WORD SEGMENTATION THROUGH CROSS-LINGUAL WORD-TO-PHONEME ALIGNMENT

Felix Stahlberg, Tim Schlippe, Stephan Vogel, Tanja Schultz

Karlsruhe Institute of Technology — Cognitive Systems Lab

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation — مؤسسة قطر عضو في
Qatar Foundation

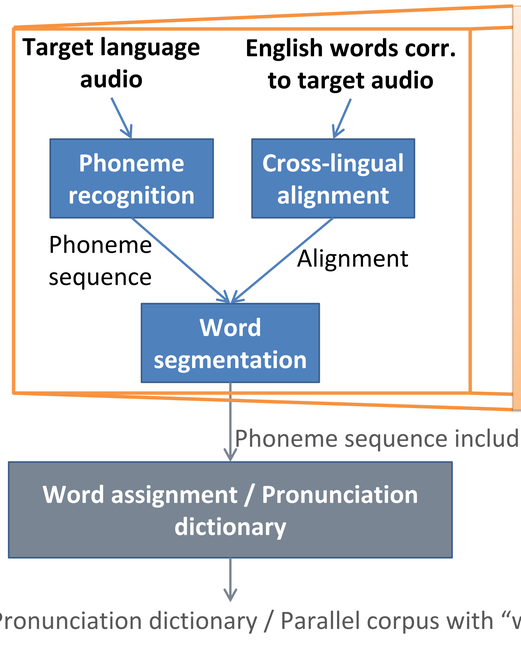
1. Overview

Long-term Goals

- Bootstrap speech technology for non-written and under-resourced languages
- Given
  - Audio data
  - Their written translations in another language (e.g. English)
- Collect training data for ASR and MT systems rapidly and at low cost
  - Pronunciation dictionary
  - Parallel corpus, language model

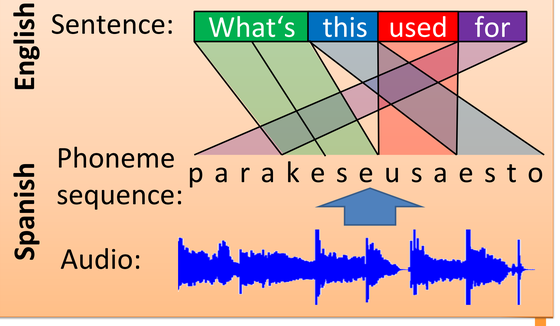
Goal of this Paper

- Segment phoneme sequences into word units using the written translations
- Simulate phoneme recognition errors realistically
- Compare our cross-lingual word segmentation method to monolingual ones, e.g. Adaptor Grammars (Johnson, 2008)



Cross-Lingual:

- GIZA++, **Model 3P**

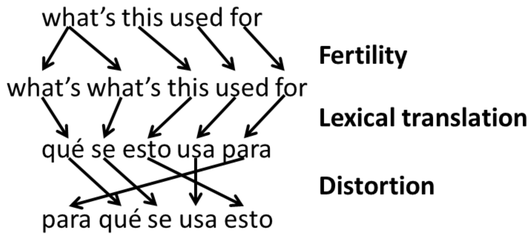


2. Cross-Lingual Alignment

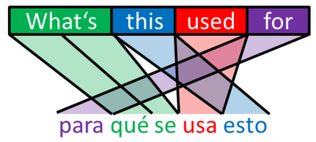
IBM Model 3

Problem: Generative story does not fit word-to-phoneme alignment

Generative Story



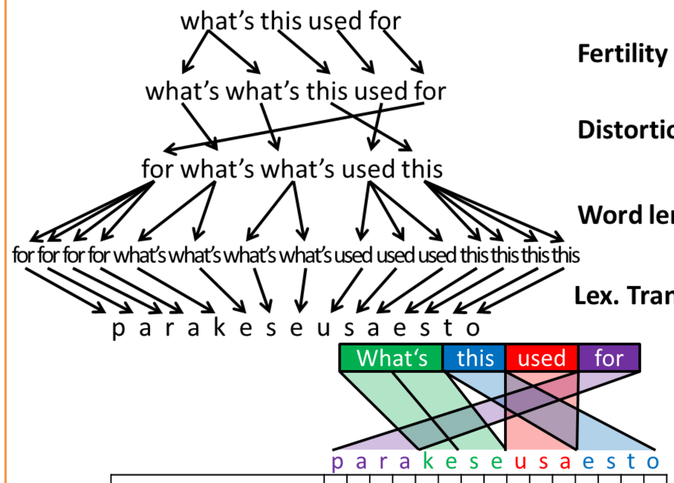
Alignment



English word position (i)	4	1	1	3	2

Model 3P

- Extends generative story of IBM Model 3 with additional steps
- Uses GIZA++ alignments to initialize Model 3P parameters
- Then our PISA alignment tool[1)] applies EM algorithm



English word position (i)	4	4	4	4	1	1	1	1	3	3	3	2	2	2	2
Target word position (π_{ik})	1	–	–	–	2	–	3	–	4	–	–	5	–	–	–
Target word length (ψ_{ik})	4	–	–	–	2	–	2	–	3	–	–	4	–	–	–
Phoneme position in target word (j)	1	2	3	4	1	2	1	2	1	2	3	1	2	3	4

1) The PISA alignment tool is available at http://pisa.googlecode.com/ under the GPL v3

3. Experiments and Results

Compare:

1. Adaptor Grammars (Monolingual)
2. GIZA++ word-to-phoneme alignments
3. Model 3P

Experimental Setup

- English-Spanish BTEC corpus (123k sentence pairs)
- Phoneme recognition errors up to 25.3% were simulated using the confusion matrix of a Spanish phoneme recognizer trained on the Spanish portion of GlobalPhone (Schultz, 2002)

Results