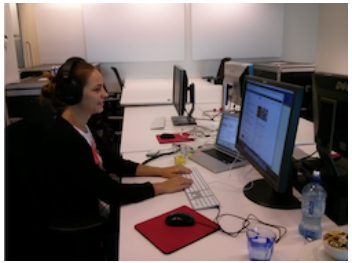
DataHound and Data Collection in the Field

Most language projects begin with recordings done using Android phones in the field. This initial supervised data is used to build our initial deployment model.

Customized **DataHound** Android Apps for Phone, Glass, and Tablets are used to capture the special audio characteristics of these devices.



How we do it



For audio

- In-field collections - in country, in the right environments using DataHound. Volunteers can also contribute audio remotely from their own devices

For transcription and linguistic data

- Transcribers and linguists use AppsEngine application, PeraPera to submit text annotations

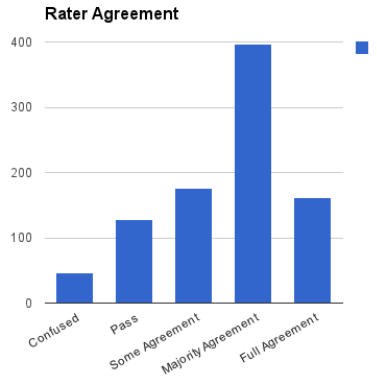
Things to consider

- local labor laws, internet connectivity, cultural expectations, written language resources, i18n capabilities (font display, segmentation)

Metrics

How to measure a speech system

- Traditional metric is Word Error (WER)
- → Count #Substitution #Insertions #Deletions



Transcribe test data

- Expensive, slow
- Measure recognizer vs human
- Problems human's don't agree ⇒ ⇒ ⇒

Metrics

Long tail problem

- Head of distribution easier to recognize (i.e not so useful to test on)
- Manual test sets don’t cover long tail

Alternative testing strategies

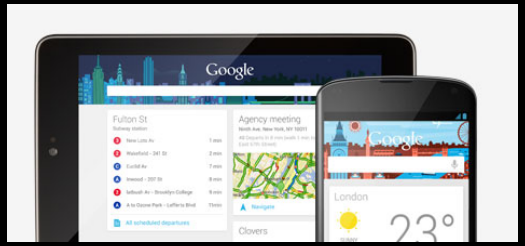
- Side x Side testing
  - Track differences between system A / B
- Live experiments
  - Track live metrics: Click through rates, Correction rates, Retention rates

On device recognition

Making it small, very small

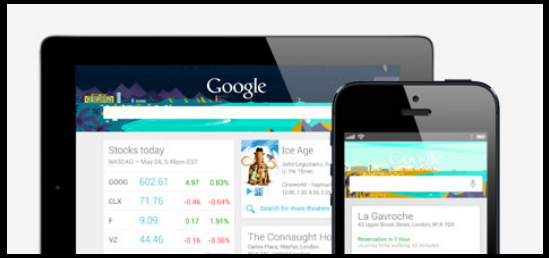
Google

On-device Recognition



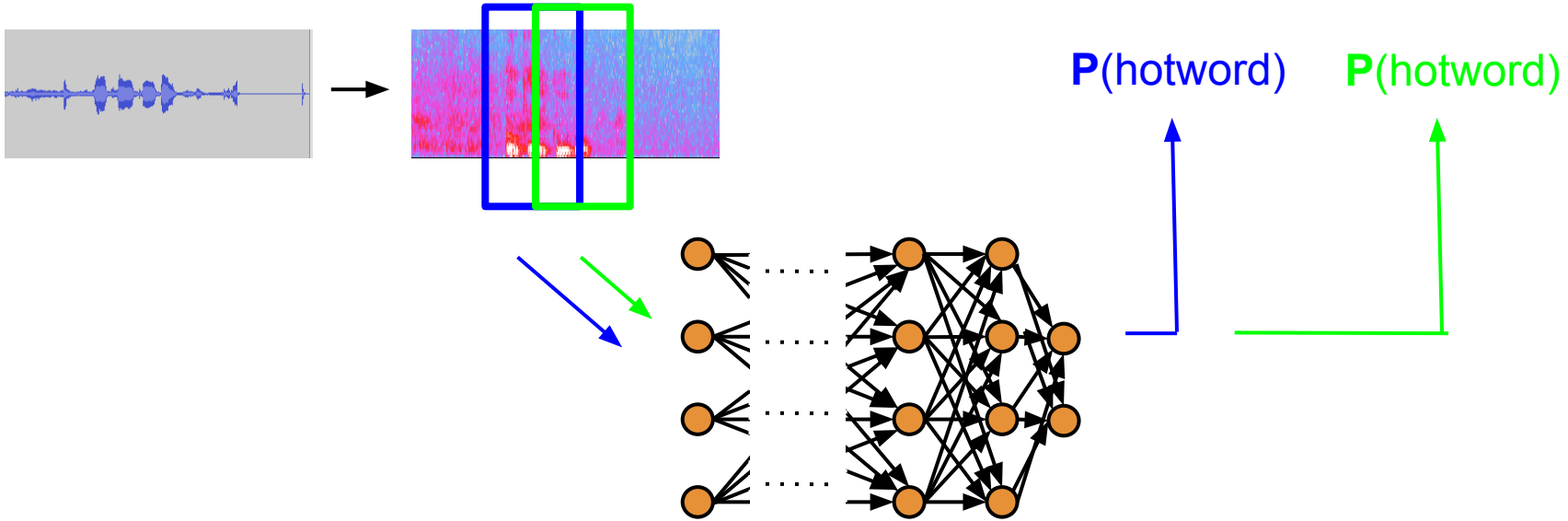
GET IT ON Google play

available in the chrome web store



Download on the App Store

Deep Neural Network **Hotword Detection**



The next 5 Billion

A challenge and an opportunity

6000 languages spoken in the world!

...but only 75-80 of these languages are a "primary written language"

Voice search languages

We cover 50 languages/dialects!, close to these 80 but…

- we have a major gap in emerging markets
  - South East Asia
  - Africa
  - India

- These are the markets where google's **next 5B users** are located!

Google

Why focus on Emerging markets?

Just a few data points

- 60 Million mobile users came online in *Asia in 3 months*
  - It is like adding the whole UK to the mobile network!
- In the past two years India internet users doubled from 100M to 200M users
  - the same growth took 6 years in the USA
- Consider this: Most of the people in the world who aren't online yet…. live in Asia or Africa

Conclusion / Q&A