N-GRAM FST INDEXING FOR SPOKEN TERM DETECTION

CHAO LIU

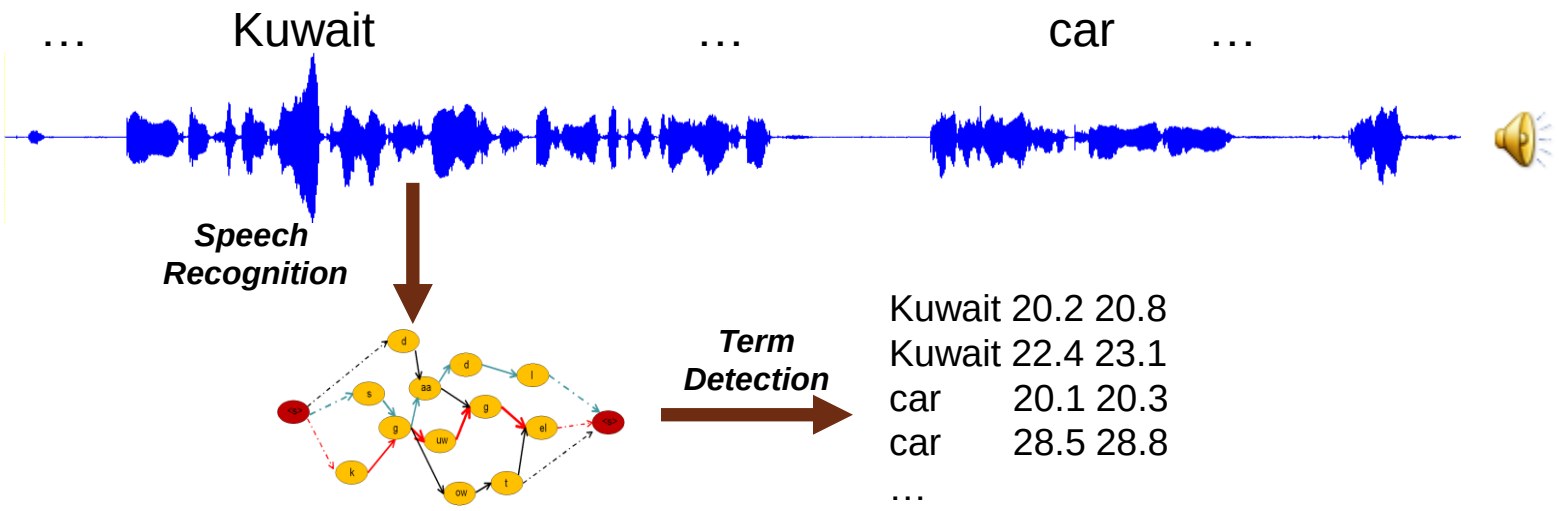
SEP 24, 2012

CONTENTS

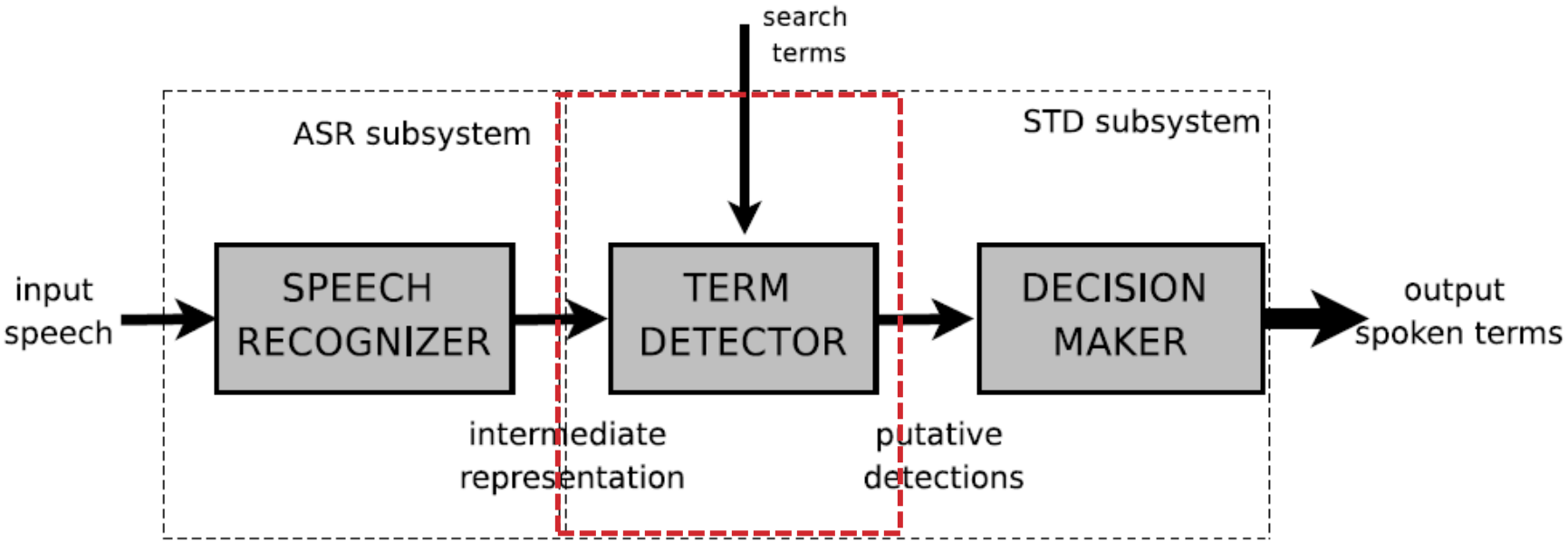
CONTENTS

SPEECH TERM DETECTION

- Spoken Term Detection (STD) : find **all of the occurrences of a specified "term"** in a given corpus of speech data.(NIST)
  - Term：a sequence of one or more words. For example: "car", "New York".
  - System output: **location** of the term in audio, a **score** indicating how likely the term exists.
  - Evaluation：both **speed** and detection **accuracy**.



