

Selected topics from ASR research for many languages at Tokyo Tech

Sadaaki Furui
Tokyo Institute of Technology
Department of Computer Science
furui@cs.titech.ac.jp

Introduction

6912 languages

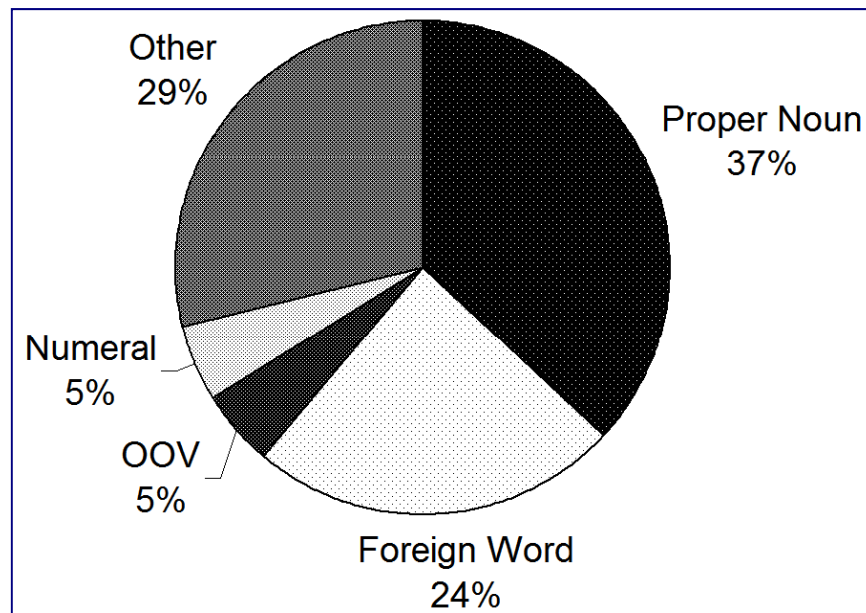


Source: <http://www.ethnologue.com>

1. Adaptation to pronunciation variations in Indonesian spoken query-based information retrieval

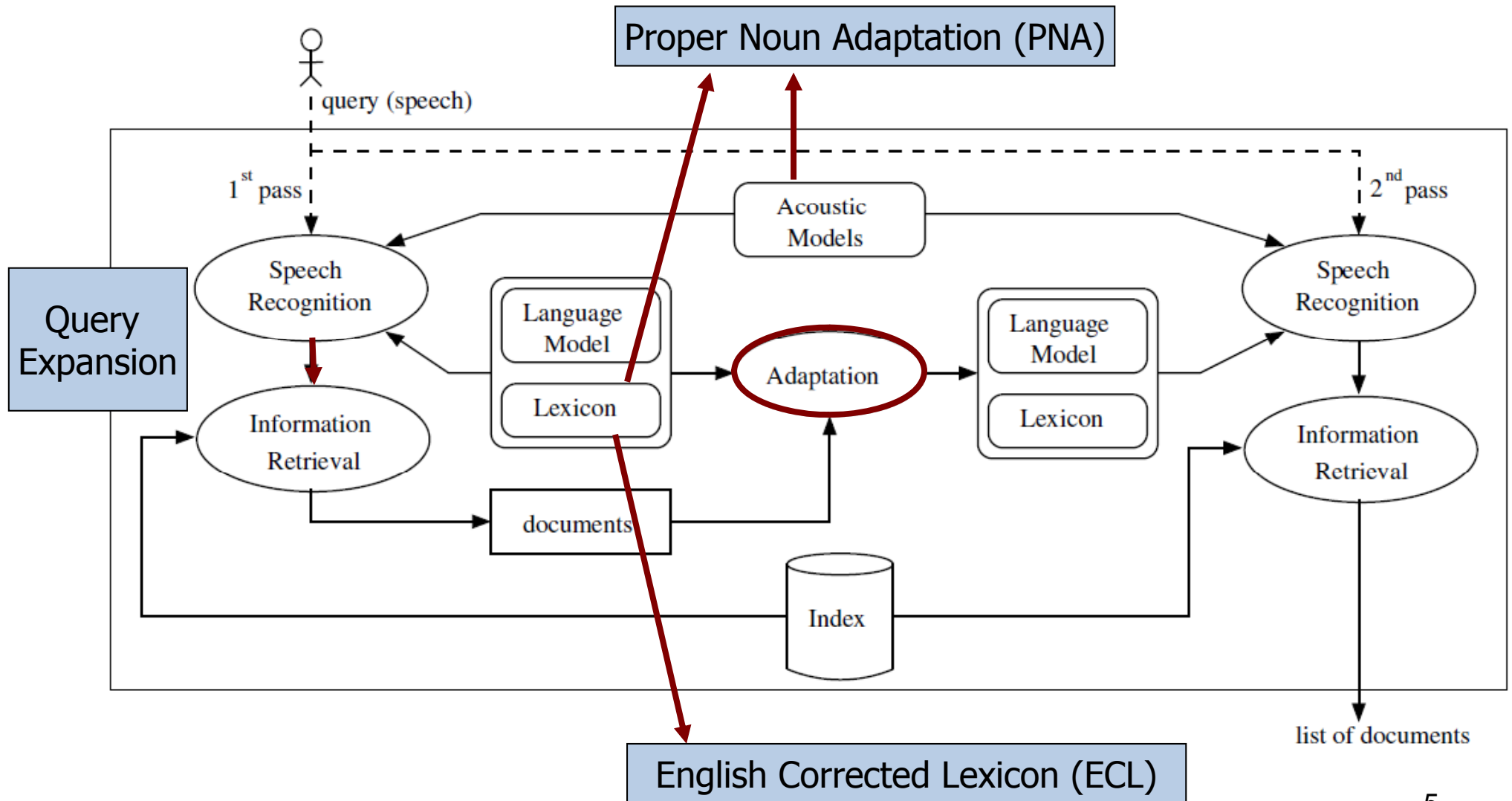
Background

- **Pronunciation variations** in Bahasa Indonesia reduce the word recognition accuracy (Variation sources: dialects, proper nouns, foreign words)
- Recognition errors especially for important terms (**proper nouns and foreign words**) reduce the performance of spoken query-based Information Retrieval (IR)



Error analysis of the baseline ASR system [D. P. Lestari, 2006]

