ASR for resource-deficient languages

Sadaoki Furui

Tokyo Institute of Technology Department of Computer Science

6912 languages



Source: http://www.ethnologue.com 2

- Motivation
 - Create robust speech recognition systems for resource deficient languages
- Why bother?
 - Everybody speaks English... right?

Nei - いいえ - Nej - Não - Non - Mhai - Lay - No - Mai



Outline

- WFST-based Icelandic ASR using machine translation
- Thai LVCSR for broadcast news
 transcription
- Indonesian LVCSR for read newspaper articles

WFST-based Icelandic ASR Using Machine Translation

The Icelandic language

- Germanic language
- Written Icelandic has not changed much since 13th century
- Currently nearly without dialects
- The Icelandic language is one of the most important heritage that the country has
 - The speakers are proud of the language and want to preserve it; (language Purism, since 19th century)
 - New Icelandic words are created instead of using foreign words

Tölva (computer) combines tala ("digit; number") and völva ("seeress").

- Highly inflected language (with four cases)
- Three genders

- We want to create an Icelandic Weather Information Speech Recognition System
- We are facing a huge problem!
 - Data Sparseness

Native speakers in millions



• How can resource deficient languages be helped??





How can resource deficient languages be helped?



- Using translated data may be useful !!
 - Manual (often hard work)
 - Machine Translation
 - Sentence-by-sentence – Needs parallel corpus
 - Word-by-word (WBW)
 - Only a dictionary is needed (often easy to obtain)

Translation method

• Word-by-Word (WBW) translation is expected to be useful for closely grammatically related languages



Previous work

- English corpus is
 translated into Icelandic
- Translation-based LM and sparse text-based LM are interpolated





Current research: Two setups



WFST (Weighted Finite State Transducer)based "T³ decoder"



- Small flexibility
 - Difficult to change partial models

13

Structure of the T³ decoder





Icelandic and English LMs are combined using a WBW translation transducer as follows:





Icelandic and English LMs are combined using a WBW translation transducer as follows:



Training text data

The Jupiter corpus

 (a weather information corpus
 developed by [!]]]
) was used as the English rich corpus





Weather information domain

Evaluation data

Attribute	Acoustic Corpus
No. male speakers	10
No. female speakers	10
No. utterances / keywords	400 / 844
Time (minutes)	12



- Why keyword detection?
 - Real applications often use keyword detection
 - Easy to compare results in different languages



Conclusion (Icelandic ASR)

- Our method has shown improvement using WBW translation in a WFST network
 - English output (3.4% absolute)
 - Icelandic output (2.2% absolute)

Thai LVCSR for Broadcast News Transcription

Characteristics of the Thai language

พ.ต.ท.ทักษิณ ชินวัตร นายกรัฐมนตรี เปิดเผยถึงกรณี การชุมนุมประท้วงการเจรจา เขตการค้าเสรี (เอฟทีเอ) ไทย-สหรัฐฯ ครั้งที่ 6 วันที่ 9-13 ม.ค.นี้ ที่ จ.เชียงใหม่ และเรียกร้องให้นำรายละเอียดการเจรจาเข้าพิจารณาใน สภาผู้แทนราษฎรก่อนว่า ไม่จำเป็น สภาไม่มีผู้เชี่ยวชาญ และไม่มีกฎหมายกำหนด

- No word boundary
- No specific rule to insert spaces within a paragraph

Background on Thai speech recognition research



Development of Thai BN LVCSR system

- Research on BN transcription system for Thai falls behind other languages
 - English: 1995 (Stern, 1997)
 - Japanese: 1997 (Matsuoka et al., 1997)
 - Mandarin: 1998 (Guo et al., 1998)
 - Italian: 2000 (Federico et al., 2000)
- We need to speed up our research activities to catch up with others



- 1. Development of Thai BN corpora
- 2. Development of a Thai BN LVCSR system

Development of Thai BN corpora

- BN speech corpus
 - About 17 hours of TV news transcribed
 - Structure information (section, speaker turn) and property tags (speaker's name, gender, speaking style, noise) annotated
 - Containing around 224k words
- BN transcript text corpus
 - About 35 hours of TV news transcribed
 - Containing around 573k words
- Newspaper text corpus
 - Covering about 5 years of news (2003-2007)
 - The size is much larger than the BN transcript text corpus

