Towards Human Translation Guided Language Discovery for ASR

Sebastian Stüker
SLTU Workshop

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Introduction

- Training LVCSR systems require large amounts of training data
- Training the acoustic models requires large amounts of transcribed audio recordings
- Usually the necessary training material is recorded and then manually transcribed
- For phoneme based systems a pronunciation dictionary needs to be created: Needs time and money
- For many languages writing systems do not even exist and would need to be created for an ASR system
- Often the output of an ASR system is not needed directly but is part of a more complex system, e.g. a speech-to-speech translation system
- No “correct” word transcription necessary, only one that is suitable for further processing

Collecting Training Data from Translators

- Instead of manually annotating data on explicitly collected training data, collect the data on the fly
- Valuable, parallel data for training speech-to-speech translation systems is produced in real-life in human mediated translation scenarios (e.g. two people communication via an interpreter)
- We assume that one of the languages involved is a well-known language and the other one is an unknown, less prevalent language for which we want to create an ASR system.
- The speech from the well known language can be transcribed automatically, the speech from the unknown language not.
- Speech from the unknown language might not even be transcribed on word level, e.g. because no script exists or no expert for transcription.
- Phoneme Based transcription possible, maybe even automatically.
- In this work focus on exploratory experiments creating a suitable dictionary from the translation data.

Related Work

- Besacier et. all proposed speech translation based between phonemes in the less prevalent languages and words in the well known language in 2006
- In order to achieve good translation quality they proposed a monolingual word discovery algorithm operating on the phoneme string
- In our work we try to utilize the parallel information for discovering the words in the phoneme string of the new language.

Word Alignment

- Word alignments know from the field of statistical machine translation for training the translation model of a recognizer
- Source string with J words $s^J = s_1,s_2,\ldots,s_J$ and target string with I words $t^I = t_1,t_2,\ldots,t_I$
- A word-to-word A alignment between the two strings is defined as a subset of the Cartesian product of the word positions:

$$A \subseteq \{(i,j) : j = 1,...J; i = 1,...,I\}$$

- Usually each source word assigned to exactly one target word
- Thus alignments can be written as

  $a = a_1, \ldots a_J$

Word Alignment

- Alignments can be found with the help of statistical alignment and statistical translation models from SMT
- Similar as in ASR the probability in SMT is composed of a language model and a translation model $P(s^J|t^I)$
- Incorporating an alignment between s^J and t^J gives a statistical alignment mode

$$P(s^J a^J|t^J)$$

- The translation probability can be expressed as

$$P(s_1^J|t_1^I) = \sum_{a_1^J} P(s_1^J, a_1^J|t_1^I)$$

- Alignment probability usually depends on a set of parameters Θ
- Best set of parameters found on parallel data using EM training

Word Alignment

- Using the learned parameters one can find the most likely alignment between two sentences

$$\bar{a}_1^J = \underset{a_1^j}{\operatorname{argmax}}\, P_{\bar{\Theta}}(s_1^J, a_1^J | t_1^I)$$

- Different models exist, e.g. IBM 1-5, HMM models, hybrid models etc.
- Used IBM 4 Models for our experiments

Measuring Alignment Quality

- For measuring the quality of found alignments one can use the alignment error rate (AER)
- Uses manually aligned sentence pairs as references
- Alignments a_j are not unambiguous
- a_j labeled as either sure (S) or possible (P), $S \subseteq P$
- Precision and recall for an alignment can now be determined

$$recall = \frac{|A \cap S|}{|S|}, precision = \frac{|A \cap P|}{|A|}$$

- AER is derived from the F-Measure

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

Data

- Worked an English-Spanish version of the Basic Travel Expression Corpus (BTEC)
- English takes the role of the well known language, Spanish takes the role of the less prevalent language
- 155K parallel sentences, 12K English vocabulary, 20K Spanish vocabulary
- Removed sentences that were longer than 50 words, phonemes respectively
- Removed pairs exceeding sentence length ration of 9-1

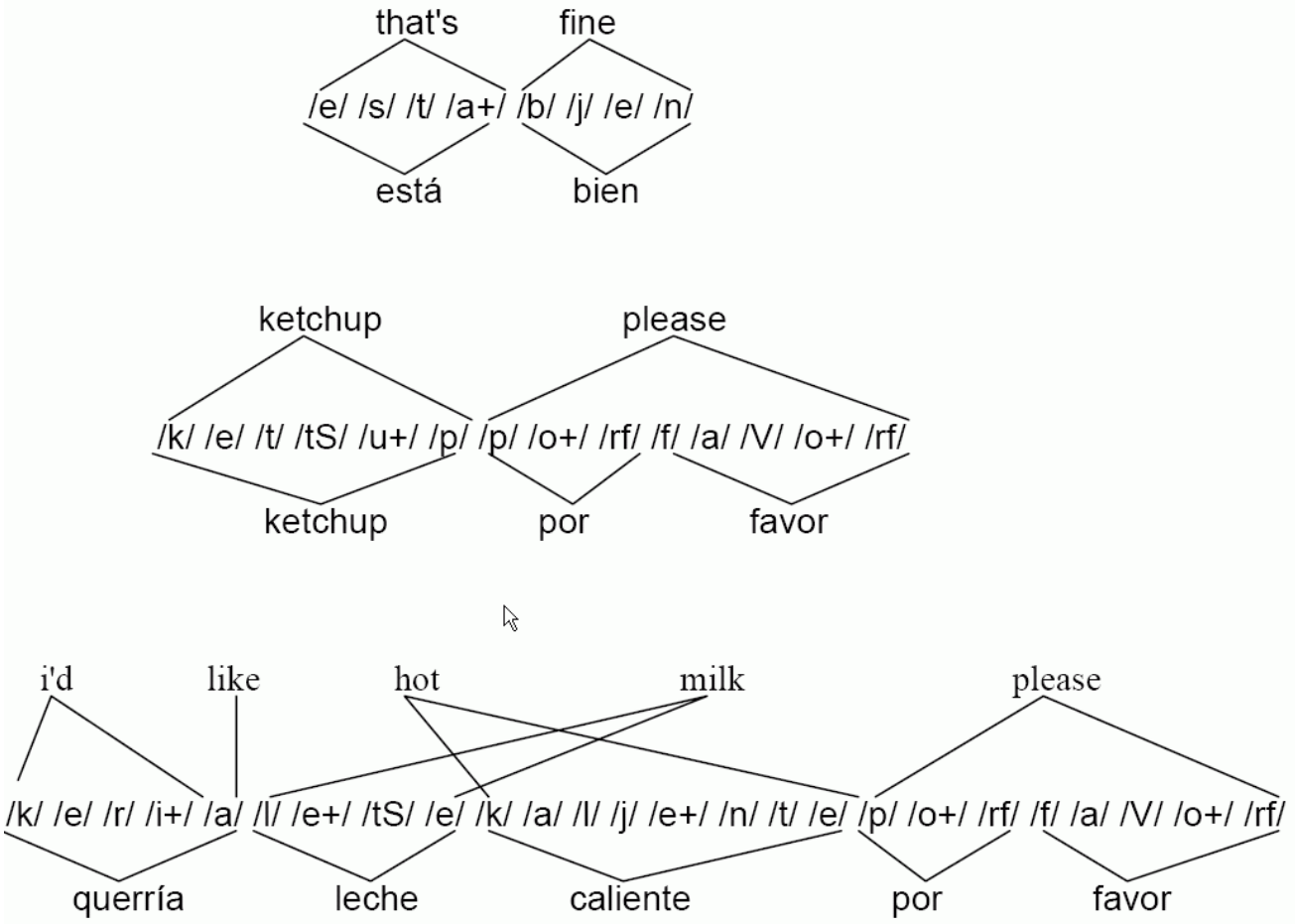
Word to Phoneme Alignment

- Used GIZA++ and Pharao training script for training IBM-4 models
- Trained model for word-to-word alignment (reference) and word-to-phoneme alignment

	Spanish phonemes	Spanish words
precision	83.5%	88.8%
recall	66.9%	75.3%
AER	25.4%	18.1%

- Degradation compared to words but still reasonable

Examples


that's
fine
/e/ /s/ /t/ /a+/ /b/ /j/ /e/ /n/
está
bien
ketchup
please
/k/ /e/ /t/ /tS/ /u+/ /p/ /p/ /o+/ /rf/ /f/ /a/ /V/ /o+/ /rf/
ketchup
por
favor
i'd
like
hot
milk
please
/k/ /e/ /r/ /i+/ /a/ /l/ /e+/ /tS/ /e/ /k/ /a/ /l/ /j/ /e+/ /n/ /t/ /e/ /p/ /o+/ /rf/ /f/ /a/ /V/ /o+/ /rf/
querría
leche
caliente
por
favor

Dictionary Extraction

- Only use alignment direction from Spanish phonemes to English words
- Every English word that is mapped to a phoneme sequence is a potential “Spanish” words
- Words mapped to the same sequence are merged
- Words mapped to none consecutive sequences are split
- Resulting dictionary contains 16K words
- 5,400 words have an exact, phonetic match in the original dictionary

Outlook

- End-to-End evaluation
- Take inverse alignment direction and heuristics into consideration
- Merge with approach from Besacier et. al. for word discovery