SPEECH TERM DETECTION

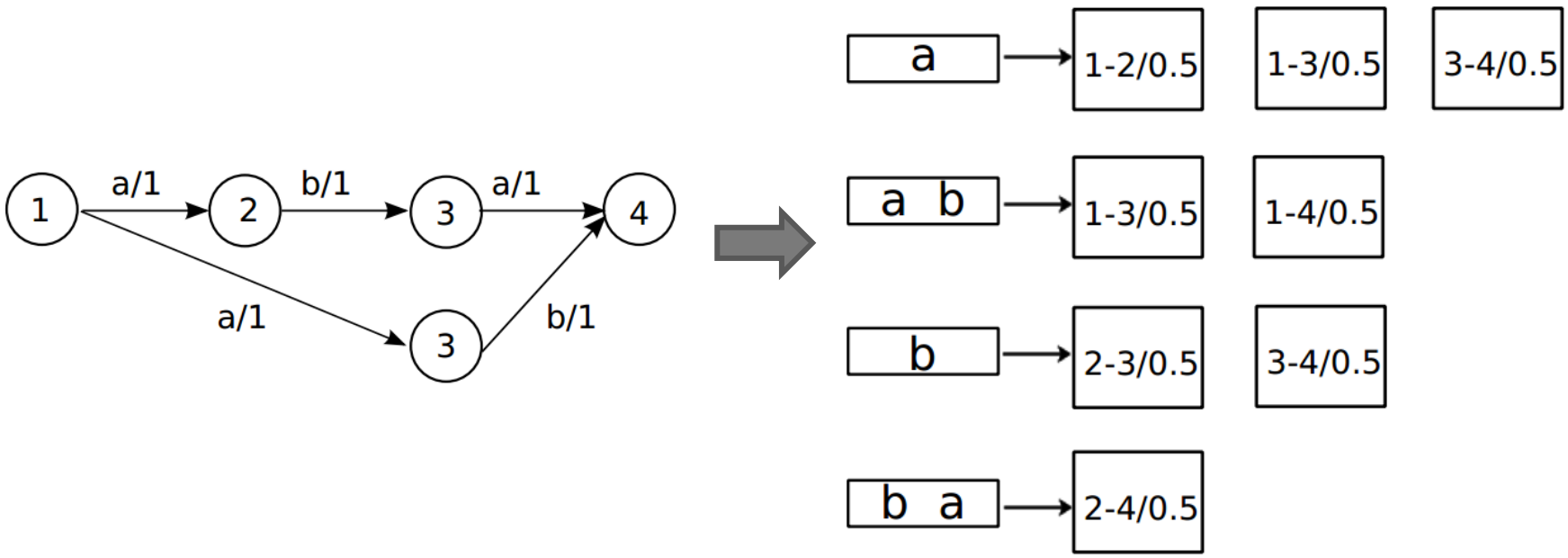


We focus on an efficient indexing scheme, which is essentially important for STD on large databases.

CONTENTS

N-GRAM INVERTED INDEXING

- Get all n-gram fragments with their confidence scores existing in the input lattice, and sort them in chronological order.
- Speed up term searching using inverted list.

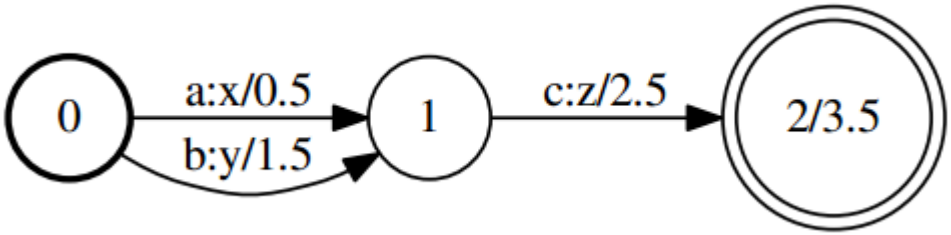


CONTENTS

FINITE STATE TRANSDUCER

- **Basic parts of FST**
  - Input label - phone / n-gram
  - Output label – time period
  - Weight – confidence

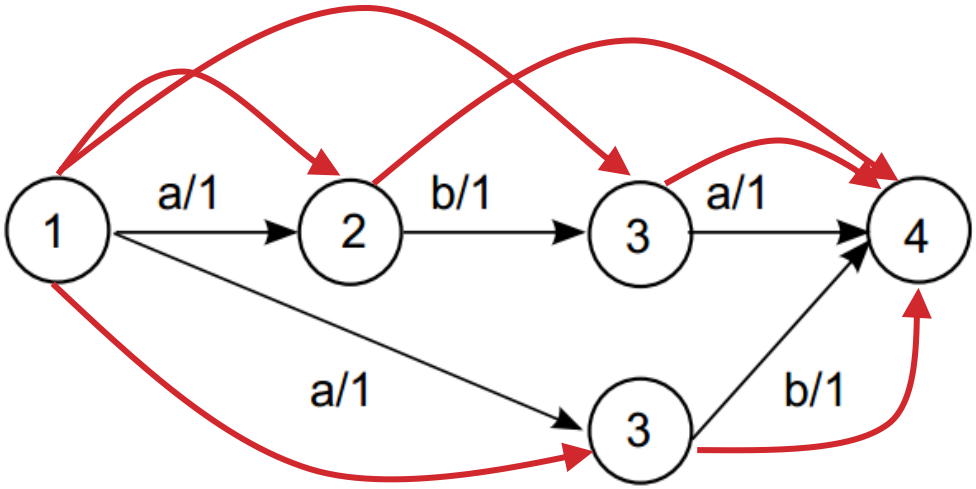
- **FST operations**
  - Determinization
  - Minimization
  - Unification



