Voice Search

Opportunities, Challenges, and Methods

8 October 2014
PROPOR 2014 - Sao Carlo, Brazil

Opportunities

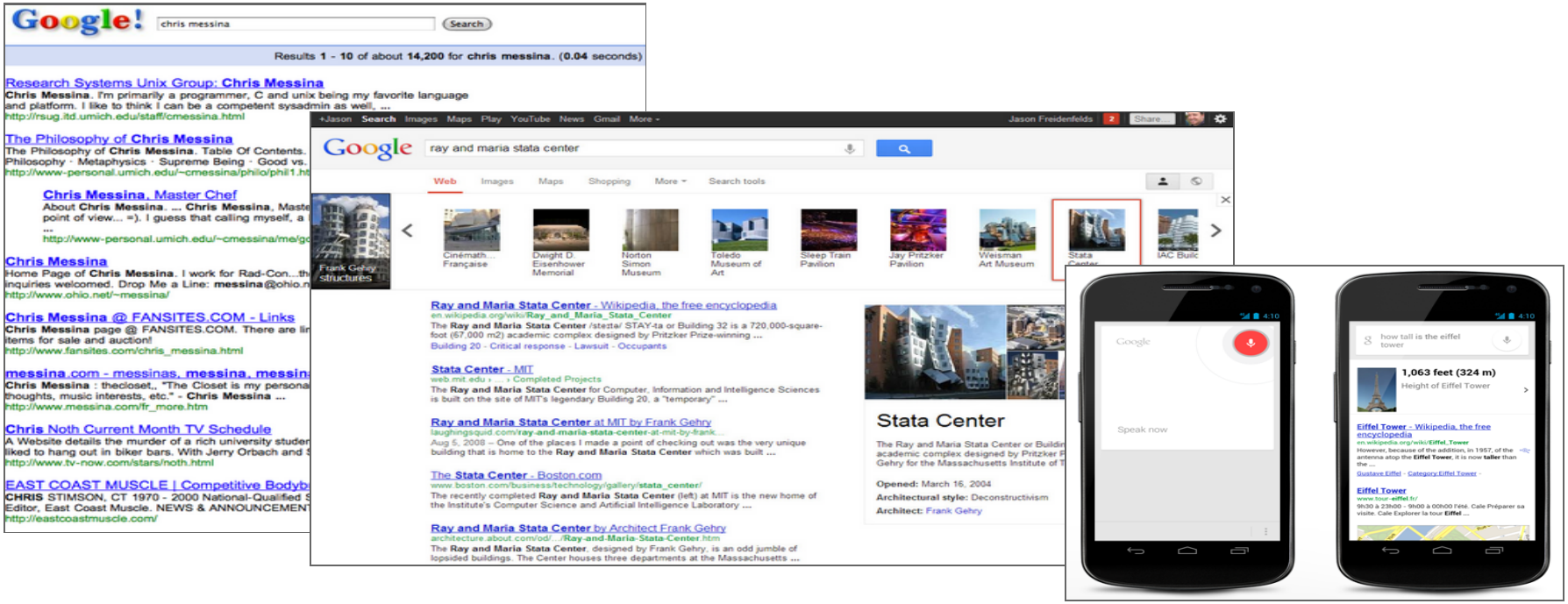
Speech is core to the future of mobile


Google
5:55
My Wishlist
Watch Now
My Movies
My TV Shows
PLAY MOVIES & TV


12:37
TUE, OCTOBER 22
12:37

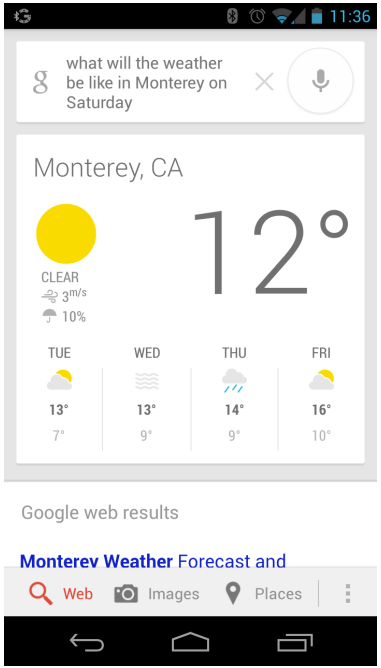
Google

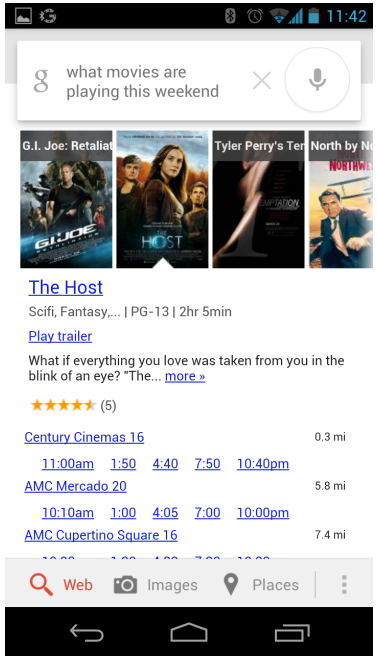
Evolution of Search

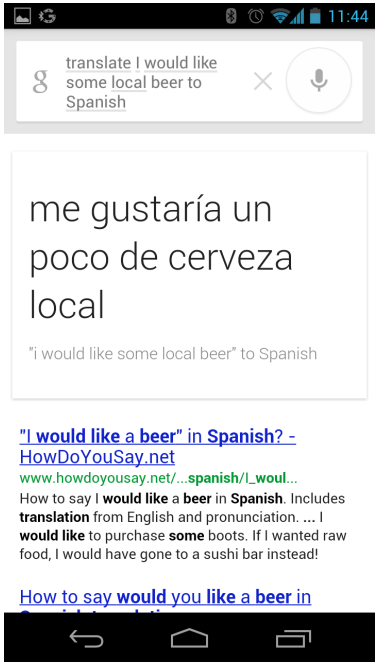


10 blue links → more than just keywords → mobile, voice

Google Voice search with spoken answers







Opportunities

Converse naturally with Google, on any platform
Accelerate how you get information & get things done

Challenges and Methods

It has to work everywhere, every time

How we're getting there.

Progress in Speech Recognition

2006

2011

TECHNOLOGY | INNOVATION, THE INTERNET, GADGETS, AND MORE. | APRIL 6 2011 4:36 PM

Now You're Talking!