Recognition units for Thai LVCSR

- Problem: To find the optimal set of lexical units for LVCSR
 - Word
 - Words are defined in a dictionary
 - Segmentation errors always occur when unknown words exist
 - Pseudo-morpheme (PM)
 - A syllable-like unit in Thai written form
 - Compound Pseudo-morpheme (CPM)
 - PMs are combined together to create a CPM
 - No dictionary is required to construct CPMs
- Solution: CPM unit is used in our system

Compound pseudo-morpheme (CPM)

• Research on CPM for Thai

- CPM was created by merging PM based on mutual information (Thienlikit et al., 2004)
- CPM was created by merging PM based on statistic and linguistic information, using a decision tree (Jongtaveesataporn et al., 2007)



Decision tree training



Segmentation of a new corpus



Attributes used in the training

Geometric average of direct and reverse bigrams

$$M(w_{i}, w_{i+1}) = \sqrt{P(w_{i+1} \mid w_{i})P(w_{i} \mid w_{i+1})}$$

- Length (number of characters) of pseudomorphemes
- Distance from some specific words
- Whether or not the two pseudo-morphemes contain some specific words
- Specific characteristic of Thai compound words

Examples of automatic merging

นายโรเบิร์ตผู้นำคอมมูนิตี้เชื่อว่า Text นาย โรเบิร์ต ผู้นำ คอมมูนิตี้ เชื่อว่า Manual Mr. Robert leader community believes that PMSEG นาย โร เบิร์ต ผู้นำ คอม มู นิ ตี้ เชื่อ ว่ Iteration 1 นาย โรเบิร์ด ผู้นำ คอมมู นิดี้ เชื่อว่า Iteration 2 นาย โรเบิร์ด ผู้นำ คอมมูนิดี้ เชื่อว่า Iteration 3 นายโรเบิร์ด ผู้นำคอมมูนิดี้ เชื่อว่า

Text Corpora for LM training



Spoken Style Speech – Thai specific

- Words expressing additional meaning or emotion – 2 words
- Words showing politeness 4 words



Ending words

LM Interpolation



 $\mathsf{P}(\mathsf{w}|\mathsf{h}) = \lambda \mathsf{P}_{\mathsf{N}\mathsf{P}}(\mathsf{w}|\mathsf{h}) + (1-\lambda)\mathsf{P}_{\mathsf{B}\mathsf{N}}(\mathsf{w}|\mathsf{h})$

Experimental conditions

Acoustic model

- Gender-dependent acoustic models
- 12 MFCCs, delta, and delta energy
- Triphones, 1000 tied-states, 8 Gaussian mixtures
- Read speech data: 40.3 hours from 68 male and 68 female
- Language model
 - Various language models were trained by using the transcript and newspaper corpora
 - 3-grams
- Test set
 - 3000 utterances were randomly selected from the BN speech corpus
 - 1033 utterances were speech without background noise
- PM Error Rate (PER) is used as a measure for recognition accuracy

Comparison among various recognition units

Language models were trained from a newspaper text corpus



- Improvement of CPM over Word unit came from:
 - Independent of a dictionary → less severe segmentation errors
 - Low out-of-vocabulary rate

Results



Sentence PP comparison

Case	Ratio of sentences that match the case	Ratio of ending words to all words in sentences that match the case
Newspaper PP > Transcript PP	52.5%	1.9%
Newspaper PP < Transcript PP	47.5%	0.3%
Newspaper PP > Interpolated PP	83.5%	2.6%
Newspaper PP < Interpolated PP	16.5%	0.1%

Effects of LM interpolation

Case	Number of PMs
$Correct \rightarrow Correct$	128
Incorrect \rightarrow Incorrect	100
Incorrect \rightarrow Correct	164
Correct \rightarrow Incorrect	2
Surrounding PMs: Incorrect → Correct	99
Surrounding PMs: Correct → Incorrect	3

- Total PMs in the test set: 24507
- Number of repaired PMs by the interpolated model: 263
- Number of newly misrecognized PMs: 5

An analysis of PP & recognition results shows that

- The newspaper language model can predict written style sentences better than spoken style sentences
- The BN transcript language model can predict spoken style sentences better than written style sentences
- Interpolating the newspaper language model with the BN transcript language model improves the prediction of spoken style sentences

Conclusion (Thai LVCSR)

- A pioneer work on Thai BN transcription system has been conducted.
- Corpora required for building a system were developed.
- Recognition unit problem was solved by using CPM units.
- A language model for spoken style speech recognition was made by combining text and speech corpora

Future work (Thai LVCSR)



Indonesian LVCSR for Read Newspaper Articles

Indonesian language



- Used in the formal occasions, such as in education and working
- Vocabulary : borrowed heavily from many languages (Sanskrit, Arabic, Portuguese, Dutch, Chinese, English, and many others).
- 32 Sounds (6 vowels, 3 diphthongs,18 original consonants, 5 consonants in loanwords).
- Word accentuation is not semantically distinctive.
- Writing system : Latin characters.
- Grammar:
 - No inflectional changes by gender, tense, or plurals/singular,
 - The basic word order is S + V + O,
 - An agglutinative language.

