Spoken Term Detection Using Multiple Speech Recognizers’ Outputs at NTCIR-9 SpokenDoc STD subtask

Hiromitsu Nishizaki  
Yuto Furuya  
Satoshi Natori  
Yoshihiro Sekiguchi  
University of Yamanashi, Japan
Outline

- Introduction
- Spoken Term Detection (STD) using multiple speech recognizers
  - Overview of our STD framework
  - Multiple speech recognizers
  - Phoneme Transition Network (PTN)-based indexing
  - Search engine and experimental result
- False detection control
  - Introducing the control parameters
  - Experimental result
- Conclusion
Introduction

- Back ground
  - Much multi-media data available
    - improved the environment on multi-media
    - improved the infrastructures
  - More efficient utterance retrieval
    - key words or phrases extraction
  - Term detection from LVCSR output
    - the out-of-vocabulary problem
    - recognition errors get worse detection performance

- Our goal

  Improving Spoken Term Detection performance
Summary of our research

Multiple speech recognizers
• Combination of “1 decoder x 2 AMs x 5 LMs”
• This made speech recognition performance better

Construction of index for STD and search engine
• Confusion Network based indexing
• Term detection using a simple term search method

STD performance evaluated on the formal-run
• The index from multiple speech recognizers’ outputs got the highest STD performance
• Introducing false detection parameters makes the STD performance more improvement

2011/12/8
NTCIR-9 SpokenDoc task
Outline

Introduction
- Spoken Term Detection (STD) using multiple speech recognizers
  - Overview of our STD framework
  - Multiple speech recognizers
  - Phoneme Transition Network (PTN)-based indexing
  - Search engine and experimental result

False detection control
- Introducing the control parameters
- Experimental result

Conclusion
STD task flow diagram

Index build phase

Speech Data

Recognition System #1

Recognition System #10

Converting to sub-word sequences

Making network-based index

network-based index

Text Terms

Phoneme Terms

Term Search engine

STD Result

Search phase

2011/12/8

NTCIR-9 SpokenDoc task
Multiple speech recognizers

LVCSR decoder
• Julius rev.4.1.3

5 types of Language Models
• Word based trigram : WBC
• Hiragana based trigram : WBH
• Syllable based trigram : CB
• A bi-syllables based trigram : BM
• Nothing : Non

10 speech recognizers

2 types of Acoustic Models
• syllable based HMM : Syl
• tri-phone HMM : Tri

each model was trained from the open data

2011/12/8 NTCIR-9 SpokenDoc task
Phoneme Transition Network (PTN)

- Phoneme-level Confusion Network based index for STD
  - It called as ``PTN’’ (Phoneme Transition Network)
  - PTN is built from multiple speech recognizers’ outputs
Example of building PTN-based index

speech utterance “Cosine” ( /k o s a i N/ )

LM/AM

Outputs of 10 recognition systems (all outputs are converted into phoneme sequence)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBC/Tri</td>
<td>k o s a i N</td>
</tr>
<tr>
<td>WBH/Tri</td>
<td>q o s u a a a N</td>
</tr>
<tr>
<td>CB/Tri</td>
<td>k o s @ a m a i a</td>
</tr>
<tr>
<td>BM/Tri</td>
<td>k o s @ a @ @ N</td>
</tr>
<tr>
<td>Non/Tri</td>
<td>k o s @ a @ @ N</td>
</tr>
<tr>
<td>WBC/Syl</td>
<td>@ @ s @ a @ @ N</td>
</tr>
<tr>
<td>WBH/Syl</td>
<td>b o s @ a a a a @</td>
</tr>
<tr>
<td>CB/Syl</td>
<td>@ @ s @ a b @ i @</td>
</tr>
<tr>
<td>BM/Syl</td>
<td>@ @ s @ a @ @ @ N</td>
</tr>
<tr>
<td>Non/Syl</td>
<td>@ @ s @ a @ @ @ N</td>
</tr>
</tbody>
</table>

Base output

PTN based index

Arc Node

Terminal Node

2011/12/8

NTCIR-9 SpokenDoc task
Search engine (no false detection control)

- Simple search engine
  - Dynamic Programming (DP) based engine
  - Both endpoints free
  - Edit distance is used for calculating DP cost between an index and a query term
- We modified the simple DP framework to adapt the PTN-based index
Example of the modified DP framework for PTN-based index (baseline technique)

- NULL transition cost is set to 0.1
- no insertion errors
- Matching! (Transition cost = 0)
- Cost: 0.3

NTCIR-9 SpokenDoc task
Experimental setup

Data for STD task
- CORE set of the STD task (about 40 hours, $144 \times 10^3$ sec.)

Query
- 50 queries for the CORE set
  - Including 31 out-of-vocabulary (OOV) queries

Evaluation measure
- Recall-Precision curve
- F-measure at the maximum point of the curve

2011/12/8  NTCIR-9 SpokenDoc task
### Indices for STD

- **Two types of Index**

<table>
<thead>
<tr>
<th>Index</th>
<th># of hypothesis</th>
<th>type of index</th>
<th>How to make</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>1</td>
<td>Phoneme-base</td>
<td>1-Best output of “CB/Tri”</td>
</tr>
<tr>
<td><strong>PTN</strong></td>
<td>10</td>
<td>Phoneme-base</td>
<td>10 types of output</td>
</tr>
</tbody>
</table>

Baseline STD is performed by the simple DP on the transcription of “CB/Tri.”
STD results

Maximum F-measure = 55.6%

Maximum F-measure = 71.4%
Outline

Introduction

Spoken Term Detection (STD) using multiple speech recognizers
  - Overview of our STD framework
  - Multiple speech recognizers
  - Phoneme Transition Network (PTN)-based indexing
  - Search engine and experimental result

False detection control
  - Introducing the control parameters
  - Experimental result

Conclusion
Robust for false detections

- False detection control for more STD improvement
- Our approach generates many false detections because of:
  - using multiple speech recognizers’ outputs
  - using a network-based index

Two types of control parameters!

Voting
- The number of recognizers outputting the same phoneme on the same arc

ArcWidth
- The number of arcs between successive two nodes

NTCIR-9 SpokenDoc task
False detection control parameters

Voting

A phoneme from more recognizers may have better confidence

PTN based index

ArcWidth

The less number of arcs may enhance the reliability of the recognized phonemes

2011/12/8

NTCIR-9 SpokenDoc task
Experimental results (with false detection control)

Maximum F-measure = 72.5%

Maximum F-measure = 71.4%
Conclusion

Summary

• Using multiple speech recognizers for STD
  • Multiple recognizers make STD performance better
  • Integrating multiple recognizers’ output into PTN was very powerful to improve the performance

Future works

• Improving index
  • Reduction of unnecessary information

• Improving search engine
  • Developing new control parameters in the STD engine
  • Customizing the engine depending on an inputted query

2011/12/8 NTCIR-9 SpokenDoc task
Thank you for your attention

Our poster will be posted at the poster session tomorrow