ac->xz / 6.5
bc->yz / 7.5

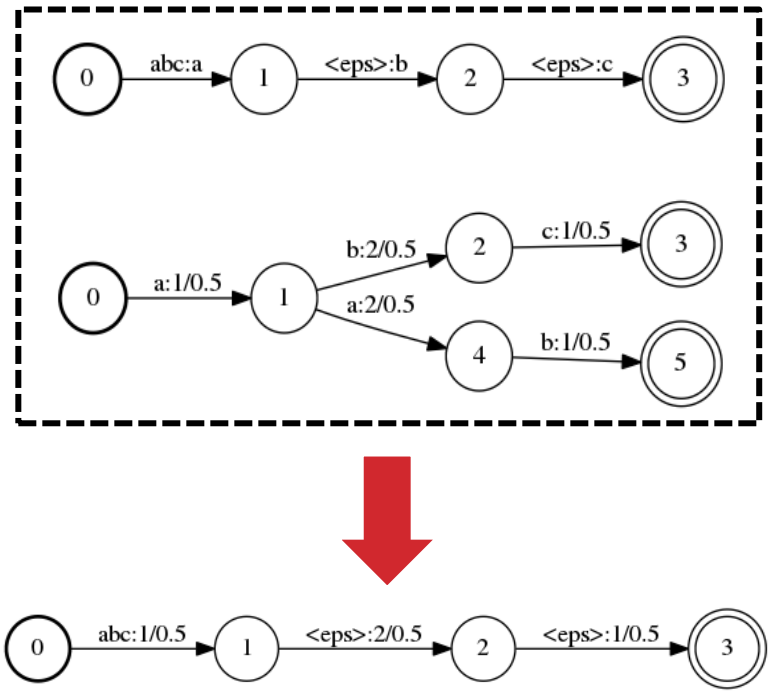
FST INDEXING

- **Convert lattice to FST by linking initial and final states to all other states.**



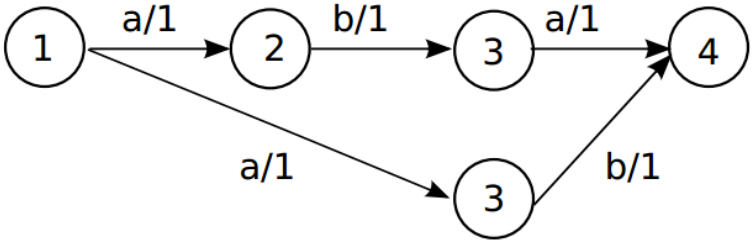
FST INDEXING

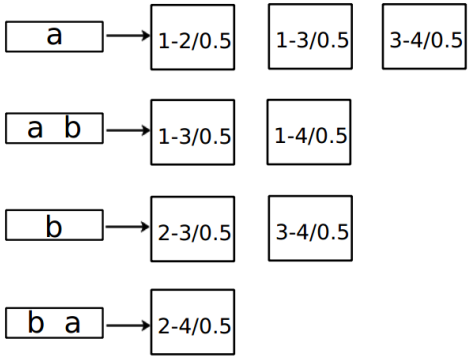
- Compile searching term to FST
- Do composition on term FST and utterance FST