Corpora development: Text corpus



- Automatic parsing:
 - remove all of the SGML-like tags
 - remove the title and ID of the document
 - add "<S>" and "" tag to separate sentences
 - transform numbers into words, for example "103" to "seratus tiga"
 - transform all capitalized alphabet into noncapitalized alphabet
 - remove all punctuation symbols, except ",",".","!" and "?", were changed into "."
 - delete the text inside the "(" and ")" symbol
 - fix incorrect written words.
- Manual correction:
 - split long sentences into several sentences
 - merge two or more short grammatically incorrect sentences into a correct grammatical sentence.

Corpora development: Speech corpus



Automatic phoneme labeling:

- Build and employ the grapheme to phoneme tool

- Add symbol /silB/ and /silE/ (silence beginning and ending sentences)

- Add symbol /sp/ (short pause) between words

Corpora development: Dictionary



• Selection criteria: words that occur in the text corpus for more than 3 times

- Employed to abbreviation words
- Number of vocabulary: 41,389 words (The Indonesian standard electronic dictionary consists of 38,143 words)

Baseline: Experimental conditions

Acoustic model:

- Using the leave-one-out method
 - 10 experiments:
 - Training: 18 speakers, 293 sentences
 - Testing : 2 speakers, 35 sentences
- Using the 1st-12th order MFCC, delta MFCC coefficients, and energy
- Context-dependent HMMs with 32 Gaussian mixtures per state
- Language model:
 - 2-gram and 3-gram
 - 3-gram perplexity: 87.8
 - OOV (out-of-vocabulary) rate: 1.3%.

Baseline: Experimental results



Baseline: Error analysis



Proper noun adaptation (PNA)

- Supervised adaptation based on the MLLR technique using 8 classes
- Adaptation data: proper noun utterances
 extracted from the training set
 - 742 proper nouns/speaker
 - 13K proper nouns/experiment

Improvement by proper noun adaptation (PNA)



Proper noun (PN) recognition results

Baseline system

PNA system



• PN error rate for the baseline system : 33.1%

• PN error rate for the PNA System : 25.5%

Speaker adaptation (SA)

- Supervised adaptation based on the MAP technique
- Applied to both baseline models and proper noun adapted models.
- Adaptation data:
 - Baseline models: 293 sentences/speaker
 - Proper noun adapted models: 742
 words/speaker

SA: Experimental Results



Conclusion (1) (Indonesian LVCSR)

- One of the first LVCSR systems for Indonesian language was built.
- A 14.5 hours bi-phonetically balanced corpus and a 610K sentences of the text corpus were collected.
- The dictionary generated from the collected text corpus which consists of 41,389 words covers almost all standard Indonesian vocabulary.
- The average recognition accuracy of the baseline system was 80.0%.

Conclusion (2) (Indonesian LVCSR)

- Proper noun recognition was acoustically more difficult than non-proper noun recognition.
 Supervised adaptation for proper noun-specific acoustic models based on the MLLR approach was proposed. In average, 1.8% improvement (absolute value) was achieved.
- Supervised speaker adaptation was performed using a MAP approach. It increased the accuracy by 3.8% on average.

Future Works (Indonesian LVCSR)

- To investigate the effect of various Indonesian dialects to the recognition result, a new speech corpus representing various dialects in Indonesia needs to be collected.
- To make the result of this research applicable to practical conditions, new text and speech corpora need to be collected and some adaptation should be conducted.
- Developing practical systems.

Summary

- Resource-deficient language LVCSR is important and difficult.
- We have tried various languages, including Icelandic, Thai and Indonesian.
- Various techniques, such as translation, interpolation and adaptation, have shown improvement in recognition accuracy using a limited amount of speech data from the resource-deficient language.

spurningar - 質問 - Fragen - spørgsmål - questions