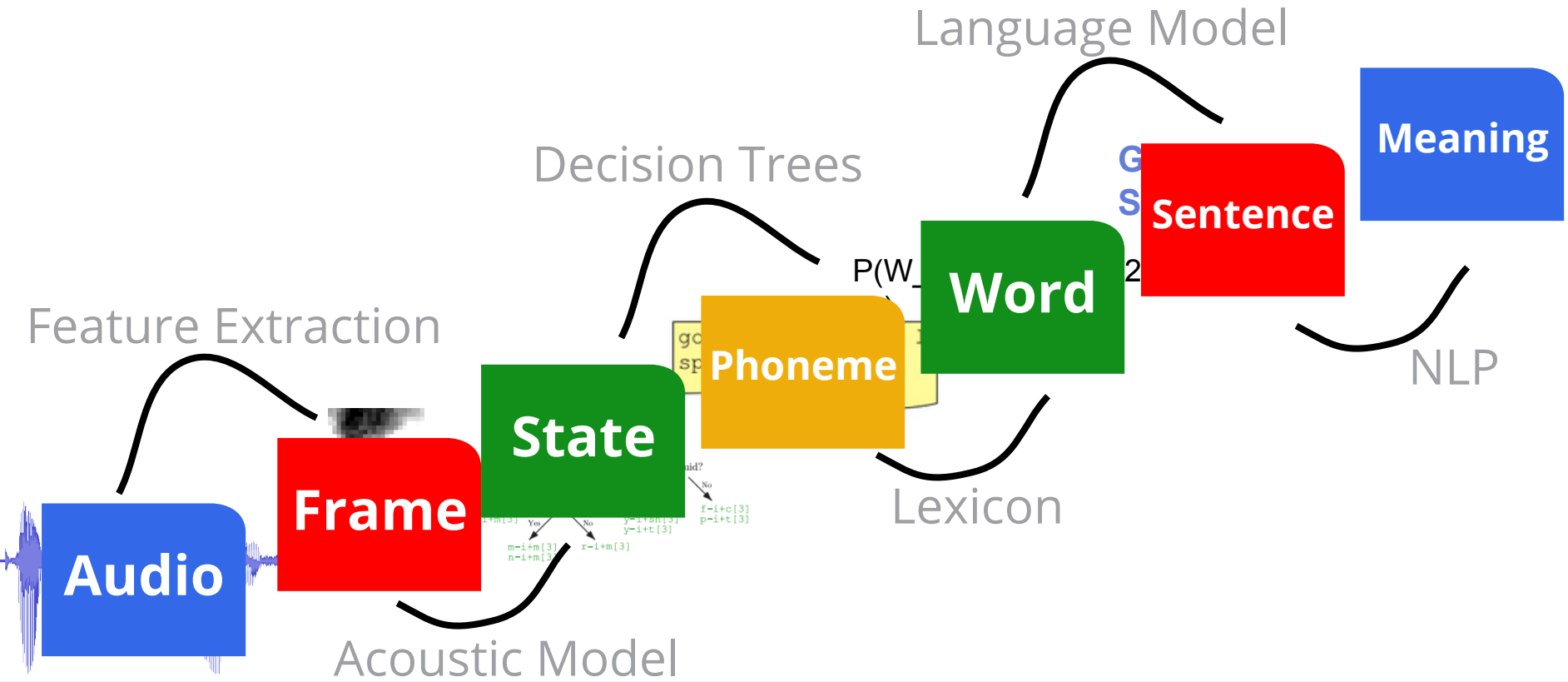
Google has developed speech-recognition technology that actually works.