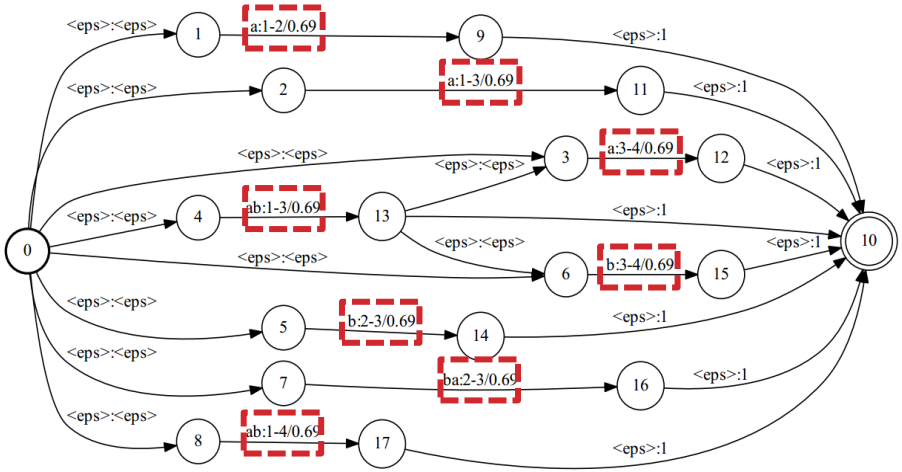


CONTENTS

N-GRAM FST INDEXING



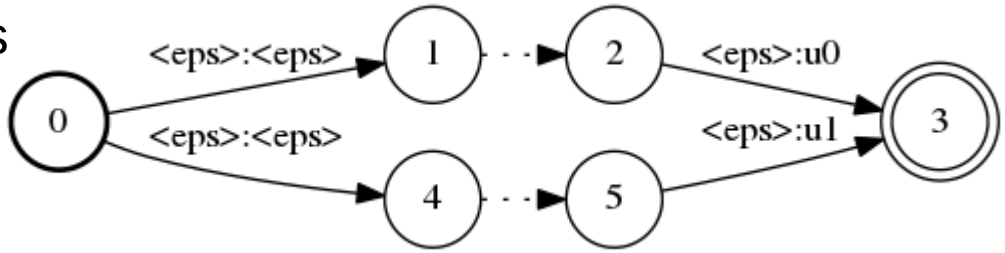




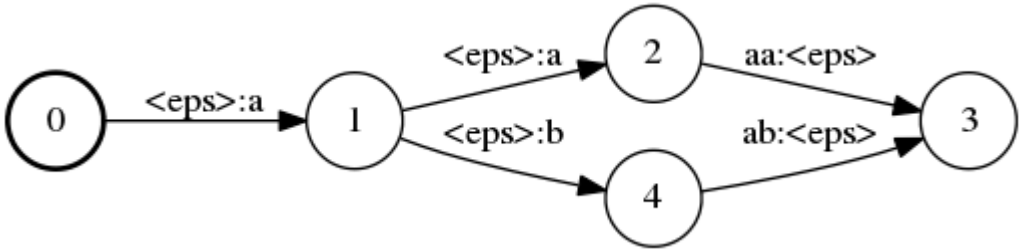
- **Convert lattice to N-gram fragments**
- **Compile N-gram fragments to FST**

OPTIMIZATION

- **Standard operations**
  - Determinize, Minimize, RmEpsilon.
  - Viewing it as an acceptor, encoded label (Allauzen and Mohri, 2004)
- **Union**
  - Corpus



