Combination of DTW-based and CRF-based Spoken Term Detection on the NTCIR-11 SpokenQuery&Doc SQ-STD Subtask

NTCIR

Hiromitsu Nishizaki, Naoki Sawada, Satoshi Natori (University of Yamanashi)
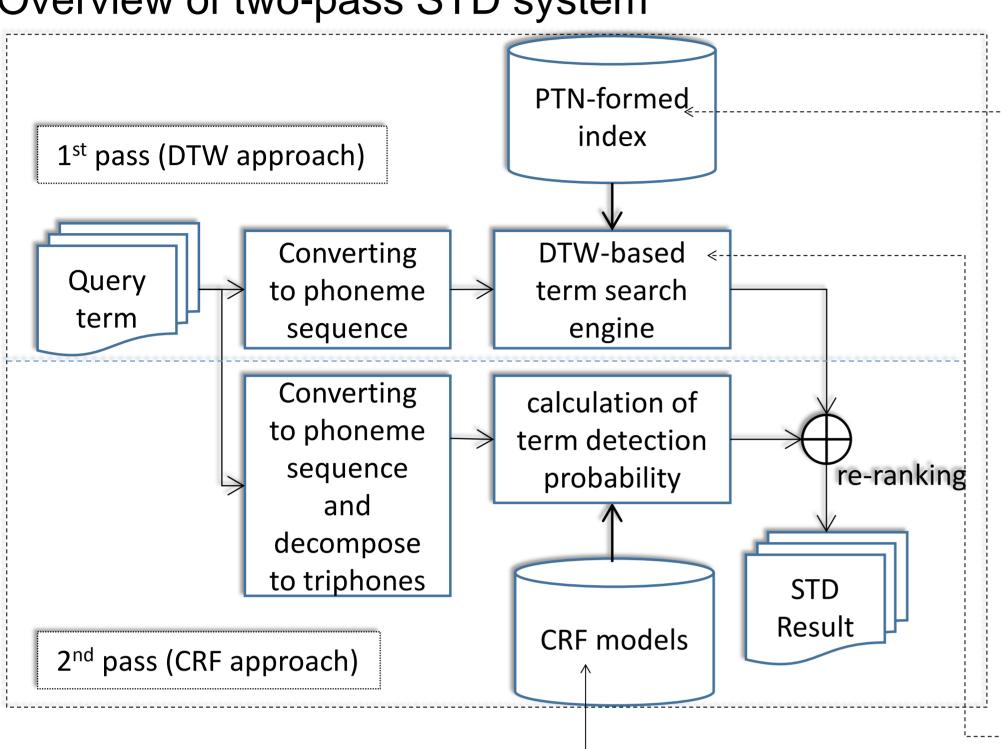Kentaro Domoto, Takehito Utsuro (University of Tsukuba)

① Introduction

- We proposed a two-stage STD method
  - Combination of Dynamic Time Warping (DTW) based and CRF-based methods

- The DTW-based method outputs so many false detections
- Goal of this research:
  - To control false detections by the DTW method using CRF-based triphone detection framework
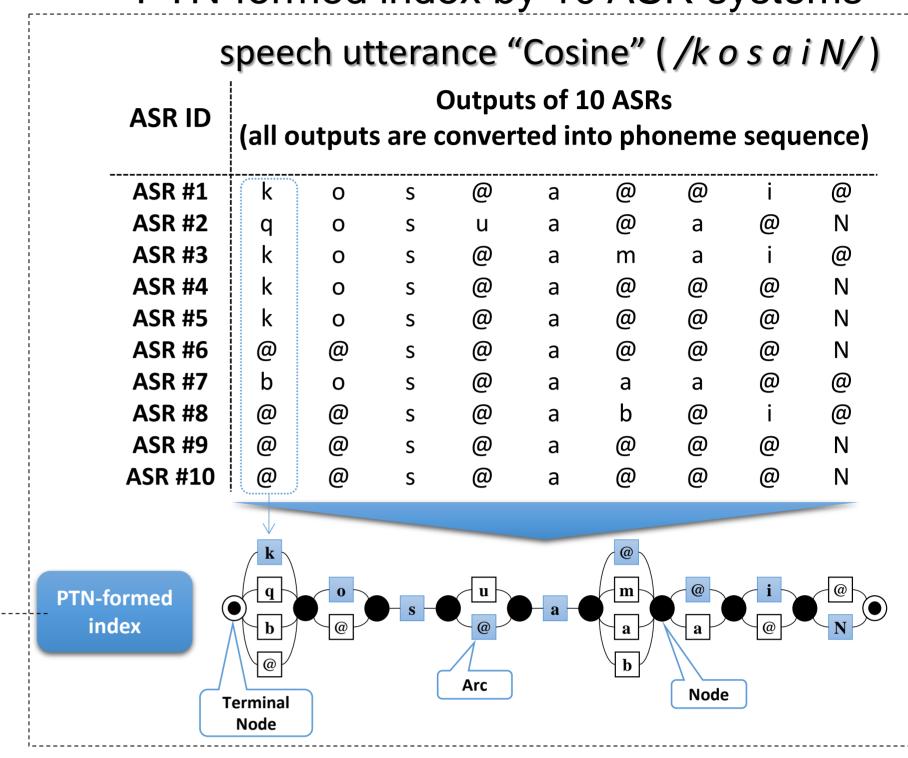
② Two-pass STD framework