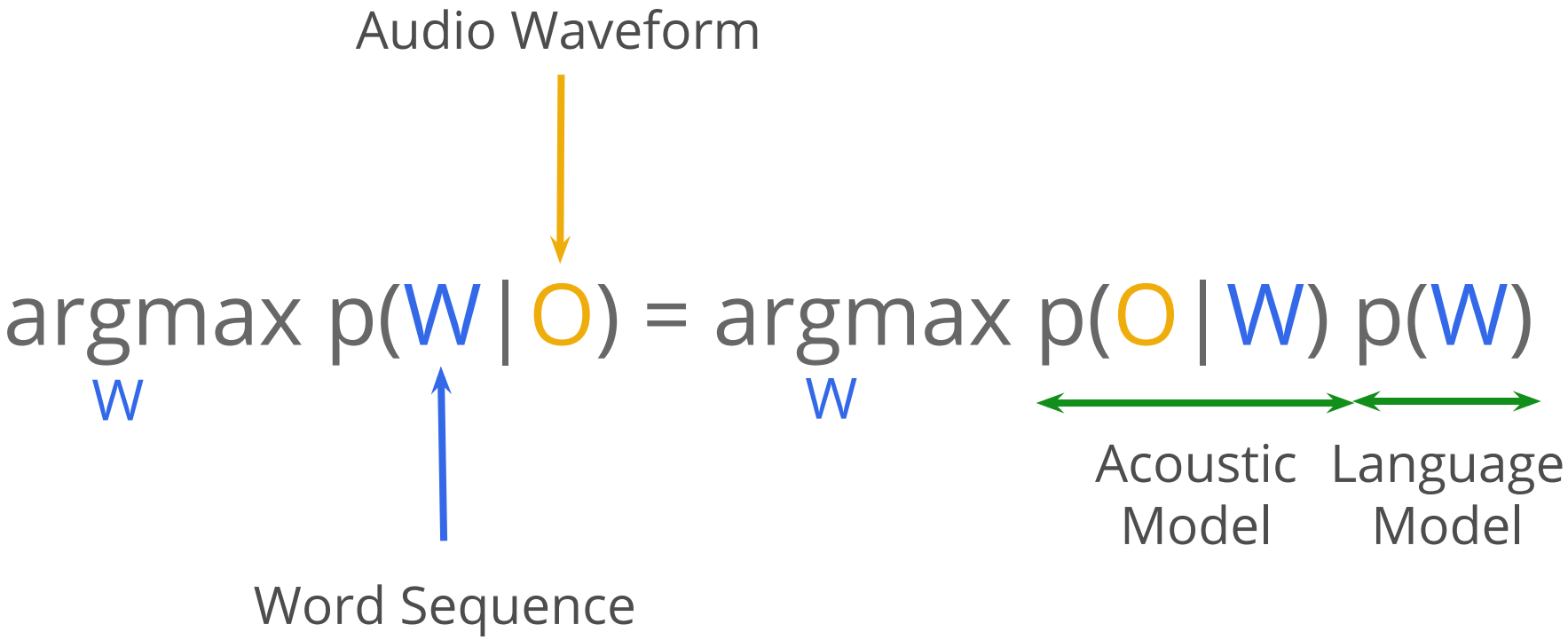
24 4 0

By Farhad Manjoo

Google

Speech Recognition Levels




Audio Waveform
argmax p(W | O) = argmax p(O | W) p(W)
W
W
Acoustic Model
Language Model
Word Sequence

Weighted Finite State Transducers (FSTs)

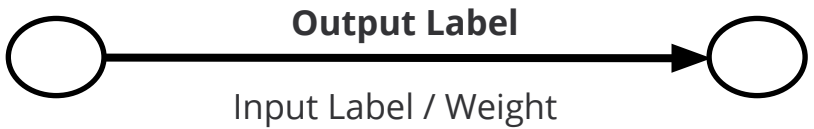
Challenge: How to efficiently represent the very large phonetic, lexical, and language models needed for Voice Search and other LVR tasks.

WFSTs one way to express this as probabilistic transductions.

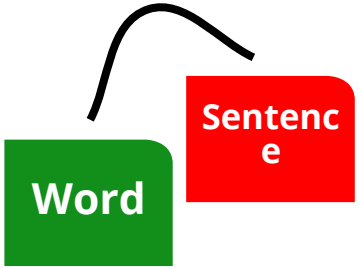
A mathematically sound way to express probabilistic graphs and algorithms over them. (e.g. Viterbi, forward-backward)

Powerful algorithms to combine and optimize these graphs.

Graphical representation:

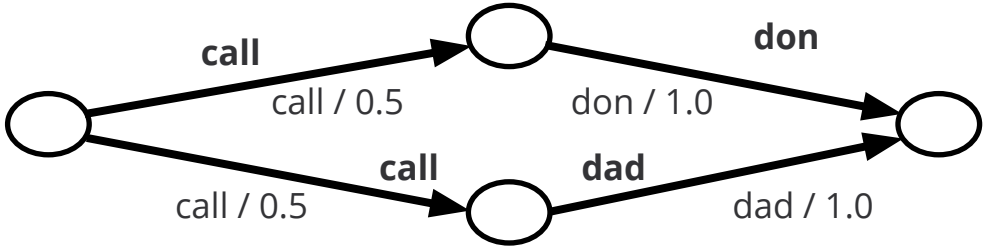


Mehryar Mohri, Fernando Pereira, and Michael Riley. "Weighted finite-state transducers in speech recognition." Handbook on Speech Processing and Speech Communication, Part E: Speech recognition, Springer-Verlag. 2008.

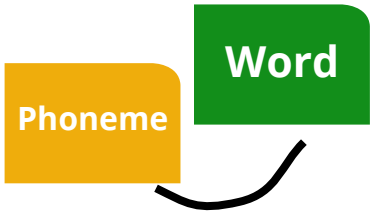


Language Model

- Toy Example:
  - “call don”
  - “call dad”



- In reality:
  - 100M+ states
  - estimated using billions of training examples.