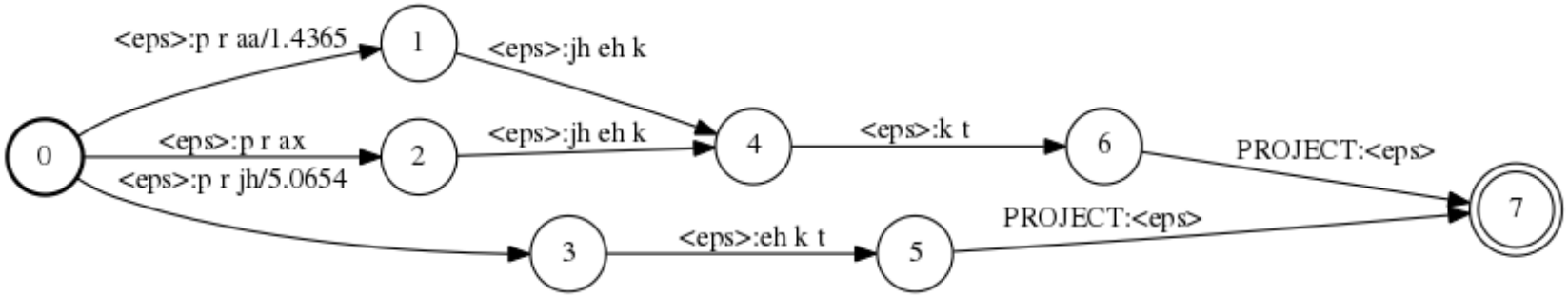
  - Terms



FUZZY SEARCH

- **OOV words with uncertain pronunciation / mispronounced**
  - N-best pronunciation prediction (Wang and King, 2011)

- **Just union FSTs together**



CONTENTS

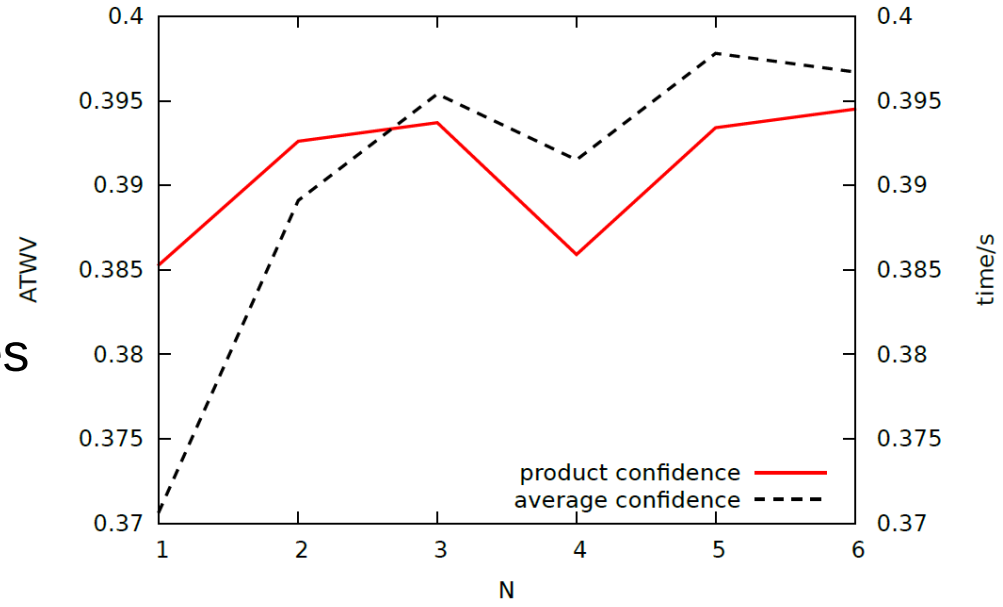
BACKGROUND

- **The ASR system was built with corpora used for train AMI RT05s ASR system**
  - 80.2 hours of speech for acoustic model (AM) training
  - 521M words of text for language model (LM) training
  - Phone Error Rate (PER) is 40.49%
  - Average lattice density is 805 nodes / second
- **STD Experiments were performed on RT04s and RT05s data sets**
  - 489 INV terms and 67 OOV terms for development
  - 255 INV terms and 484 OOV terms for evaluation

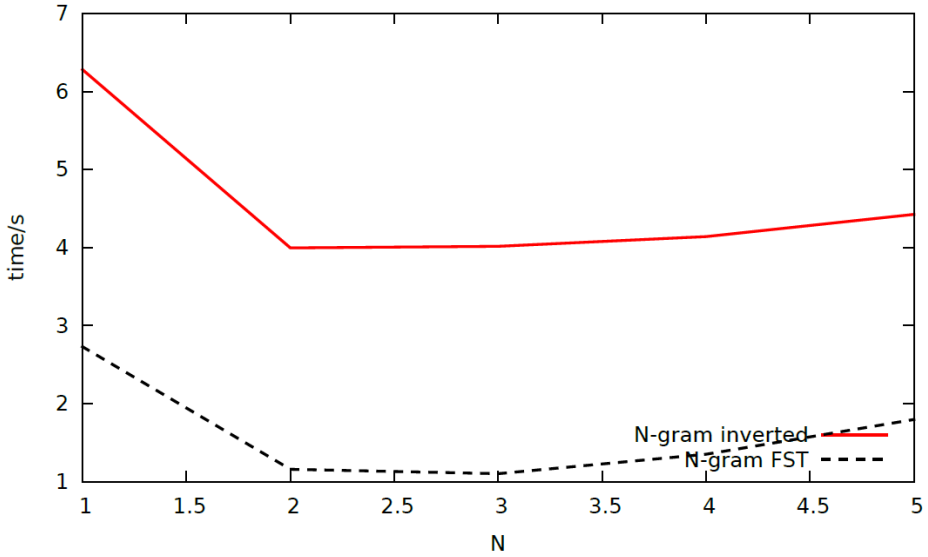
EXPERIMENTS

- **Metric for accuracy：Actual Term Weighted Value**
  - $ATWV = 1 - \underset{term}{average}\{P_{Miss}(term) + \beta \cdot P_{FA}(term)\}$

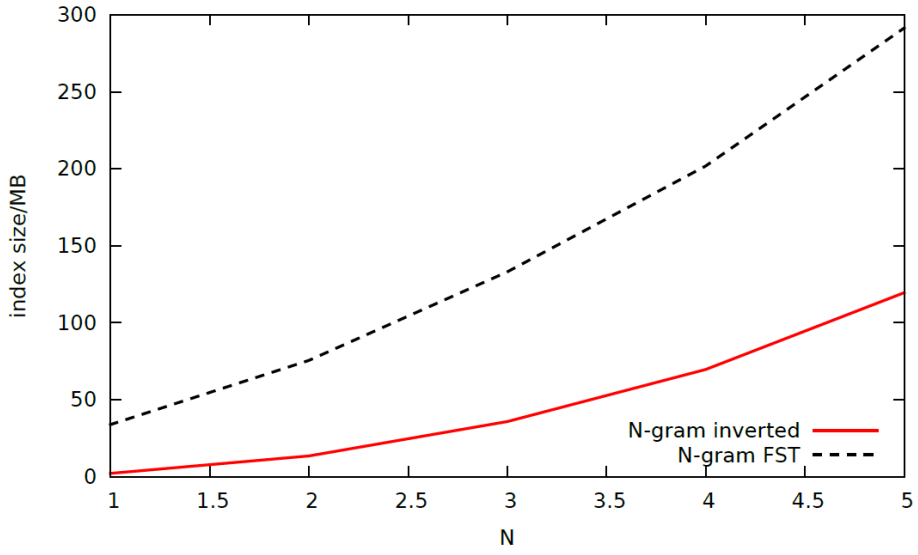
- **Relevant factors**
  - N
  - Confidence measures



COMPARISON

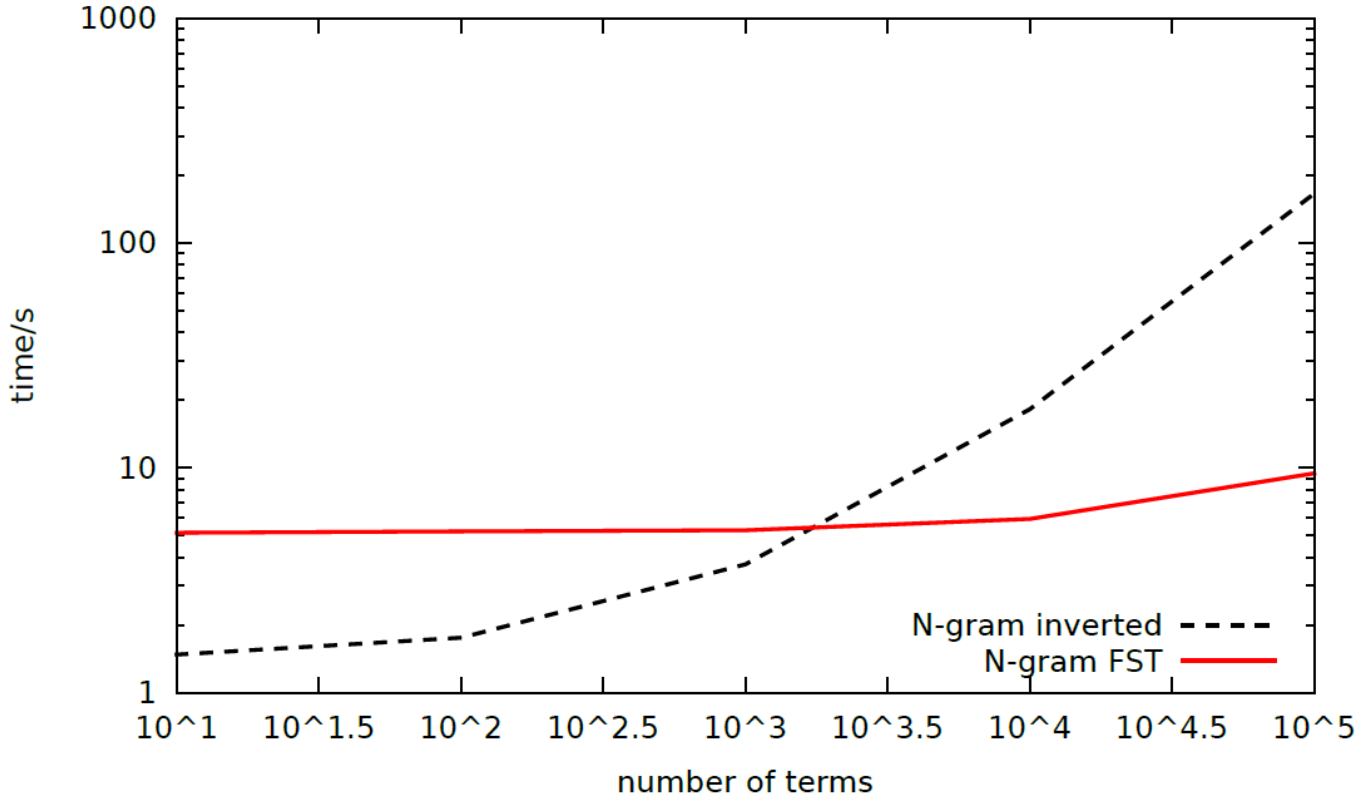


Searching efficiency



Index size

COMPARISON


1000
100
10
1
time/s
10^1
10^1.5
10^2
10^2.5
10^3
10^3.5
10^4
10^4.5
10^5
number of terms
N-gram inverted
N-gram FST

RESULTS ON EVAL SET

INV terms	ATWV	Index size/MB	Time/s
Lattice	0.4782	483	>10^3
FST indexing	0.4782	959	16.6
N-gram inverted indexing	0.5310	226	6.0
N-gram FST indexing	0.5310	943	5.9

OOV terms	ATWV	Index size/MB	Time/s
Lattice	0.2191	483	>10^3
FST indexing	0.2191	959	19.0
N-gram inverted indexing	0.2813	226	9.7
N-gram FST indexing	0.2813	943	6.0
FST indexing / fuzzy	0.2305	959	81.1
N-gram inverted indexing / fuzzy	0.3156	226	401.9
N-gram FST indexing / fuzzy	0.3156	943	30.4

CONTENTS

CONCLUSION

- **Compared with conventional FST indexing, N-gram FST indexing provides better STD performance by relaxing phone connectivity.**
- **Compared with the conventional N-gram inverted indexing, this approach is faster and possesses advantages of FSTs in terms of solid theory and rich tools.**
- **N-gram FST indexing shows significant improvement while doing fuzzy search.**

THANK YOU !

Q&A