System overview



Database development

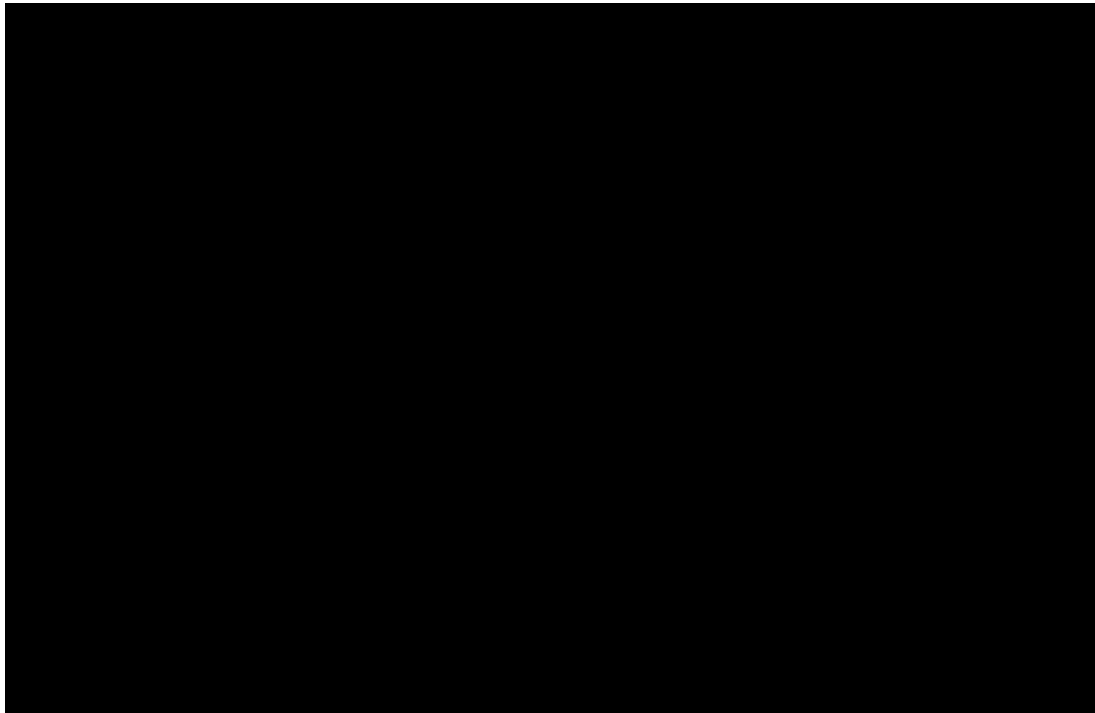
- Since there is no standard evaluation corpus for spoken query IR in Bahasa Indonesia, we developed a test set of spoken queries for experiments.
 - Documents and query databases are stored in the TREC format:
 - Collected documents : Newspaper, Magazine A, and Magazine B
 - Spoken queries: 5400 Indonesian spoken queries uttered by 20 speakers (11 males and 9 females)

Source	Speakers	Short	Medium	Long	Total
Newspaper	20	35	35	35	2100
Magazine A	20	35	35	35	2100
Magazine B	20	20	20	20	1200

- Exhaustive relevance judgments

Proper noun adaptation (PNA)

- Aim: to model the acoustic variations in uttering proper nouns
- Developing accurate pronunciations for proper nouns is difficult in many languages, including Indonesia



origins in the

PNA procedure

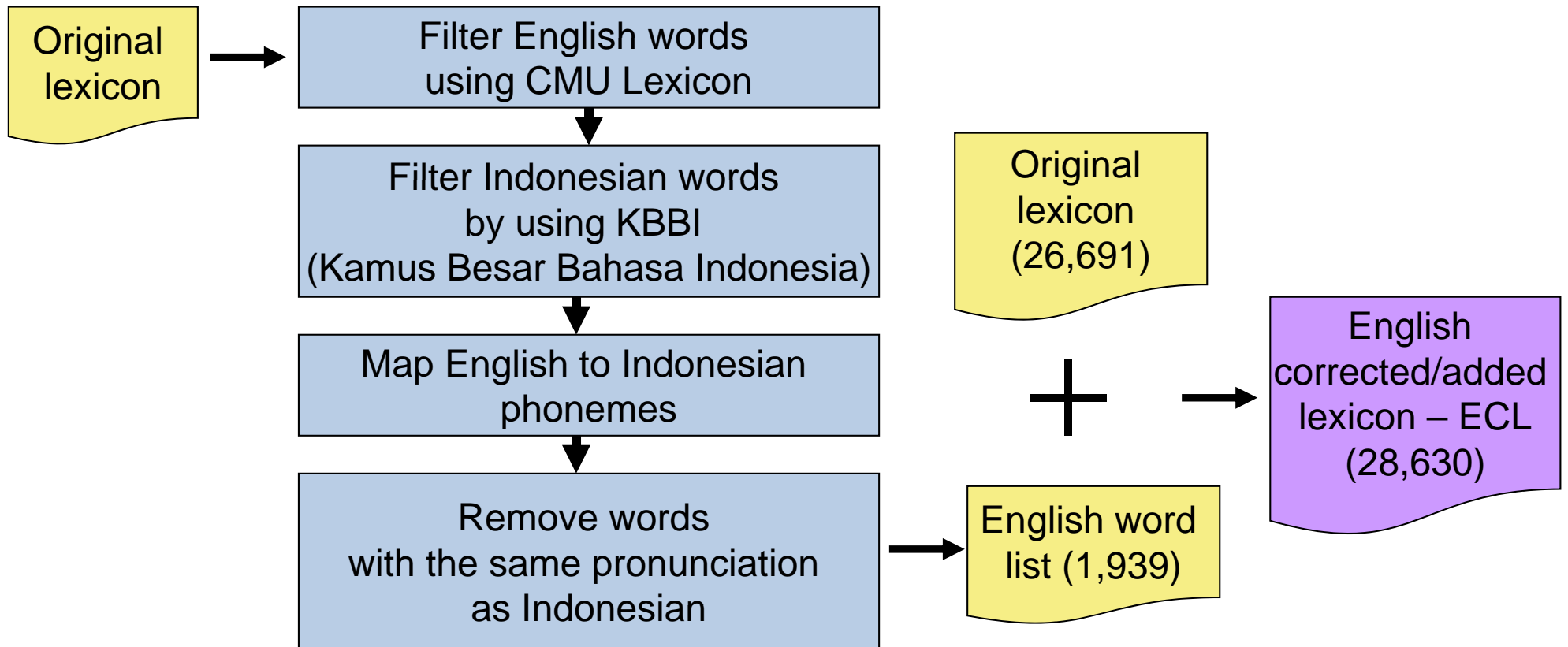
- Extract proper noun words from the speech corpus (14,840 words).
- Make proper noun specific triphone HMMs:
 - Supervised adaptation based on MLLR using eight regression classes
- Combine the baseline HMMs and the proper noun specific HMMs
- Add the proper noun pronunciation to the baseline lexicon (Using the proper noun dictionary provided in the Indonesian Standard Dictionary, we found 3,216 proper nouns in the baseline lexicon)

English to Indonesian phoneme mapping (EIPM)

- Add the pronunciation of English words to the Indonesian lexicon using rules modified from [S. Sakti, 2005]

Eng	Ind	Eng	Ind	Eng	Ind
aa	a	f	f	oy	oy
ae, eh	e	g	g	p	p
ah	e2	hh	h	r	r
ao	o	ih, iy	i	s	s
aw	aw	jh	j	sh	sy
ay	ay	k	k	t, th	t
b	b	l	l	uh, uw	u
ch	c	m	m	v	f
d, dh	d	n	n	w	w
er	e r*	ng	ng	y	y
ey	e y*	ao, ow	o	z, zh	z
n+y	ny	k+h	kh	ii	i*
ng-k	ng*	sha	sie2*		

EIPM procedure



ASR accuracy

	Baseline	PNA	PNA+ECL
Accuracy (%)	75.1	76.8	77.8

PNA: Proper noun adaptation

ECL: English corrected/adapted lexicon

IR experiments

- Compared the Vector Space Model (VSM)-based IR[G. Salton, 1975] and the Inference Network (IN)-based IR [H. R. Turtle, 1991] using the tf-idf weights
- MRR scores (%)

	VSM	IN
Base (1 best)	62.2	71.1
PNA (1 best)	63.6	72.0
PNA+ECL (1 best)	64.9	73.2
PNA+ECL (5 best)		71.8
PNA+ECL (5 best with occurrence weighting)		74.0
Text Query	77.3	80.5

Summary

- **Pronunciation variations** in Bahasa Indonesia mainly come from three sources: dialects, proper nouns, and foreign words.
- Proposed a **proper noun adaptation** based on MLLR to increase the proper noun recognition rate.
- Proposed rule-based **English-to-Indonesian phoneme mapping** to increase the English word recognition rate.
- **IN-based IR** outperforms the vector space model IR both for text queries and spoken queries.

2. Thai broadcast news (BN) LVCSR

Characteristics of the Thai language

พ.ต.ท.ทักษิณ ชินวัตร นายกรัฐมนตรี เปิดเผยถึงกรณีการ
ชุมนุมประท้วงการเจรจา เขตการค้าเสรี (เอฟทีเอ) ไทย-
สหรัฐฯ ครั้งที่ 6 วันที่ 9-13 ม.ค.นี้ ที่ จ.เชียงใหม่ และ
เรียกร้องให้นำรายละเอียดการเจรจาเข้าพิจารณาในสภา
ผู้แทนราษฎรก่อนว่า ไม่จำเป็น สภาไม่มีผู้เชี่ยวชาญและไม่มี
กฎหมายกำหนด

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

Development of Thai BN corpora

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

Lexical units for Thai LVCSR

- **Problem:** To find an optimal set of lexical units for LVCSR
 - **Word**
 - Manually defined and stored in a dictionary
 - OOV problem
 - Segmentation errors always occur when unknown words exist
 - **Pseudo-morpheme (PM)**
 - A syllable-like unit in Thai written text
 - Good coverage, but pronunciation depends on context
 - **Compound Pseudo-morpheme (CPM)**
 - Frequently adjoining PMs are combined to create CPMs
 - No dictionary is required to construct CPMs
- **Solution:** **CPM** unit is used in our system

CPM construction

- Frequently adjoining PMs are merged into CPMs based on the geometric average of the forward and backward bigrams:

$$M(w_i, w_{i+1}) = \sqrt{P_f(w_{i+1} | w_i) P_r(w_i | w_{i+1})}$$

- CPM lexicon has a high coverage within a limit of vocabulary size
- Lower acoustic confusability than PM
- Almost unique, context-independent pronunciation

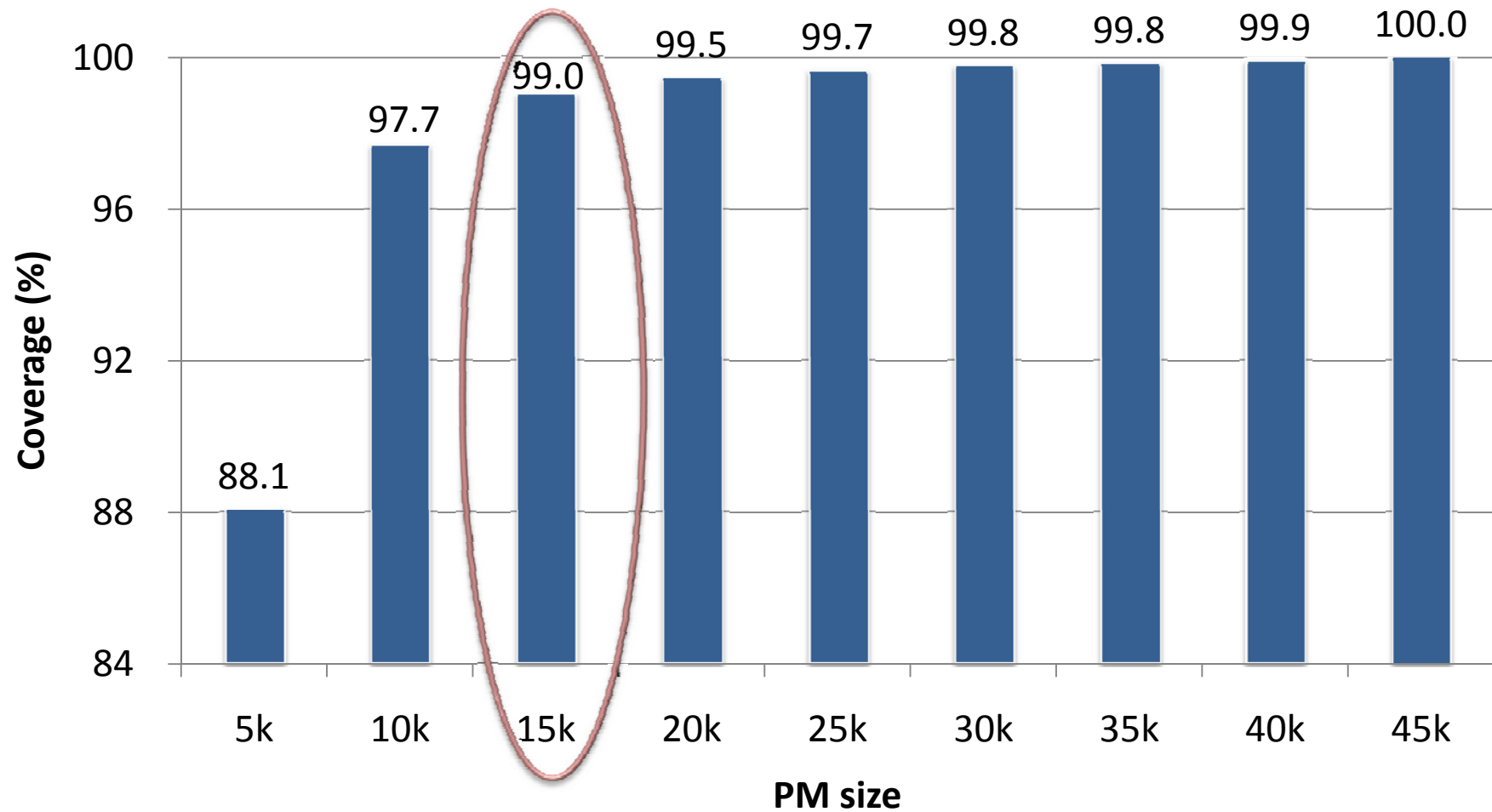
Examples of automatic merging

Text	นายโรเบิร์ตผู้นำคอมมูนิตีเชื่อว่า
Manual	นาย โรเบิร์ต ผู้นำ คอมมูนิตี เชื่อว่า Mr. Robert leader community believes that ...
PM	นาย โร เบิร์ต ผู้ นำ คอม มู นิ ตี เชื่ อ ว่า
Iteration 1	นาย โรเบิร์ต ผู้นำ คอมมู นิตี เชื่อว่า
Iteration 2	นาย โรเบิร์ต ผู้นำ คอมมูนิตี เชื่อว่า
Iteration 3	นายโรเบิร์ต ผู้นำคอมมูนิตี เชื่อว่า

Experimental conditions

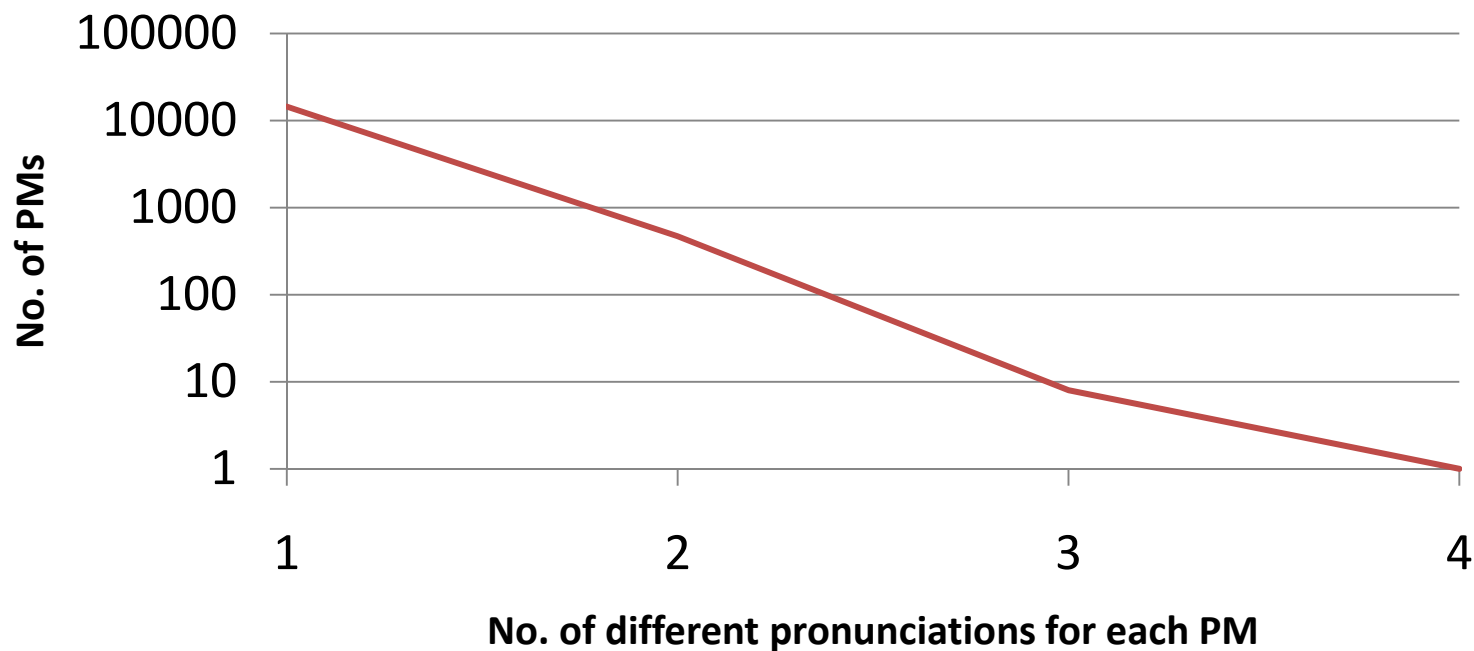
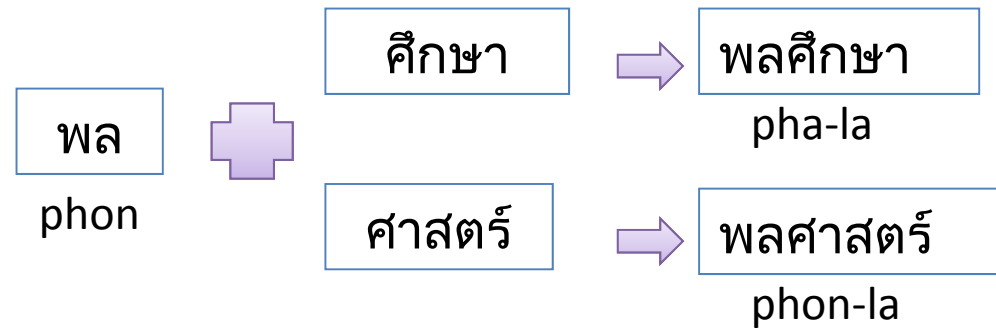
- **Acoustic model**
 - Gender-dependent acoustic models
 - 12 MFCCs, delta, and delta energy
 - Triphones, 1000 tied-states, 8 Gaussian mixtures
 - Read speech data: 40.3 hours by 68 male and 68 female speakers
- **Language model**
 - Various language models trained by using BN transcript and NP corpora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

The number of PMs is finite



* NP text containing around 141M PMs

Context-dependent PM pronunciations



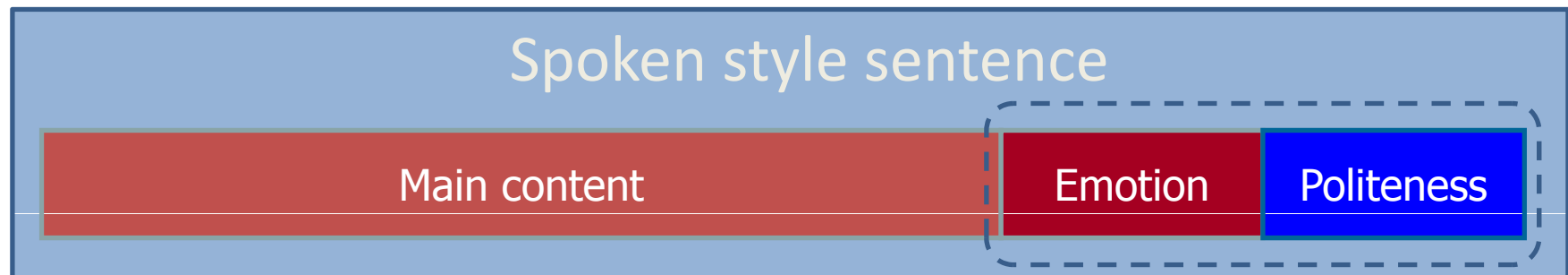
Comparison of results for the 3 lexical units

System	PM Error Rate (%)	Perplexity	OOV rate (%)
PM (5-gram)	26.6	206.4	0.5
Word (3-gram)	24.9	178.5	2.2
CPM (3-gram)	22.1	97.9	0.4

* Lexical units and LMs were built from an NP text corpus

Thai spoken style utterances

- Spoken-Style Ending Words (SSEW):
 - Words expressing **additional meaning or emotion**
 - Words showing **politeness**
 - Usually located at the end of the sentence



**Spoken-Style Ending Word
(SSEW)**

Rule-based speaking style classification

- Text is classified as
 - Spoken style text, when an SSEW appears in the content
 - Written style text, otherwise

Corpus	Written style	Spoken style
BN	57.0%	43.0%
NP	99.3%	0.7%

Differences of text styles

- Difference of text styles in BN and NP corpora

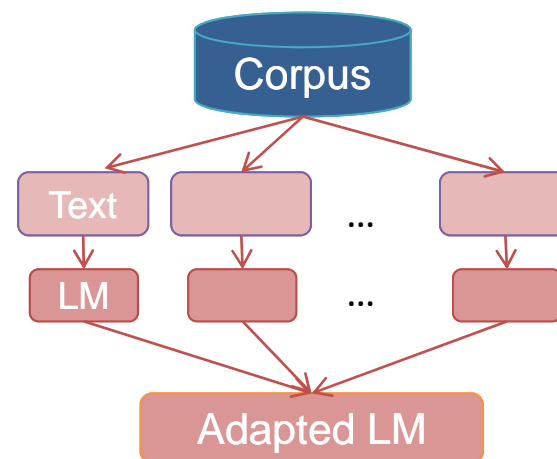
Test set type	LM	Perplexity	PM Error Rate (%)
Newspaper	NP	191.9	11.4
	BN	828.3	17.6
Broadcast news	NP	339.2	22.2
	BN	296.9	20.8

- Difference of speaking styles in Thai BN

Speaking style	LM	Perplexity	PM Error Rate (%)
Written style	NP	246.3	17.2
	BN	307.5	18.4
Spoken style	NP	795.2	35.8
	BN	270.5	27.2

Unsupervised LM adaptation framework

- Multi-pass recognition
- Text is clustered by
 - BN or NP styles
 - Speaking styles
 - Topics by using K-means algorithm
- A specialized LM is trained for each text cluster
- Adapted LMs are obtained by interpolating the specialized LMs



Various adaptation schemes


Adaptation Type	Text clustering by	Number of LM components	PM Error Rate (%)
Baseline	-	1	20.2
Style	Text source (BN and NP styles)	2	19.0
	Speaking style	2	19.2
	Text source + speaking style	4	18.2
Topic	Topic	8	19.5
Topic and Style	Topic + text source	16	18.0
	Topic + speaking style	16	18.5
	Topic + text source + speaking style	32	18.3

Summary

- A pioneering Thai BN speech recognition system was developed
- The first Thai BN speech and text **corpora** were developed
- Lexical units for Thai LVCSR were investigated and **Compound Pseudo-morpheme (CPM)** was found to be most suitable
- **Topic and style LM adaptation** for Thai BN recognition task was found to be effective

3. Vocabulary expansion through automatic abbreviation generation for Chinese voice search

Abbreviation problem in Chinese voice search

- **Abbreviations** are frequently used in spoken query for efficiency and simplicity.  **OOV problem**
- In Chinese, long named entities are frequently abbreviated using sub-words. The abbreviation process is more complex than English (like Georgia Tech for Georgia Institute of Technology).
- We focus on modeling variations of Chinese query abbreviation based on either word units or sub-word units.
 - No requirement for a huge amount of training data
 - Predicting variations for unseen entries in a statistical way
 - Covering sub-word segmentation variations

Proposed method

- Formalize abbreviation generation as a tagging problem and use **CRF (Conditional Random Field)** as the tagging tool.
 - Character based tagging
 - Little constraint
 - Rescoring by a length model and web information
- Perform **vocabulary expansion** by adding the automatically generated abbreviation candidates to the vocabulary.

Chinese abbreviation

- **Three formalization methods** (different from English acronyms)

- Reduction
- Elimination
- Generalization

Full-name	abbreviation	English explanation
招商 银行	招行	China merchants bank
清华 大学	清华	Tsinghua university
陆 军, 海 军, 空 军	三 军	army, navy and air force

- This research considers the abbreviations with 99% coverage which
 - select characters from the (original) full-name
 - preserve the character order in the full-name

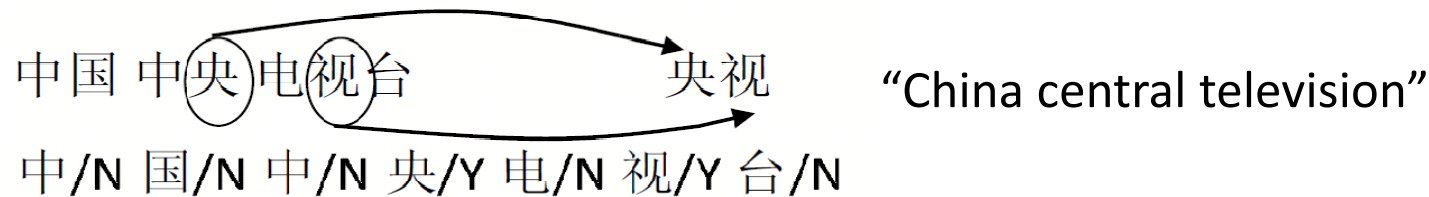
Abbreviation modeling by CRF

- CRF (Conditional Random Field) method
 - A statistical method that has been used in various NLP tagging tasks

$$P(L|C) = \frac{1}{Z(C)} \exp \left(\sum_{t=1}^T \sum_k \lambda_k f_k(l_t, l_{t-1}, C, t) \right)$$

in which, $C = c_1 c_2 \dots c_T$ is an observation sequence, $L = l_1 l_2 \dots l_T$ is a label sequence; $f_k(l_t, l_{t-1}, C, t)$ is the k^{th} feature function, λ_k is the corresponding weight, and $Z(C)$ is a normalization factor.

- Formalizes abbreviation generation as a tagging problem.

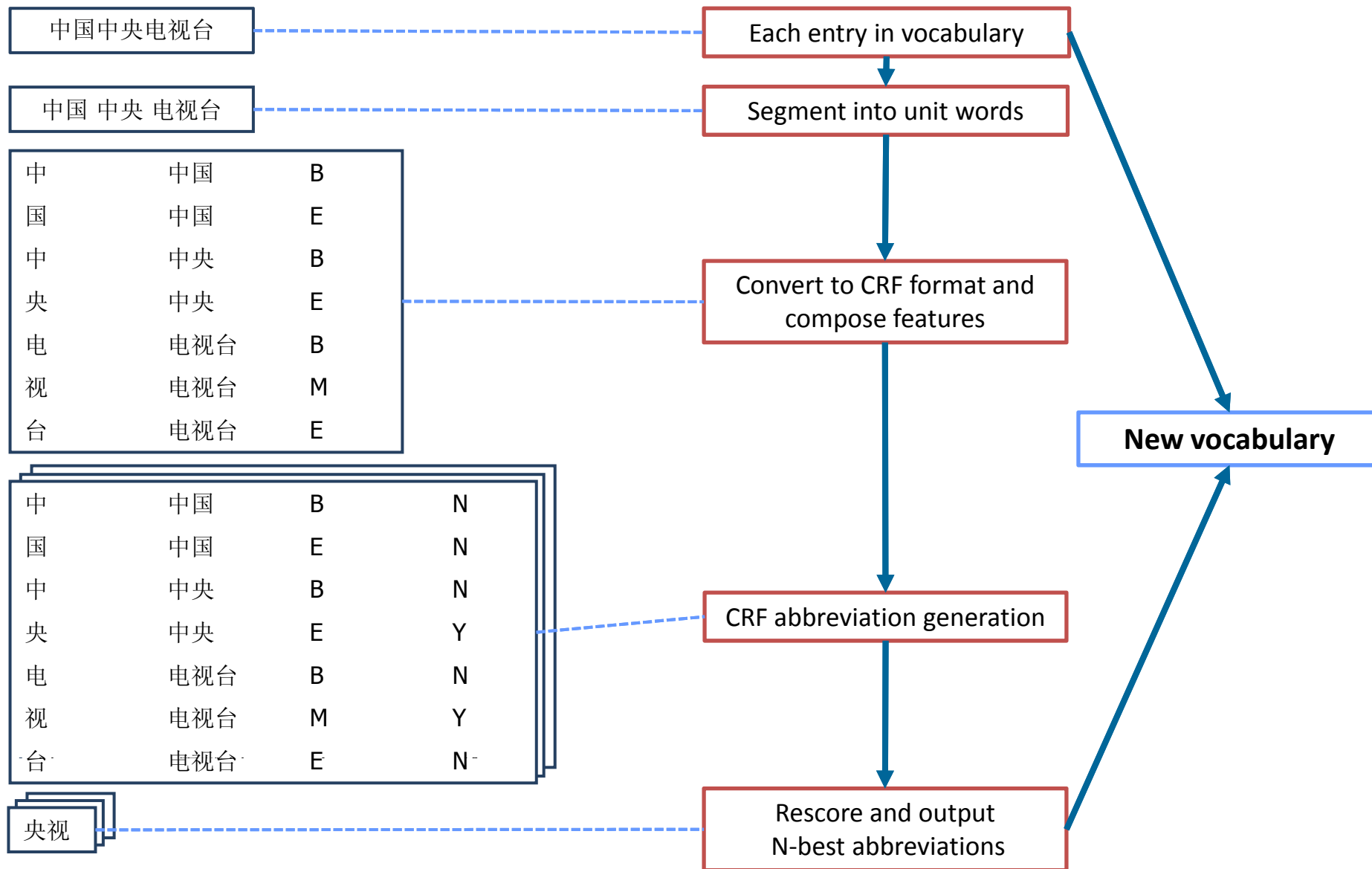


Feature selection

- Pre-processing
 - Segment the full-name into unit words
 - Using a 2-tag CRF segmenter trained by “Penn Chinese Treebank” corpus
- Features for CRF:
 - 1. Current character
 - 2. Current word
 - 3. Position of the current character in the current word
 - 4. Combination of the above features 2 and 3

System structure

China central television



Incorporating a length model

- Abbreviation process:
 - Model length mapping relation as discrete probabilities $P(M|L)$, in which L is the length of a full-name and M is the length of an abbreviation.
 - Re-rank the N-best results of CRF by the length model.
- Considering the abbreviation process as a product of two steps:
 - Model the length of abbreviation according to the length model
 - Model the abbreviation, given the length and the full-name

$$\tilde{A} = \underset{A}{\operatorname{argmax}} P(A, M|F, L) \approx \underset{A}{\operatorname{argmax}} P(M|L)P(A| F, M)$$

Use Bayesian rule

$$P(A|F, M) = \frac{P(A, M|F)}{P(M|F)} = \frac{P(A, M|F)}{\sum_{\text{Length}(A')=M} P(A', M|F)} = \frac{P(A|F)}{\sum_{\text{Length}(A')=M} P(A'|F)}$$

F : fullname

L : fullname length

A : abbreviati on

M : abbreviati on length

Improving by a web search engine

- Using a search engine:
 - “abbreviation” + “full-name” as a query
 - Get the number of hits, #hit
- Criterion for re-ranking
 - $(\text{Probability}_{\text{by CRF+Length}}) \times \text{\#hit}$
- In practice
 - We re-rank the top-30 results from the previous step
 - To limit the number of access to web search engine

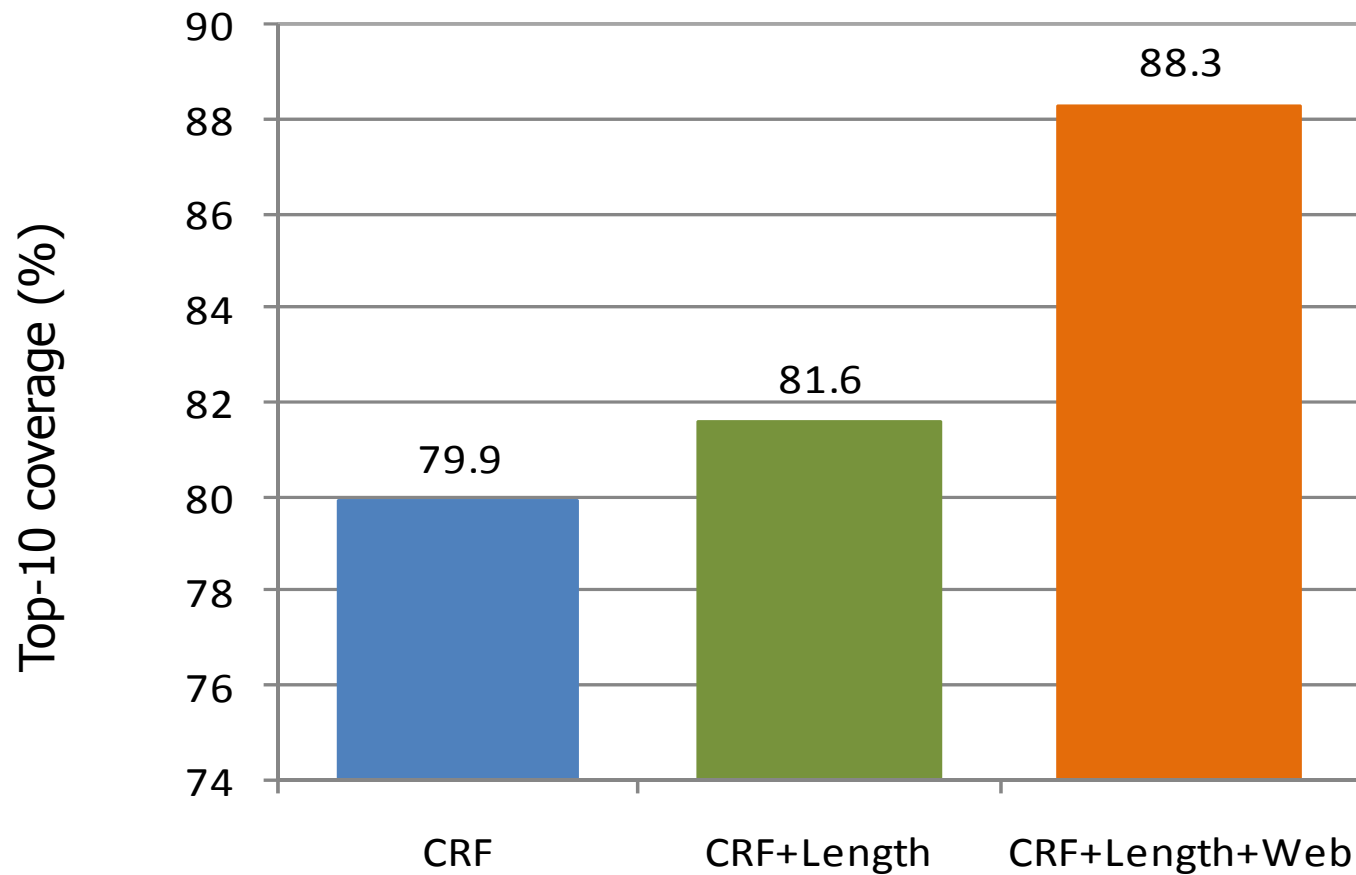
Experiments on abbreviation modeling

- Database
 - “Modern Chinese abbreviation dictionary”
 - Wikipedia
- For experiments
 - Training data: 1298 pairs
 - Test data: 647 pairs
- Length statistics of training data
- Evaluation metric
 - Abbreviation coverage in top-10 candidates

length of full-name	length of abbreviation					sum
	2	3	4	5	>5	
4	107	1	0	0	0	108
5	89	140	0	0	0	229
6	96	45	46	0	0	187
7	60	189	49	16	0	314
8	48	29	60	3	6	146
9	10	47	35	12	2	106
10	18	11	29	8	6	73
others	21	43	38	17	14	133
average length of the full-name					7.27	
average length of the abbreviation					3.01	

Experimental results with different methods

(Test set: 647 abbreviations)



Experiments on voice search

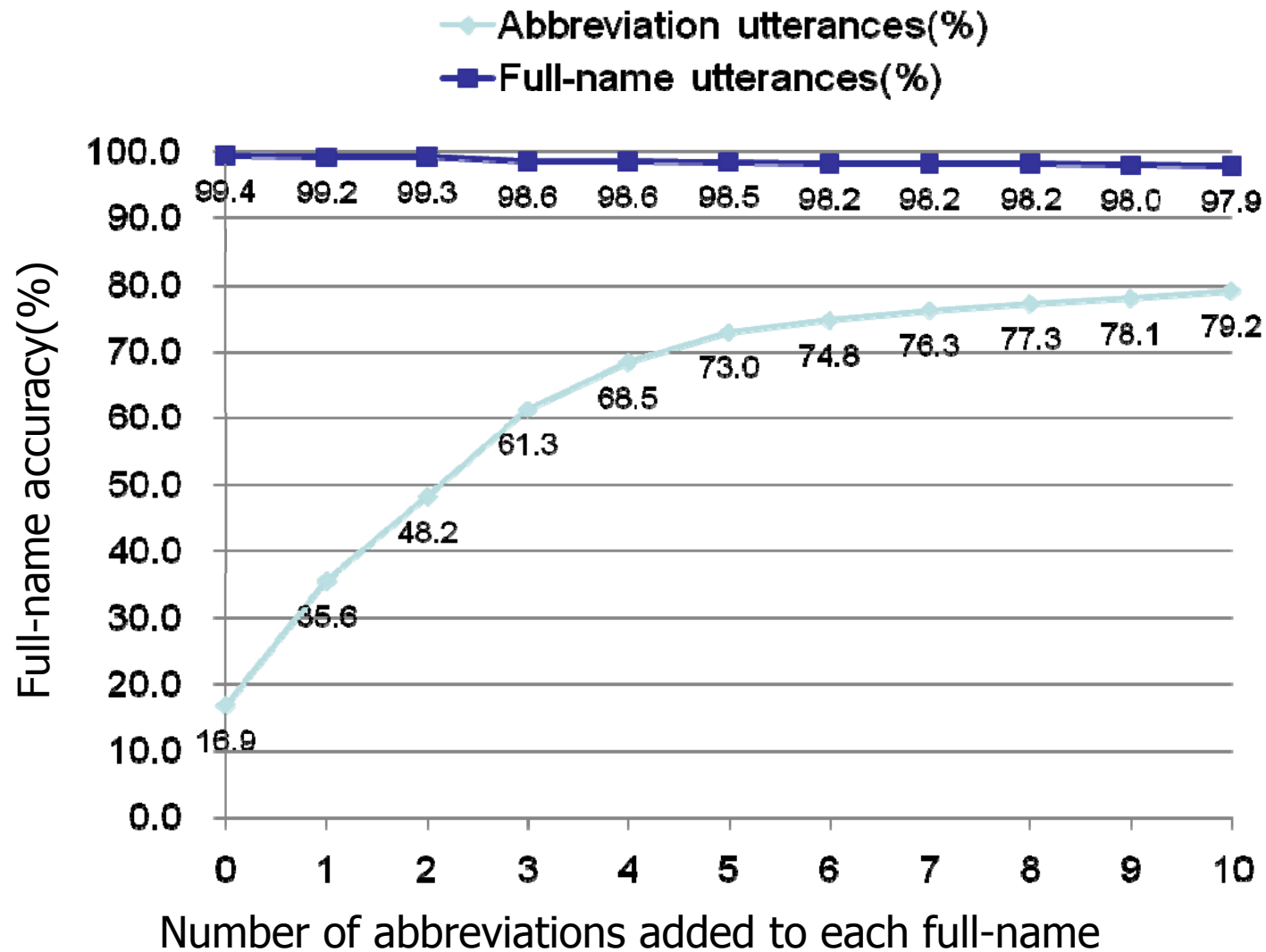
- Data collection -

- 400 named entity full-names from the test set used in our pervious experiment
- 20 subjects X 80 entities (1600 entities)
 - Each entity has 4 subjects to generate abbreviations
- Procedure:
 - 1. A full-name is displayed.
 - 2. The subject writes down an abbreviation and reads it.
 - 3. If there is another abbreviation, go to 2.
 - 4. Proceed to the next full-name and go to 1.
- Output: 783 unique abbreviations, about 2 for each entity in average

Voice search experiments

- **Acoustic model**
 - Toneless triphone models, trained by 150 hours Chinese speech data
- **Vocabulary**
 - Full-names (400) + automatically generated abbreviations (0-10 for each full-name)
- **Test data**
 - Abbreviation utterances (generated by subjects)
 - ✓ 2204 utterances by 20 speakers, 783 unique abbreviations
 - Full-name utterances
- **Evaluation metric**
 - Search accuracy for correct full-names using ASR output

Results of voice search experiments

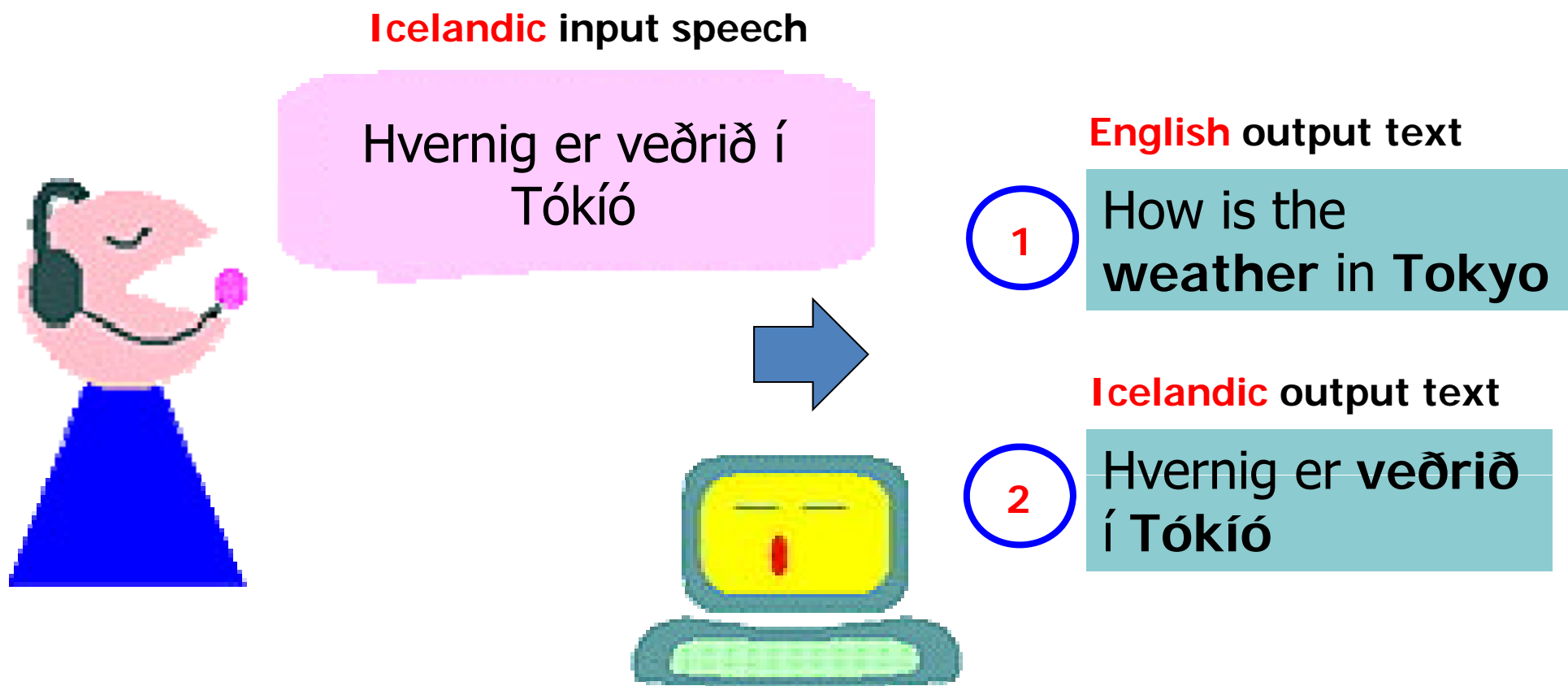


Summary

- The **CRF** works well in generating abbreviations for organization names.
- Both the **length model** and the **web search engine** further improve the performance.
- **Vocabulary expansion** through adding the generated abbreviations significantly improves voice search accuracies.

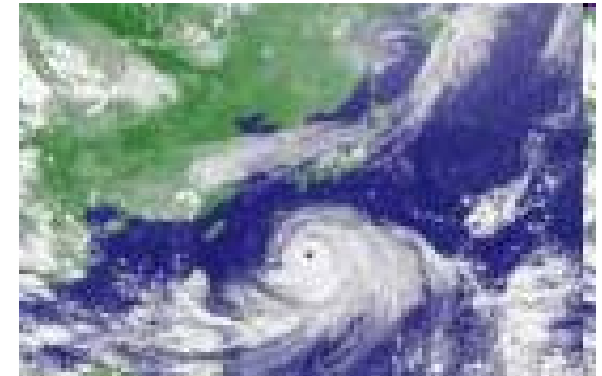
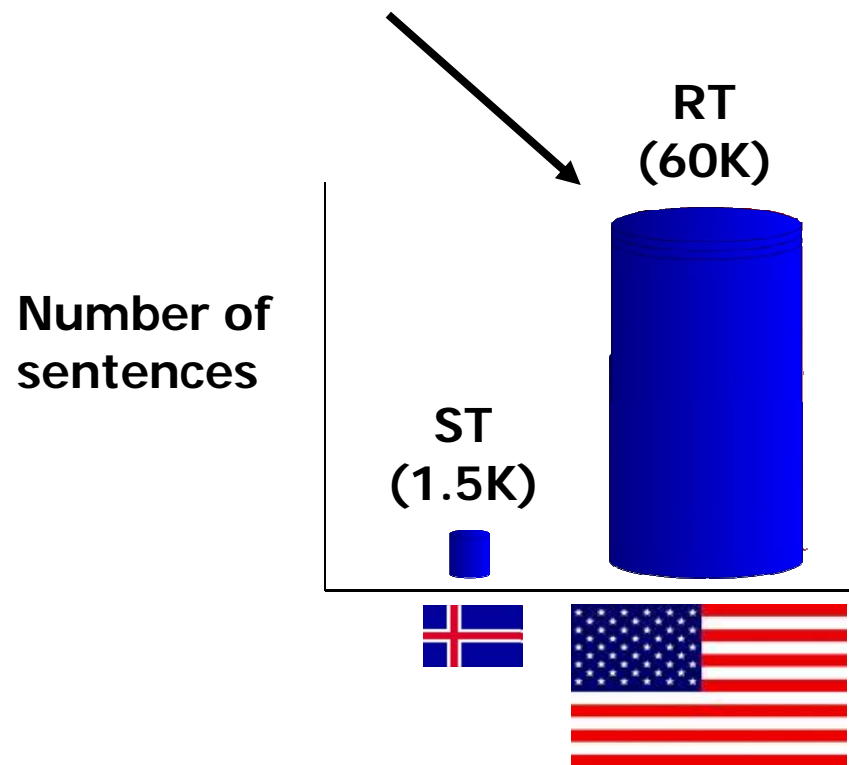
4. WFST-based Icelandic ASR using machine translation

ASR for resource-deficient languages: Icelandic case (two setups)



Training text data

- The Jupiter corpus
(a weather information corpus developed by MIT) was used as a  English rich corpus



Weather information domain

RT: Rich text
ST: Sparse text

Method 1

(Icelandic to English)

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

Traditional format

$$\tilde{T} = \operatorname{argmax}_T \max_W P(O|W) P(W|T)^{\lambda_{ST}} P(T)^{\lambda_{RT}}$$

WFST format

Icelandic
speech



$$H \circ C \circ L \circ G_{ST} \circ Tr \circ G_{RT}$$



English
Text

Method 2

(Icelandic to Icelandic)

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

Traditional format

$$\widetilde{W} = \operatorname{argmax}_W \max_T P(O|W) P(W|T)^{\lambda_{ST}} P(T)^{\lambda_{RT}}$$

WFST format

Icelandic
speech



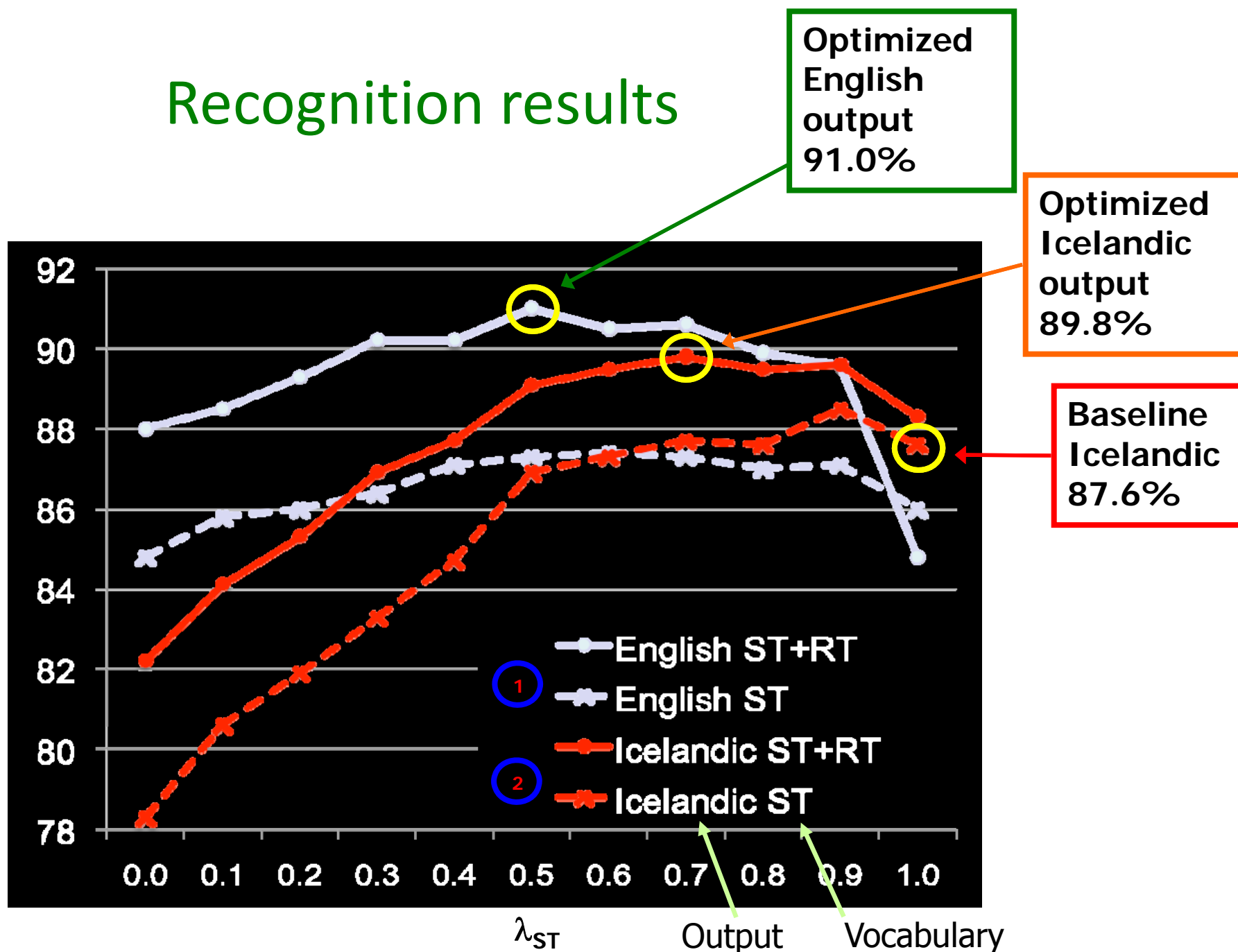
$$H \circ C \circ L \circ G_{ST} \circ \pi(T_r \circ G_{RT})$$



Icelandic
Text

Recognition results

Keyword
Accuracy
Rate (%)



Summary

- Icelandic language was selected as a resource deficient language.
- A language model (LM) trained on a translated text from a resource rich language (English) was interpolated with a sparsely trained LM in the target language.
- This method was implemented in a WFST network to output on the fly the translation of the recognized speech.
- This strategy reduces the development time of the speech recognition system if the backend system is already available for the resource rich language.

Overall summary

1. Adaptation to pronunciation variations in **Indonesian** spoken query-based information retrieval
2. **Thai** broadcast news (BN) LVCSR
 - Lexical units for Thai LVCSR
 - Topic and style-adapted language modeling for Thai broadcast news ASR
3. Vocabulary expansion through automatic abbreviation generation for **Chinese** voice search
4. WFST-based **Icelandic** ASR using machine translation