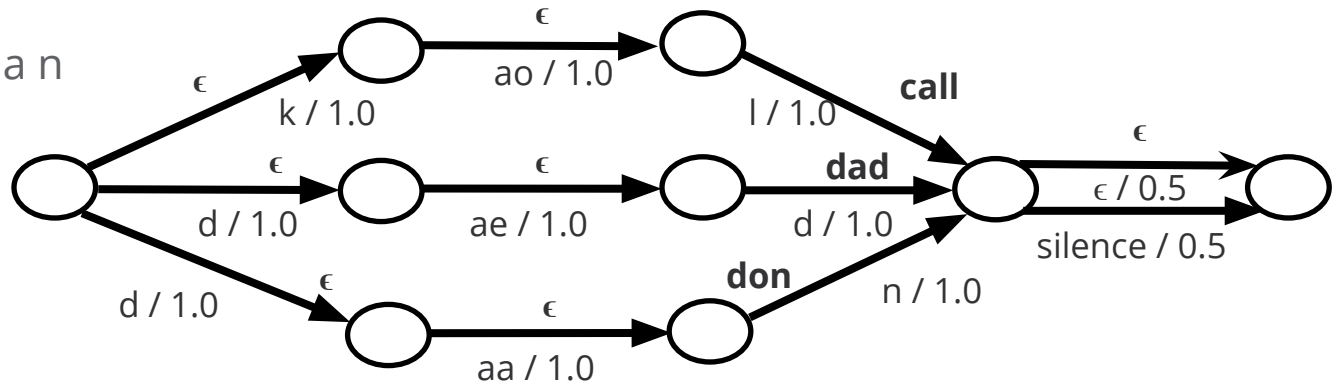


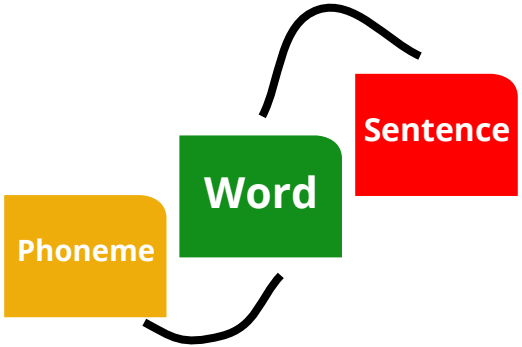
Lexicon

call: k ao l

dad: d ae d

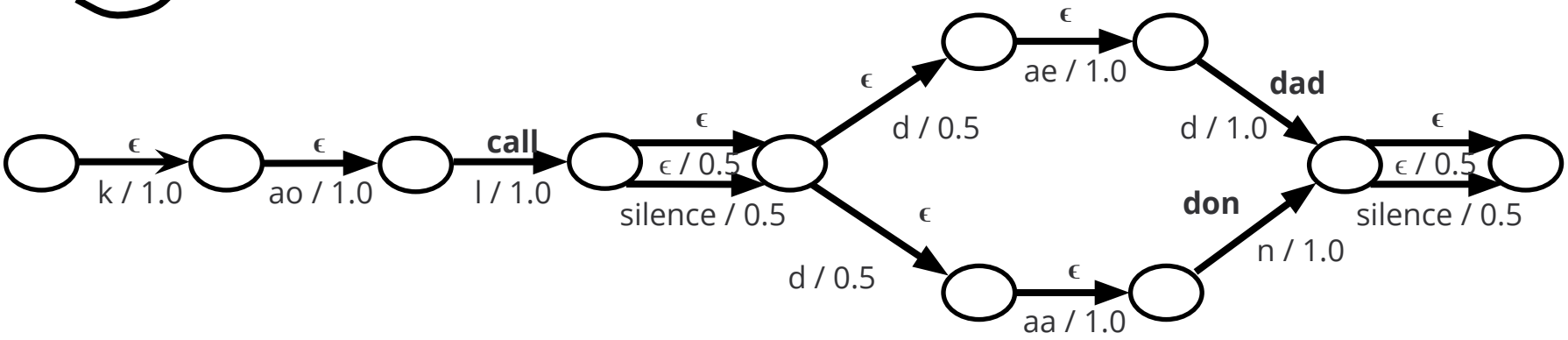
mom: d aa n

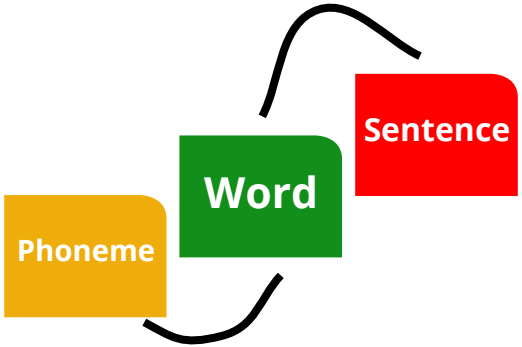




Cascading Transductions: Composition

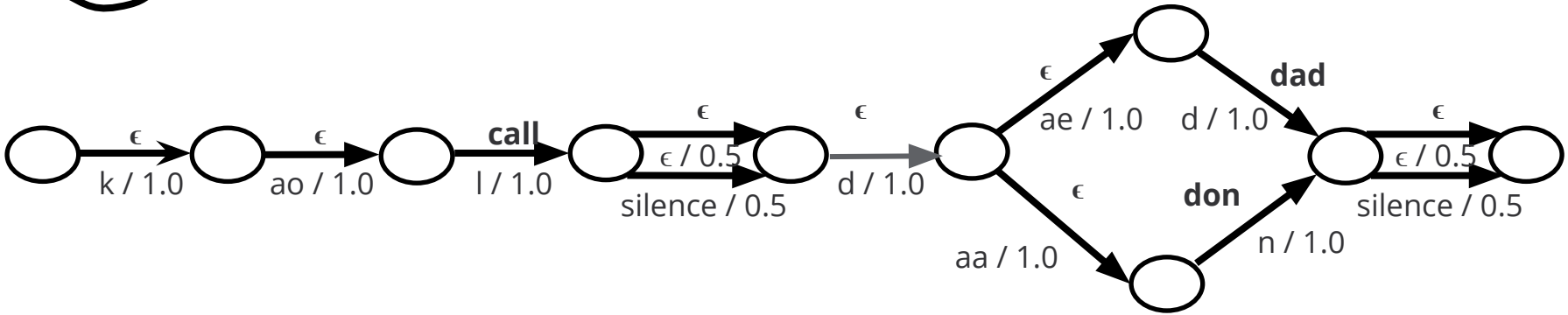
- Map *output* labels of Lexicon to *input* labels of Language Model.
- Join and optimize end-to-end graph.





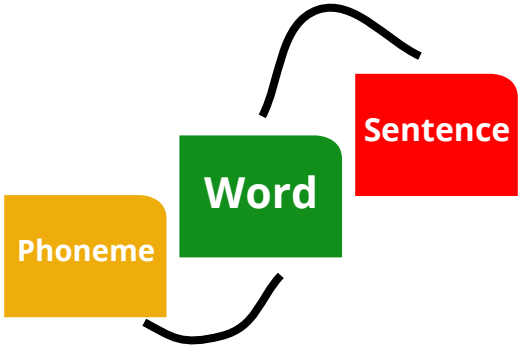
Optimization via Determinization

- Redundancy causes excessive search
- Determinization creates an equivalent transducer that has no two transitions from a state with the same label



Other operations: Minimization, Epsilon removal, Weight pushing.

Uses our open-source OpenFst: *www.openfst.org*



Context-Dependent Phonemes

- Acoustic models of phonemes built in **context** (typically triphonic): d_aa_n: */aa/ preceded by /d/ and followed by /n/*
- Application of context-dependency to the decoder graph very naturally implemented by composition with a *context-dependency* finite-state transducer: