EECS E6870: Lecture 12: Special Topics – Spoken Term Detection

Stanley F. Chen, Michael A. Picheny and Bhuvana Ramabhadran
IBM T. J. Watson Research Center
Yorktown Heights, NY 10549

stanchen@us.ibm.com, picheny@us.ibm.com,
bhuvana@us.ibm.com

December 1, 2009

What is it?

- **Search for specific terms in large amount of speech content (key word spotting)**
- **Enable open vocabulary search**
- **Applications:**
  - Call monitoring
  - Market intelligence gathering
  - Customer analytics
  - On-line media search

Something like this………


http://www2.criticalmention.com - Critical TV 3.0 - Microsoft Internet Explorer
CRITICAL MENTION
Back
Forward
User: lsimpson2 Customer: ! CM New Feature
Log off
Home
Alerts / Searches
Library
My Account
Help
Search
Advanced Search
Expanded Search Results
Back to Search Results
Minutes Back 10 5 1
Reset Timeline
Minutes Forward 1 5 10
ALJZ (ALJZ-T) Supervolcanoes
International Programming,
Category: Special Nature
Sep. 15, 2005 5:47 AM EDT
Local Broadcast Time: 5:47 AM EDT
being away from the battel field makes it hard to predict who is right and who is worng, but i think they should be involved in building thier country and writing the constitution and the whole political movement in iraq.reporter: so do you think its better to be involved in the peaceful movement? yes this is beneficial
Right-click here to download video
ALJZ (ALJZ-T) Supervolcanoes
International Programming,
Category: Special Nature
Sep. 15, 2005 5:48 AM EDT
Local Broadcast Time: 5:48 AM EDT
what do yopu think about the conflict of cultures? i think that there are some people who try to overspeak about this issue but let me tell you that islam is against that idea and islam is an open-minded religon and it tolerates the difference
Right-click here to download video
Stopped
Play
Pause
Stop
Generate URL
Order: CD Transcript Gallery
E-Mail: Video Clip & Text Text
Copy selected video to Library:
- Choose a folder -
Submit
View Sample Video Gallery
Go Back to Search Results
Shorten/Lengthen Clip
[Customer Service Live Chat]
Session Details
Log off
Internet
start
Task: training sessio...
http://www2.critical...
Inbox - Microsoft Ou...
Presentation1
Critical Mention CEN...
10:41 AM

Historically…….

- Keyword spotting (KWS)
  - In the 90s….
    - Use of filler models (parallel set of phone HMMs)
    - Likelihood ratio comparisons
    - Phone lattices for spoken document retrieval
    - Two step approach
      - Coarse step: identify candidate regions quickly
      - Detailed step: Better models to zero in on region of interest

- Phone decoding and its various flavors
- LVCSR

Historically…….

- Unreliable transcriptions: high error rate in one best transcripts
    - Search on lattices and/or confusion networks (CN)

- Efficient indexing and search algorithms
    - General Indexation of Weighted Automata [Saraclar 2004, Allauzen et al., 2004]
    - Posting list [JURU/Lucene] [Carmel et al. 2001, Mamou et al. 2007]

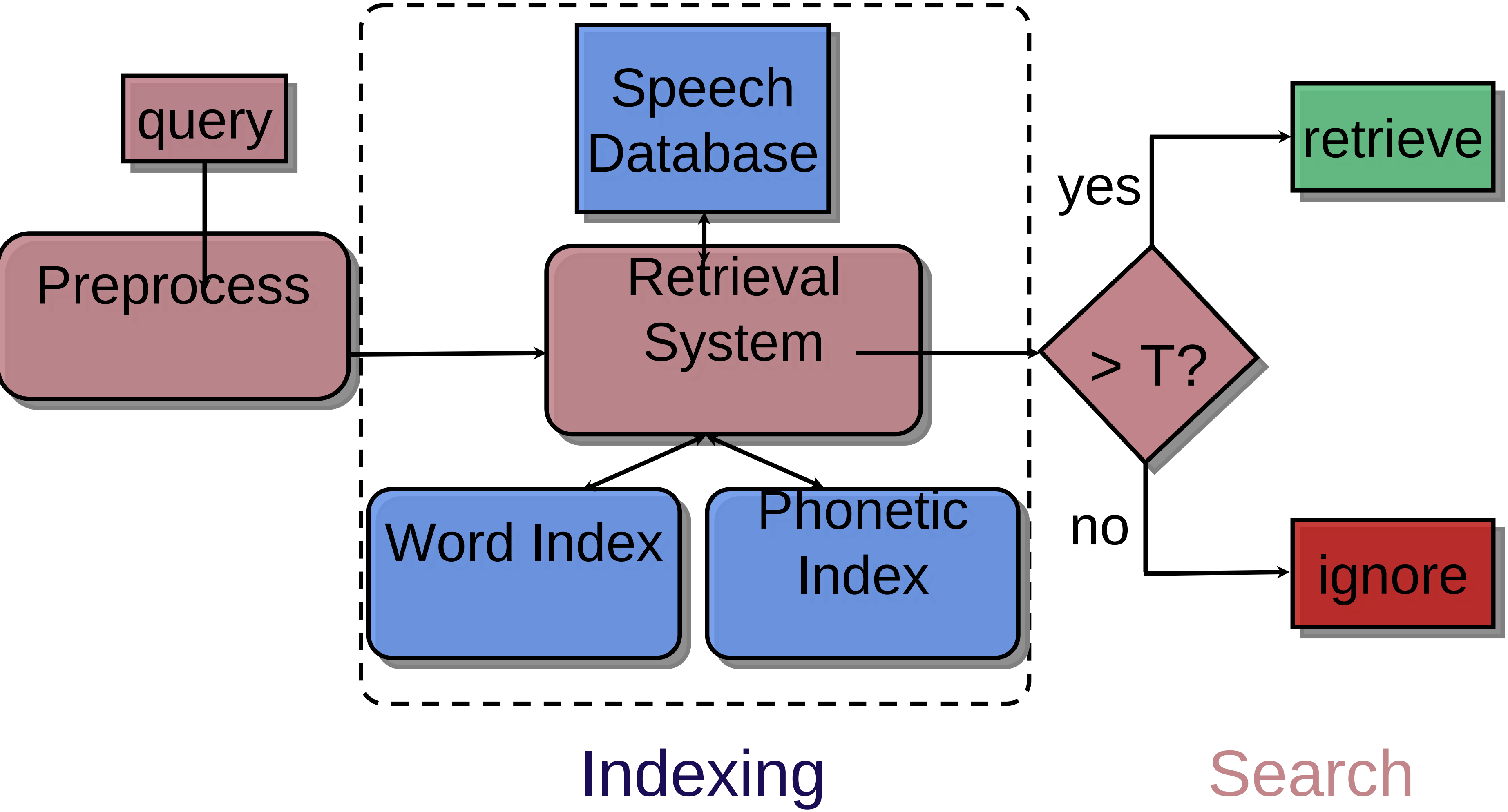
- Out Of Vocabulary queries: information bearing words
    - OOV pronunciation modeling [Can et al. 2009, Cooper, et al, 2009]
    - Search on subword decoding [Saraclar and Sproat 2004, Mamou et al, 2007, Chaudhari and Picheny, 2007]

Out of Vocabulary Terms

- ASR vocabulary might not cover all words of interest
  - **Information bearing words**
  - **Loss of context impacts word error rate**
  - **Special interest for spoken term retrieval**
- Challenges in OOV detection and recovery
  - **Rare foreign terms with a diverse set of pronunciations**
  - **Confusability with similar sounding in-vocabulary term**
  - **Language model information is missing**

Representing and detecting OOV terms

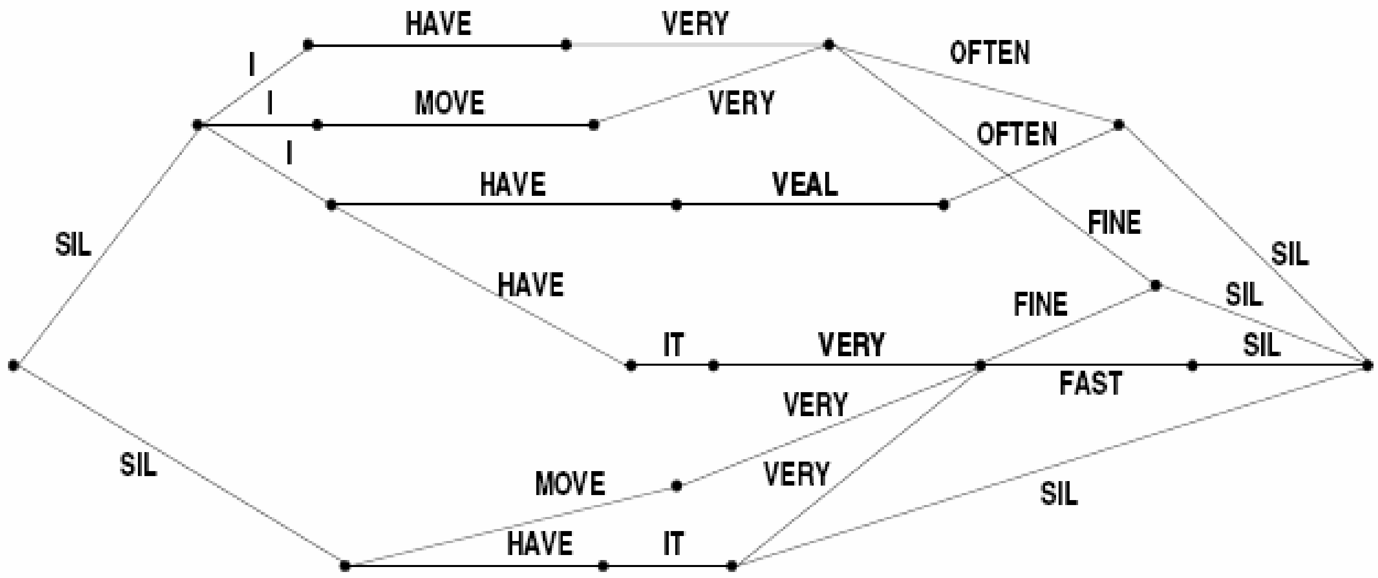
- **Use a combination of word and subword units** :
  - Identify set of words and subword units (fragments) for good coverage
  - Represent LM text as a combination of words and fragments
  - Build a Hybrid Language Model and Lexicon
  - Acoustic models for hybrid system are the same as word-based LVCSR system
- **Example** :
  - < s > THE WORKS OF ZIYAD HAMDI WERE RECENTLY AUCTIONED< =s >
  - < s > THE WORKS OF Z_IY Y_AE_D HH_AE_M D_IY WERE RECENTLY AUCTIONED < =s >


query
Preprocess
Speech
Database
Retrieval
System
Word Index
Phonetic
Index
yes
> T?
no
retrieve
ignore
Indexing
Search

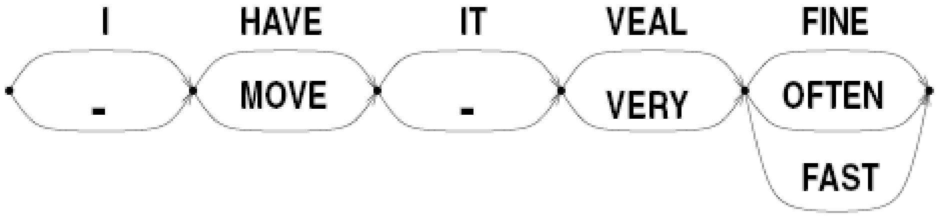
What speech Recognition output structures do we index?

- **1-best : I HAVE IT VEAL FINE**

- **Lattice:**



- **Word Confusion networks (WCN):**



Evaluation Metrics

- The basic idea is to count misses and false alarms for each query and to average this number across all queries
  - F-measure: Trade-off between Precision and Recall
  - Number of False Alarms per hour
  - In a task like distillation in GALE, false alarms may not matter as long as the first page of results contains at least an entry on what you are looking for…
  - Average Term Weighted Value: Weighted average of misses and false alarms

Indexing Architectures

- **JURU/Lucene :**
  - Extension of information retrieval methods for text (text-based search engine)
  - Use posting lists to store time , probabilities and index units
  - Compact representation but not very flexible
- **Transducer based :**
  - Represent indices as transducers
  - More flexible at the cost of compactness

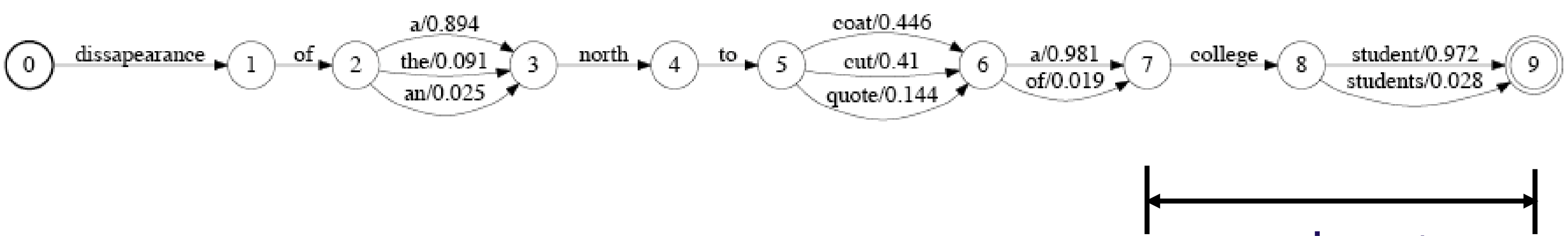
What can you do with an FST-based indexing system?

- **Allows us to search for complex regular expressions**

[healthcare 0.6, health care 0.4] [reform 0.8, plan 0.2]

- **Easy to do fuzzy matching**

- **We can search using audio snippets: query-by-example (QbyE)**



snippet

NIST Spoken Term Detection Evaluation

- Detection Task
  - Count misses and false alarms for each query
  - Average across all queries

- Broadcast News
- Telephone Speech
- Conference Meetings

- Actual Term-Weighted Value (ATWV)

$$ATWV = 1 - \frac{1}{N_{terms}} \sum_{t \in terms} (P_{miss}(t) + \beta \cdot P_{fa}(t))$$

$$P_{miss}(t) = 1 - \frac{N_{corr}(t)}{N_{true}(t)} \qquad P_{fa}(t) = \frac{N_{spurious}(t)}{Total - N_{true}(t)}.$$

B=1000, False alarms are heavily penalized

Actual Term Weighted Value [NIST STD 2006 Evaluation Plan]:

$$ATWV = 1 - \frac{1}{Q}\sum_{Q}^{q=1} P_{miss}(q) + \beta P_{FA}(q)$$

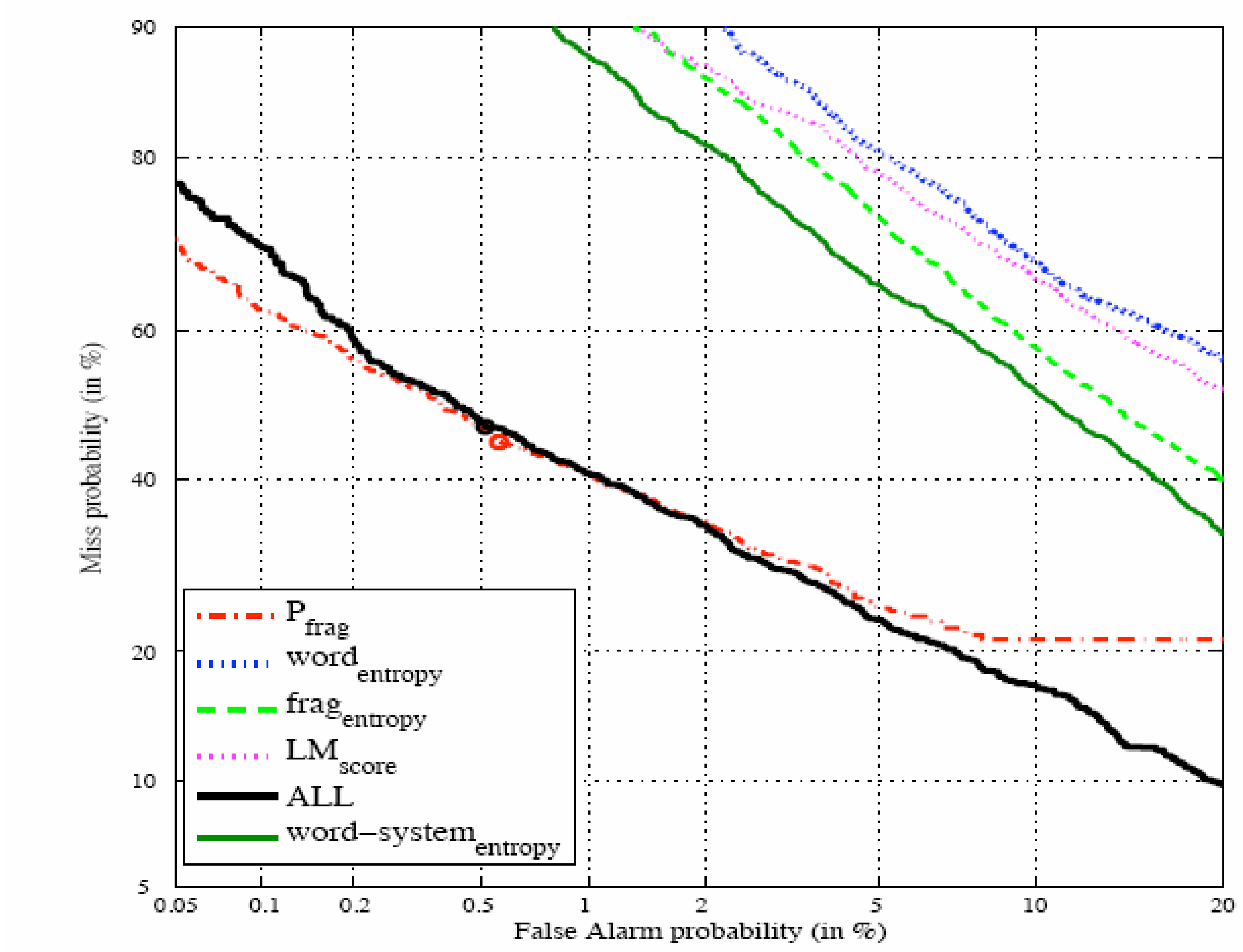
$$P_{miss}(q) = 1 - \frac{N_{corr}(q)}{N_{true}(q)} \qquad P_{FA}(q) = \frac{N_{spurious}}{T - N_{true}(q)}$$

$$\begin{aligned}
Q &= \text{number of queries} \\
N_{true} &= \text{occurrences in reference} \\
N_{spurious} &= \text{spurious instances retrieved} \\
N_{corr} &= \text{correctly retrieved instances} \\
\beta &= \text{user defined parameter, in STD 06 Eval } \beta = 999.99 \\
T &= \text{seconds of audio (secs)}
\end{aligned}$$

Word-Fragment Hybrid systems

- **Posterior probability of fragments in a given region is a good indicator of presence of OOVs**
- **Hybrid systems represent OOV terms better in phonetic sense then pure word systems or pure phonetic systems**

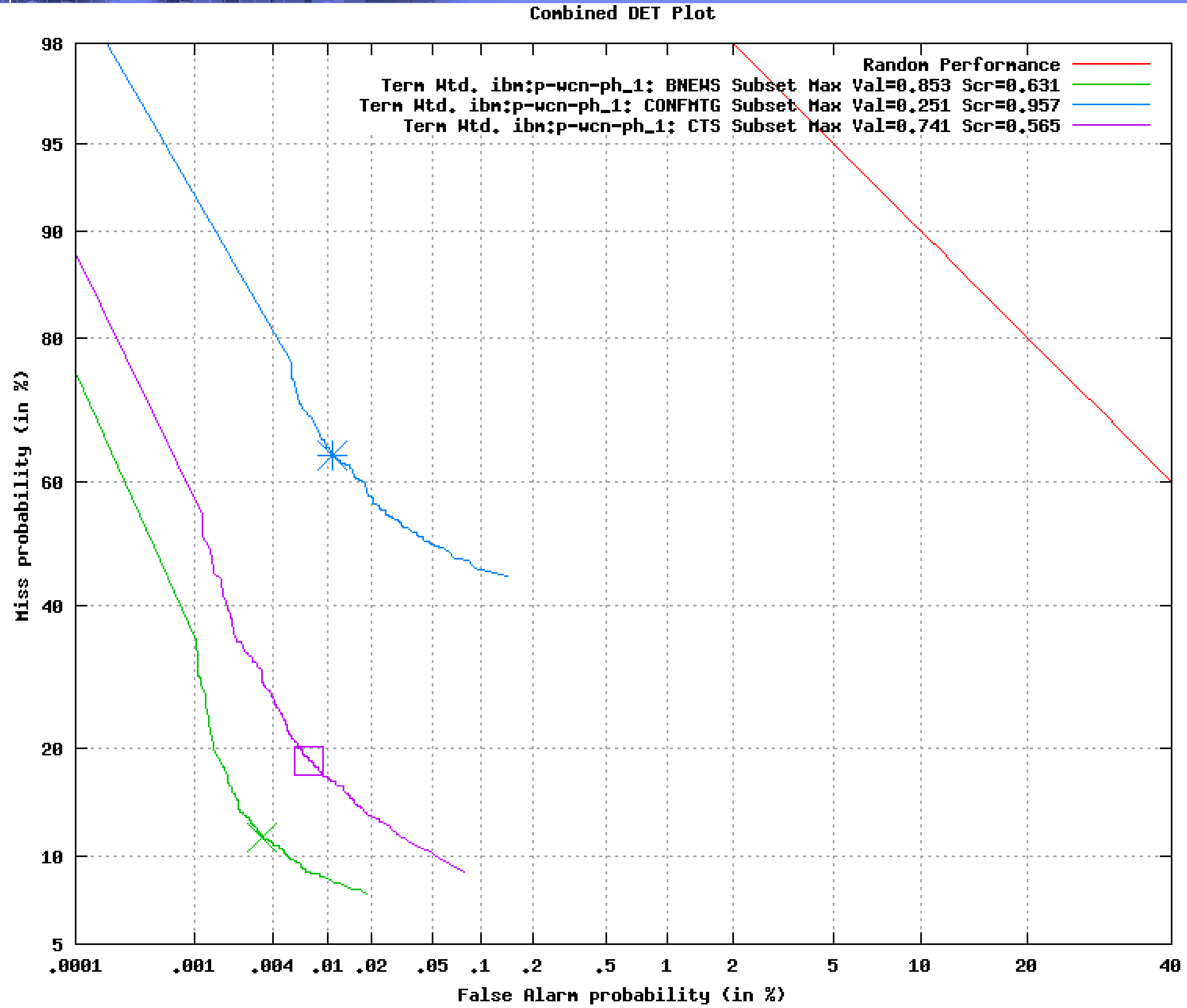
OOV Detection with hybrid systems



NIST 2006 Evaluation (English)

	system	BN	CTS	CONFMTG
TWV	Dry-Run P	0.8498	0.6597	0.2921
ATWV	Eval P	0.8485	0.7392	0.2365
MTWV		0.8532	0.7408	0.2508
ATWV	Eval C1	0.8485	0.7392	0.0016
MTWV		0.8532	0.7408	0.0115
ATWV	Eval C2	0.8293	0.6763	0.1092
MTWV		0.8293	0.6763	0.1092
ATWV	Eval C3	0.8279	0.7101	0.2381
MTWV		0.8319	0.7117	0.2514

- Retrieval performances are improved using WCNs, relatively to 1-best path.
- Our ATWV is close to the MTWV; we have used appropriate thresholds for pruning bad results.

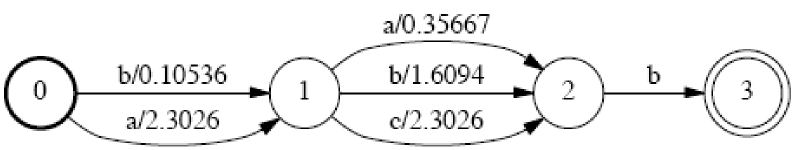

Combined DET Plot
Random Performance
Term Wtd. ibm:p-wcn-ph_1: BNEWS Subset Max Val=0.853 Scr=0.631
Term Wtd. ibm:p-wcn-ph_1: CONFMTG Subset Max Val=0.251 Scr=0.957
Term Wtd. ibm:p-wcn-ph_1: CTS Subset Max Val=0.741 Scr=0.565
98
95
90
80
60
40
20
10
5
Miss probability (in %)
.0001
.001
.004
.01
.02
.05
.1
.2
.5
1
2
5
10
20
40
False Alarm probability (in %)

WFST-based indexing

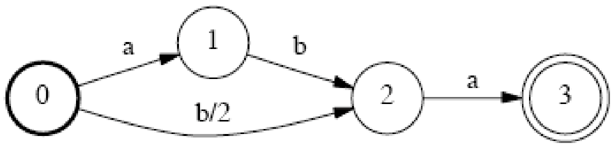
Recipe: preprocess lattices, build index, search

– Preprocess:

(1)



(2)

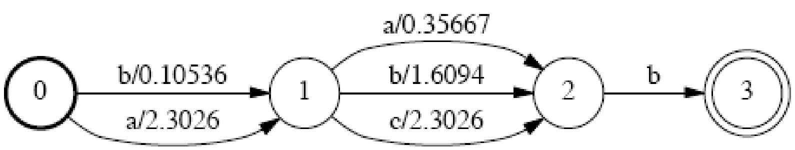


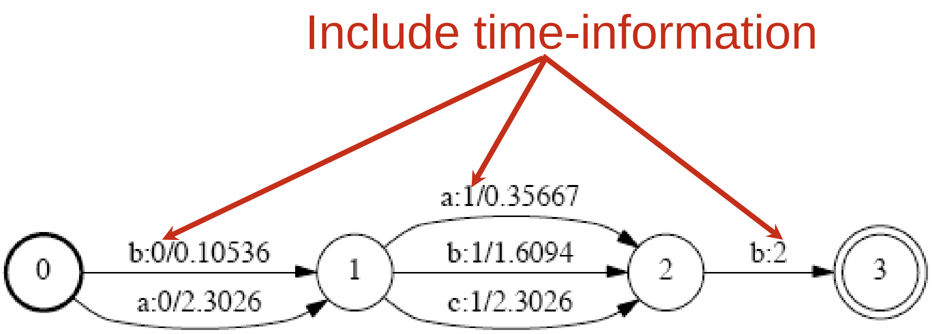
WFST-based indexing

Recipe: preprocess lattices, build index, search

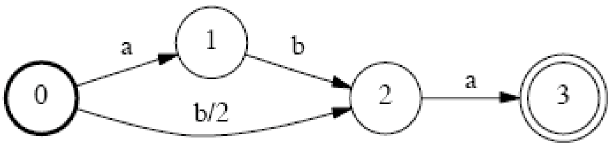
– Preprocess:

(1)





(2)



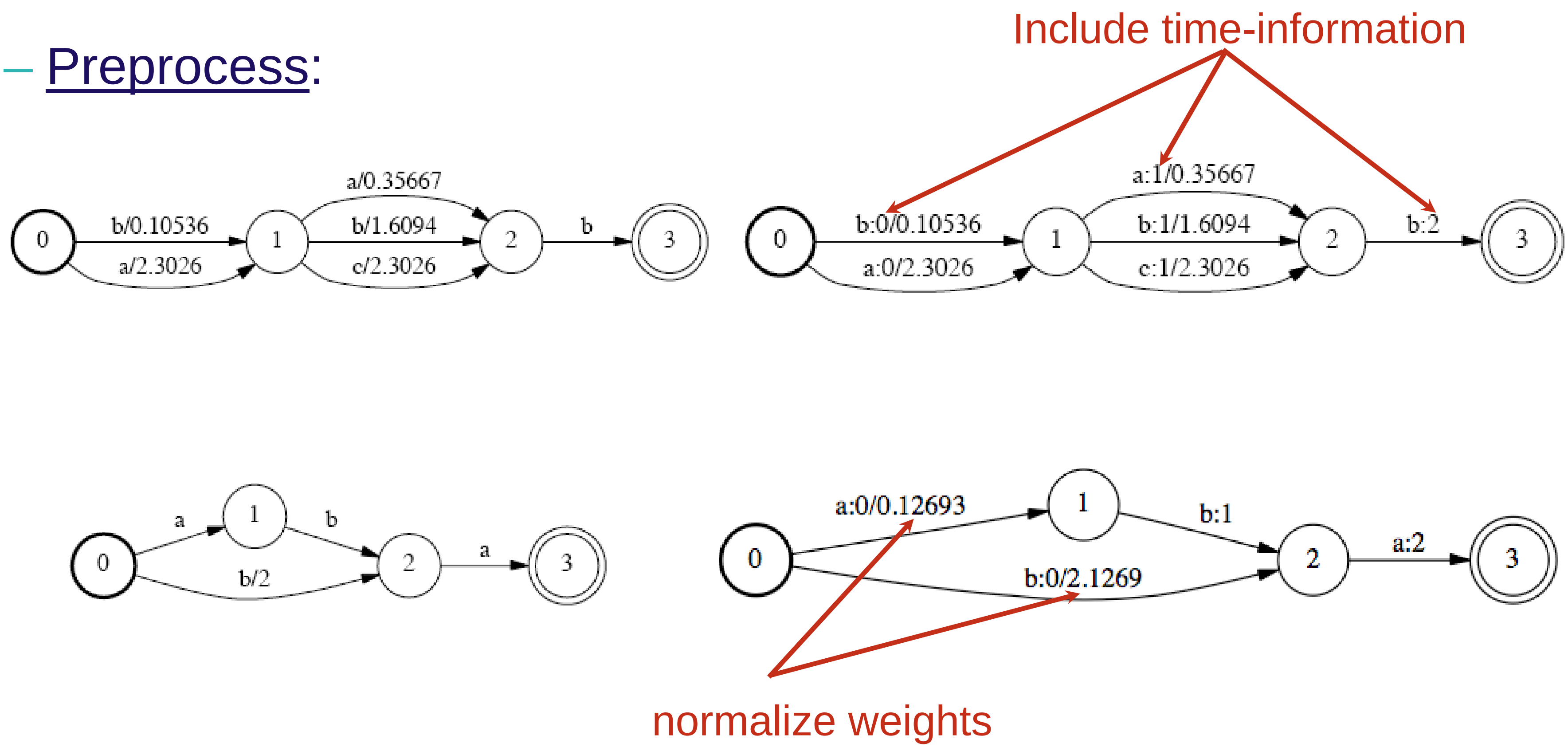
An Example: preprocess

Recipe: preprocess lattices, build index, search

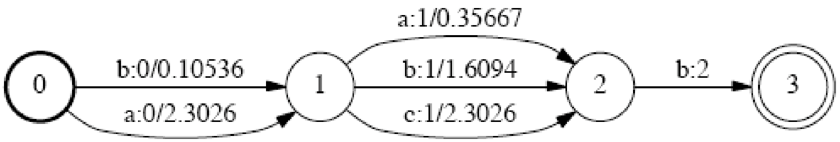
– Preprocess:

(1)

(2)

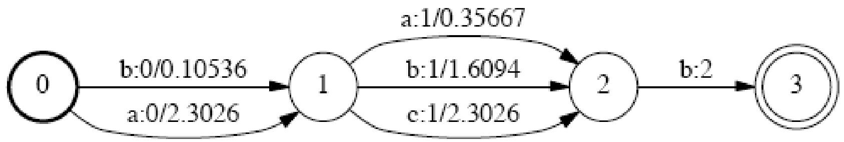


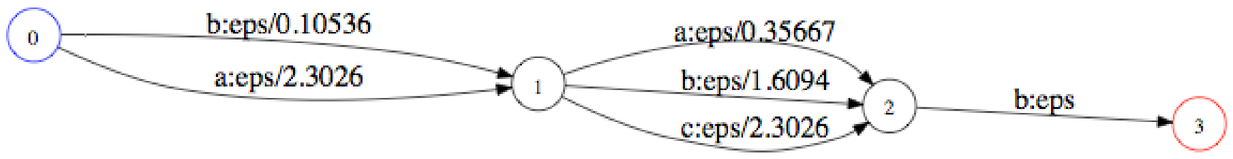
WFST-based indexing: an example

(1) 

WFST-based indexing: an example

(1)

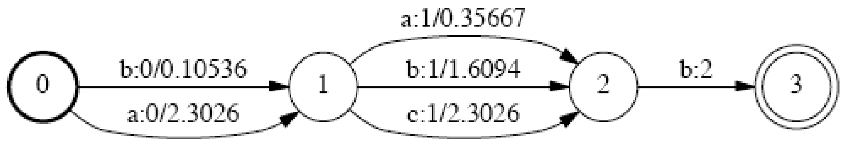


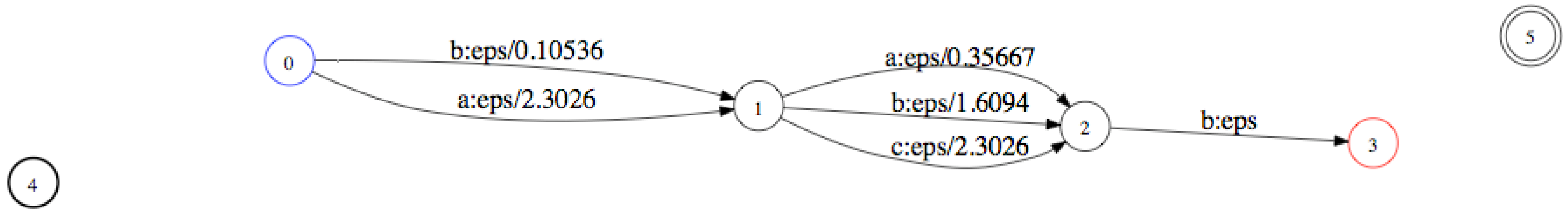


set output labels to “eps”

WFST-based indexing: an example

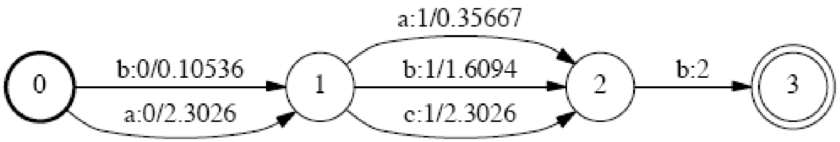
(1)

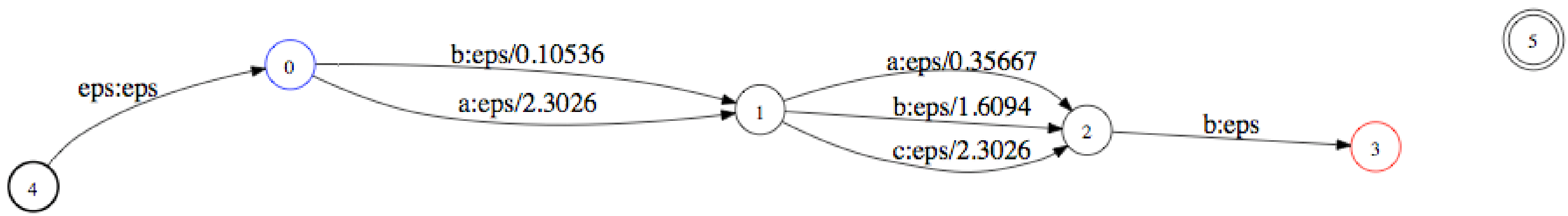




add new start state and new end state

WFST-based indexing: an example

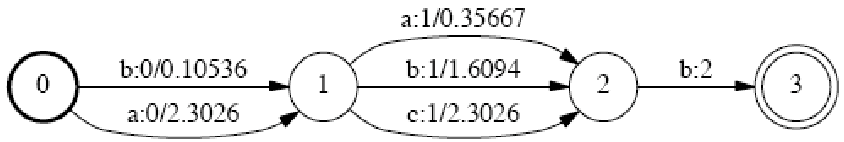
(1) 

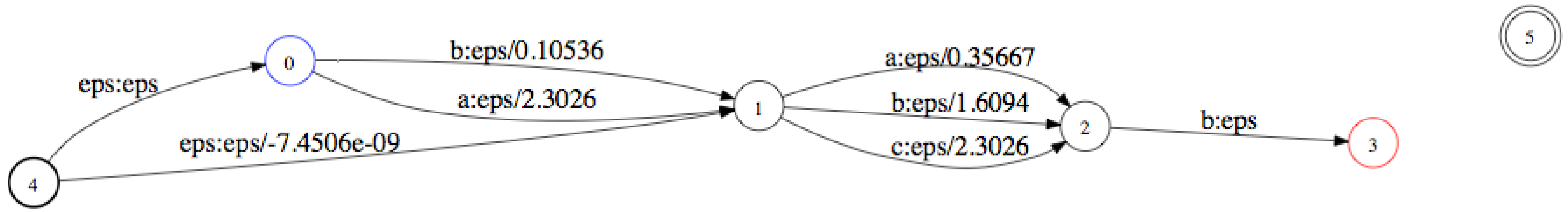


Add arc from 4 to each state S in original machine.
Weight is shortest distance in log semiring between state S to BLUE state

WFST-based indexing: an example

(1)

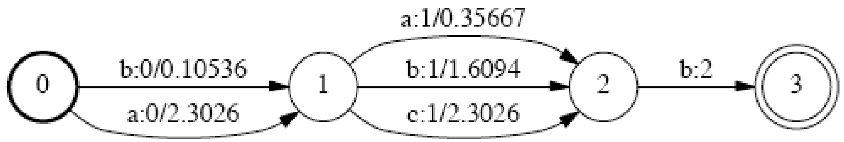


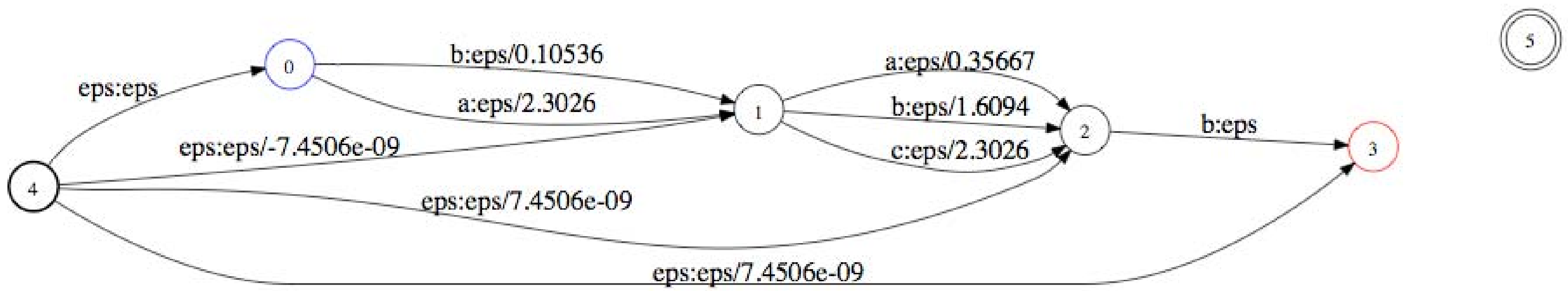


Add arc from 4 to each state S in original machine.
Weight is shortest distance in log semiring between state S to BLUE state

WFST-based indexing: an example

(1)

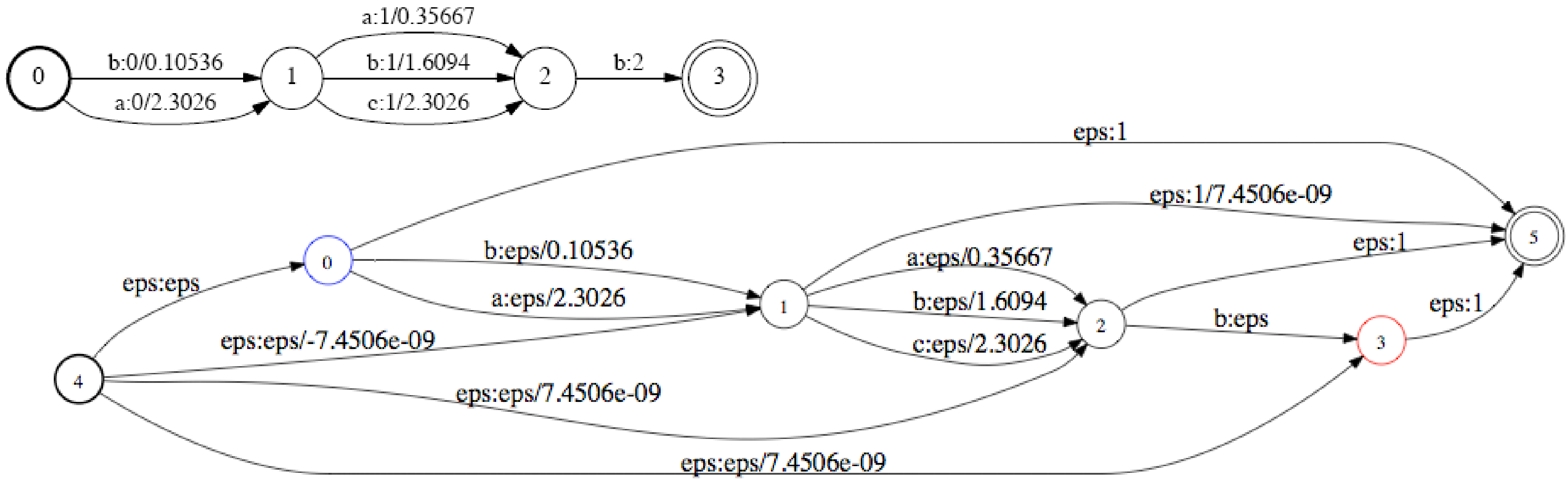




Add arc from 4 to each state S in original machine.
Weight is shortest distance in log semiring between state S to BLUE state

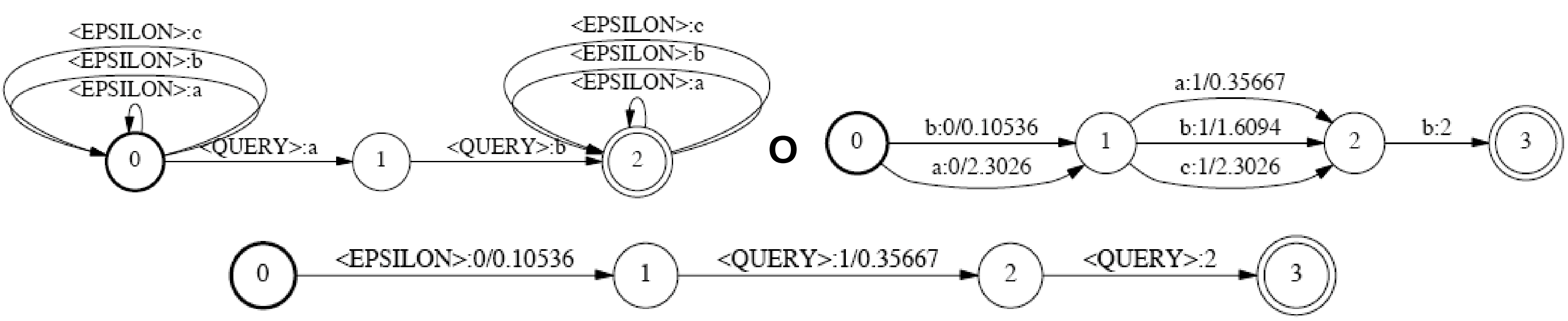
WFST-based indexing: an example

(1)



Add arc from each state S in original machine to state 5.
Weight is shortest distance in log semiring between state S to **RED** state

- **for each query in query-list**
- **compile query into string fst**
  - compose query with index fst to get utt-ids
  - padfst = pad query fst on left and right
  - for each utt-id
    - load utt-fst
    - shortest-path(compose(padded-query, utt-fst))
    - read off output labels of marked arcs



Augmenting STD with web based pronunciations

- **Generating pronunciations for OOV terms is important for spoken term detection**
- **The internet can serve as a gigantic pronunciation corpus**
- **Work done as part of CLSP 2008 workshop**
- Find pronunciations derived from the web:
  - **IPA Pronunciations: Uses International Phonetic Alphabet:**
    - Lorraine Albright / ɔl braɪt/ (Wikipedia)
  - **Ad-hoc Pronunciations: Uses informal pronunciation:**
    - Bruschetta (pronounced broo-SKET-uh)
    - Bazell (pronounced BRA-zell by the lisping Brokaw)
    - Ahmadinijad (pronounced "a mad dog on Jihad")
- **Normalize, filter and refine web-pronunciations (esp. AdHoc)**

Utility of web-pronunciations (from JHU workshop '08)

Better

Example	Pronunciations		Ref/Corr/FA/Miss	
	L2S	Web Based	L2S	Web Based
ALBRIGHT	ae l b r ay t	ao l b r ay t	276 0 1 276	276 254 20 22
GREENSPAN	g r iy n s p aa n	g r iy n s p ae n	157 0 0 157	157 85 0 72
SHIMON	sh ih m ax n	sh ih m ow n	109 0 0 109	109 98 12 11

Worse

Example	Pronunciations		Ref/Corr/FA/Miss	
	L2S	Web Based	L2S	Web Based
FREUND	f r oy n d	f r eh n d	9 3 0 6	9 3 470 6
SANTO	s ae n t ow	s ax/ey/ax/eh n t	9 2 4 7	9 2 194 7
THIERRY	th iy ax r iy	t eh r iy	7 0 16 7	7 1 1271 6

Names resemble portions of common words and prefix/suffixes

Large number of false alarms

THIERRY :: -TARY :: MILLITARY,VOLUNTARY

Experiments/Data

OOVCORP [JHU Workshop]

- Test-set:
  - 100 Hour
  - 1290 OOV queries (min 5 instances/word)
  - All queries larger than 4 phones.
- Training set (word system):
  - 300 Hours SAT system
  - 400M words, vocabulary: 83K
  - WER on RT04 BN: 19.4%
- Hybrid system:
  - Lexicon: 81.7K words and 20K fragments

DEV06

- Test-set:
  - Development set used for NIST STD 2006 Evaluation
  - 3 Hour BN
  - 1107 queries, 16 OOVs
- Training set:
  - IBM BN system
  - vocabulary: 84K

Results

DEV06

Data	P(FA)	P(miss)	ATWV
Word lattices (index:word, query:word)	0.00008	0.134	0.7991
Word CNs (index:word, query:word)	0.00007	0.094	0.8459
Hybrid lattices (index:phonetic, query:phones)	0.00008	0.240	0.6779
Merged (IV:word, OOV:phones)	0.00007	0.093	**0.8490**

OOVCORP (OOV-only queries, phonetic index)

Data	Pron model	# Best	P(FA)	P(miss)	ATWV
Word Lat	reflex	N/A	0.00004	0.638	**0.325**
Hybrid Lat	reflex	N/A	0.00002	0.639	0.342
Word Lat	L2S-weighted	6	0.00002	0.674	0.305
Hybrid Lat	L2S-weighted	6	0.00002	0.636	**0.342**

Results

DEV06

Data	P(FA)	P(miss)	ATWV
Word lattices (index:word, query:word)	0.00008	0.134	0.7991
Word CNs (index:word, query:word)	0.00007	0.094	0.8459
Hybrid lattices (index:phonetic, query:phones)	0.00008	0.240	0.6779
Merged (IV:word, OOV:phones)	0.00007	0.093	**0.8490**

OOVCORP (OOV-only queries, phonetic index)

Data	Pron model	# Best	P(FA)	P(miss)	ATWV
Word Lat	reflex	N/A	0.00004	0.638	**0.325**
Hybrid Lat	reflex	N/A	0.00002	0.639	0.342
Word Lat	L2S-weighted	6	0.00002	0.674	0.305
Hybrid Lat	L2S-weighted	6	0.00002	0.636	**0.342**

True Instances	23322
Hits	8105
False Alarms	10446

FST-based STD vs JURU/Lucene

WFST-based	**JURU-based 2006 system**
lattice and confusion networks	confusion networks
no boosting	boost posteriors based on ranking
no query-length normalization	query-length normalization
term specific threshold	global threshold

WFST-based vs JURU-based

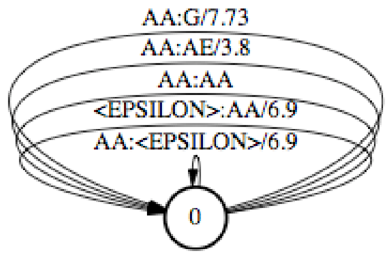
System	**Data**	**P(FA)**	**P(miss)**	**MTWV**	**ATWV**
JURU 2006	Word & Phonetic CNs	0.00005	0.108	0.8379	0.8348
WFST-based	Word CNs & Phonetic Lats	0.00007	0.093	0.8392	**0.8490**

Increasing Hits

- **Increasing hits # 1: include phonetic confusability in query**
  - Create phone-to-phone confusability matrix.
  - Model phonetic confusability using posteriors of NN-based acoustic model and aligned reference [Upendra 2009].
  - Easy to incorporate in the WFST-based framework

Increasing Hits

- **Increasing hits # 1: include phonetic confusability in query**
  - Create phone-to-phone confusability matrix.
  - Model phonetic confusability using posteriors of NN-based acoustic model and aligned reference [Upendra 2009].
  - Easy to incorporate in the WFST-based framework

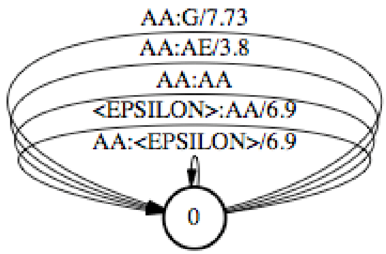


Increasing Hits

- **Increasing hits # 1: include phonetic confusability in query**
  - Create phone-to-phone confusability matrix.
  - Model phonetic confusability using posteriors of NN-based acoustic model and aligned reference [Upendra 2009].
  - Easy to incorporate in the WFST-based framework

$$oldq = \text{shortest-pathN}(q \circ L2S)$$

$$newq = \text{shortest-pathN}(q \circ L2S \circ P2P)$$



Reducing False Alarms

- **Reducing FAs #1: Query-length normalization [Mamou et al. 2007]:**

$$score(q, hit, \gamma) = p(hit)^{\frac{\gamma}{avg-duration(q)}}, \qquad \gamma \in [0, 1]$$

- **Reducing FAs #2: OOV-detection [Arastrow et al. 2009]**

–Simplest OOV detector: use posterior probabilities of fragments in a confusion bin (hybrid CN) as indicator of OOV region [frag_p > 0]

–Reduce confidence of hit if query and region do not match.

Experiments: OOVCORP

- **Increasing hits: Phone-to-Phone transducer**

	none	P2P-10best	P2P-20best	P2P-100best
ATWV	0.342	0.368	0.383	**0.3964**
% rel improv	-	7.6%	12%	**15.9%**

Experiments: OOVCORP

- **Increasing hits: Phone-to-Phone transducer**

	none	P2P-10best	P2P-20best	P2P-100best
ATWV	0.342	0.368	0.383	**0.3964**
% rel improv	-	7.6%	12%	**15.9%**

- **OOV-detection + length-normalization + cache: pron-model: P2P-20best**

OOV-det	γ-norm	cache	P(FA)	P(miss)	ATWV	improv
			0.00004	0.575	0.383	
x			0.00004	0.578	0.384	0.2%
	x		0.00005	0.555	0.394	+2.87%
		x	0.00006	0.557	0.383	0%
x	x	x	0.00004	0.551	**0.405**	+18.4%

oov-det

Query-by-Example (QbyE)

- Spoken Term Detection when the terms of interest are acoustic examples: Query by Example (QbyE).
  - **User identifies region of interest in speech stream and requests for similar examples.**
  - **User speaks query: speech to speech retrieval.**
- Focus on improving performance for Out Of Vocabulary (OOV) words.
- Demonstrates flexibility of FST-based indexing system

Query Generation for QbyE

- **Lattice Cuts : User selects a region of interest in the audio stream**
  – Represent region of interest by excising lattice corresponding to the decode for the region
  – Query representation generated by the same ASR system which generates the index
- **Isolated decodes: User presents example of audio**
  – Use lattice from an isolated decode of the audio example
- **The queries for both cases are graph structures similar to ASR lattices**
- **Pruned representation of queries found to be faster, more robust and generate lower false alarms**

Query by Example : Key results

- QbyE typically perform significantly better then textual queries for OOV terms (about 20% relative in ATWV)
- Queries represented as *lattice-cuts* from the lattices of interest yield better STD performance than *isolated-decode* queries.
- Addressing FA rates associated with multi-path queries improves performance significantly.
- QbyE can enhance performance of textual queries when using a two-pass approach