Overview of two-pass STD system

1st pass (DTW approach)

Query term → Converting to phoneme sequence → DTW-based term search engine ← PTN-formed index

Converting to phoneme sequence and decompose to triphones → calculation of term detection probability ← CRF models

⊕ re-ranking → STD Result

2nd pass (CRF approach)

PTN-formed index by 10 ASR systems

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

ASR ID	Outputs of 10 ASRs (all outputs are converted into phoneme sequence)								
ASR #1	k	o	s	@	a	@	@	i	@
ASR #2	q	o	s	u	a	@	a	@	N
ASR #3	k	o	s	@	a	m	a	i	@
ASR #4	k	o	s	@	a	@	@	@	N
ASR #5	k	o	s	@	a	@	@	@	N
ASR #6	@	@	s	@	a	@	@	@	N
ASR #7	b	o	s	@	a	a	a	@	@
ASR #8	@	@	s	@	a	b	@	i	@
ASR #9	@	@	s	@	a	@	@	@	N
ASR #10	@	@	s	@	a	@	@	@	N

PTN-formed index

Terminal Node   Arc   Node

DTW-based term detection (1st pass)

DTW matrix

DTW score 0.9

Query term (f u j i s a N)

target speech → PTN conversion → k - u - z/j - i/e - a/i - s - a/o - N

features for CRF training

current token

ASR #1	s o n o @ n (i) @ q p o N n o @ d e @ w a
ASR #2	s o n o @ n (i) @ q p o N n o @ d e @ w a
ASR #3	s o n o @ n (i) @ @ b a N n o @ u e @ w a
ASR #4	s o n o N n f u p a N n u N b e e w a
ASR #5	s o n o @ n i @ q p o N n o @ d e @ w a
ASR #6	s o n o @ m i @ @ b a N n o @ u e @ w a
ASR #7	s o n o @ n i @ q p o N n o @ d e @ w a
ASR #8	s o n o @ n i @ @ b a N n o @ u e @ w a
ASR #9	s o n o @ n i @ @ b a N n o @ u e @ w a
ASR #10	s o n o @ n i @ q p o N n o @ d e @ w a

- unigrams
- cross-ASR bigrams
- in-ASR bigrams
- in-ASR trigrams
- phoneme-based transcriptions by ASRs

O O O O O B I I I O O O O O O O O O O   BIO tags for triphone "n-e-p"

B : beginning tag of the triphone
I : inside tag of the triphone
O : outside tag of the triphone

Training example of "n-e-p" detection model by CRF

All triphones are modeled by this framework.

1st pass result

Rank	ID	Score
1	S1_0125	0.90
2	S3_0172	0.85
3	S1_0008	0.84
...	...	...
1000	S1_0482	0.01

2nd pass (final) result

Rank	ID	Score
1	S1_0008	0.91
2	S4_0219	0.90
3	S1_0125	0.79
...	...	...
1000	S3_0361	0.00

Re-ranking

$$\frac{(\gamma^2 + 1) \cdot DTW(T,i) \cdot CRF(T,i)}{\gamma^2 \cdot DTW(T,i) + CRF(T,i)}$$

Detection probability of query "f u j i s a N" at the utterance A **0.8**

$$P(T|x_i) = \left(\prod_{j=1}^{N} P_{t_j}(y|x_i) \right)$$
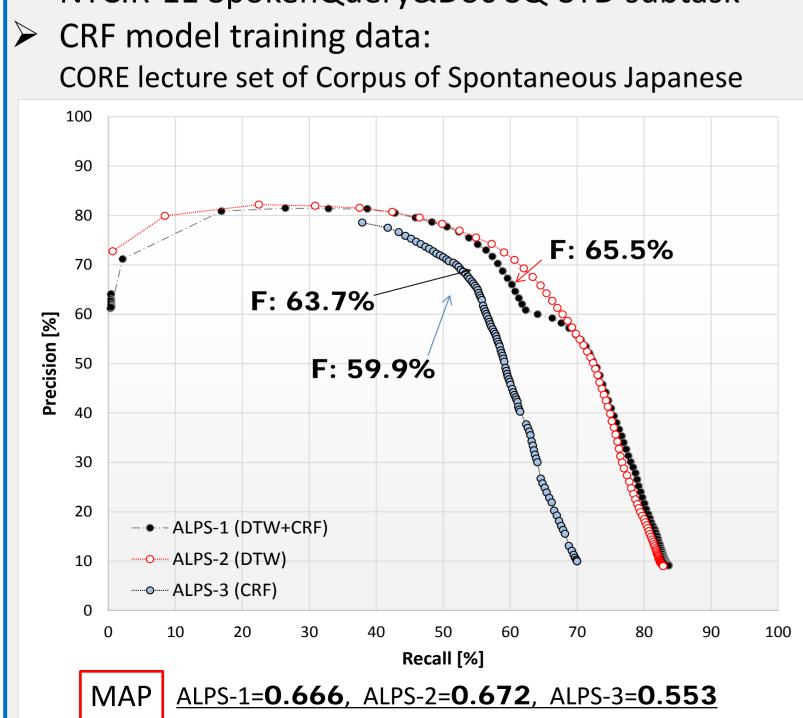
③ STD result

- Evaluation data:
  NTCIR-11 SpokenQuery&Doc SQ-STD subtask
- CRF model training data:
  CORE lecture set of Corpus of Spontaneous Japanese

F: 65.5%
F: 63.7%
F: 59.9%

Precision [%] vs Recall [%]

- ALPS-1 (DTW+CRF)
- ALPS-2 (DTW)
- ALPS-3 (CRF)

MAP ALPS-1=0.666, ALPS-2=0.672, ALPS-3=0.553

CRF-based term detection

triphone "j-i-s" detection result by CRF

Query term (f u j i s a N) → decomposing to triphones → f-u-j, u-j-i, j-i-s, ...

utterance A

ASR name (LM/AM)	Outputs of 10 ASR Systems						
WBC/Tri	f	u	j	i	s	a	N
WBH/Tri	f	u	z	u	sh	a	n
CB/Tri	s	i	z	u	y	a	n
BM/Tri	f	u	@	e	y	a	m
Non/Tri	k	o	J	u	g	o	n
WBC/Syl	f	u	J	i	z	a	N
WBH/Syl	k	e	z	e	ch	a	n
CB/Syl	@	i	t	i	s	a	q
BM/Syl	f	u	r	u	s	a	N
Non/Syl	s	u	j	i	h	a	q
B label	0.1	0.1	**0.8**	0.2	0.1	0.2	0.1
I label	0.1	0.4	0.2	**0.7**	**0.9**	0.1	0.1
O label	0.8	0.6	0.0	0.1	0.0	0.7	0.8

Probability of j-i-s
j:0.8 × i:0.7 × s:0.9   j-i-s:0.5
a:0.9
N:0.9

Detection probability of each triphone

Calculation formula

$$P_{t_j}(y|x_i) = \prod_{L=B}^{I_{tail}} P_{t_j}(L|x_i)$$

④ Conclusion

- We proposed the two-stage STD method using the combination of DTW-based and CRF-based detection approaches
- The CRF-based STD approach did not improved the DTW-based STD results
  - Because the CRF models were trained from the CSJ speeches (mismatch)