

ON KHMER INFORMATION RETRIEVAL

12 March 2011

VAN CHANNA

Kameyama Laboratory, GITS

Waseda University

CONTENTS

- Research Background
- Introduction to Khmer Language
- Building a Khmer Text Corpus
 - Methodology
 - Current Statistic
- Query Expansion Techniques for Khmer Information Retrieval
 - Proposed techniques
 - Experiment and Results
- A trainable rule-based for Khmer Word Segmentation
 - Approach
 - Experiment and Results
- Conclusion

RESEARCH BACKGROUND

- Information Retrieval (IR) system is very important for searching the any kind of information.
- No specific Khmer IR system has been implemented.
- No research on Khmer IR system has been investigate.
- A specific Khmer IR system shall be studied in order to handle the flood of Khmer information.

KHMER

- Khmer is the official language of Cambodia spoken by 15 millions in Cambodia.
- Khmer exists its own alphabet
 - Derives from an old Indian
 - None-segmented
- In modern standard Khmer script consists of:
 - 33 consonants.
 - 32 subscripts.
 - 24 dependent vowels.
 - 12 independent vowels
 - 2 consonant shifters, a dozen diacritics signs and other symbols.
- Unicode is the only Khmer standard encoding currently exists.

	178	179	17A	17B	17C	17D	17E	17F
0	ក 1780	ច 1790	ហ 17A0	ញ 17B0	ៀ 17C0	្ក 17D0	័ 17E0	្ខ 17F0
1	ខ 1781	ទ 1791	ឡ 17A1	ឌ 17B1	្គ 17C1	្ខ 17D1	៑ 17E1	្គ 17F1
2	គ 1782	ឆ 1792	ង 17A2	ត 17B2	្ង 17C2	្ខ 17D2	្ 17E2	្ឃ 17F2
3	យ 1783	ន 1793	អ 17A3	ឃ 17B3	្ឆ 17C3	្គ 17D3	៓ 17E3	្ង 17F3
4	ដ 1784	ប 1794	រ 17A4	KIV AQ 17B4	្ឈ 17C4	្ឃ 17D4	។ 17E4	្ច 17F4
5	ច 1785	ជ 1795	ត 17A5	KIV AA 17B5	្ដ 17C5	្ង 17D5	៕ 17E5	្ឆ 17F5
6	ឆ 1786	ព 1796	ឃ្ល 17A6	្ណ 17B6	្ឌ 17C6	្ច 17D6	៖ 17E6	្ជ 17F6
7	ជ 1787	ភ 1797	ខ 17A7	្ត 17B7	្ណ 17C7	្ជ 17D7	ៗ 17E7	្ឈ 17F7
8	ស 1788	ម 1798	ឌ 17A8	្ថ 17B8	្ត 17C8	្ជ 17D8	៘ 17E8	្ញ 17F8
9	ក្រ 1789	យ 1799	ឌ 17A9	្ទ 17B9	្ត 17C9	្ឈ 17D9	៙ 17E9	្ដ 17F9
A	ជ 178A	រ 179A	ឃ 17AA	្ធ 17BA	្ត 17CA	្ញ 17DA		
B	ថ 178B	ល 179B	ប 17AB	្ន 17BB	្ត 17CB	្ដ 17DB		
C	ឌ 178C	រ 179C	ប 17AC	្ប 17BC	្ត 17CC	្ឋ 17DC		
D	ណ 178D	ដ 179D	ត 17AD	្ផ 17BD	្ត 17CD	្ឌ 17DD		
E	ណ 178E	ថ 179E	ត 17AE	្ព 17BE	្ត 17CE			
F	ត 178F	ស 179F	ជ 17AF	្ភ 17BF	្ត 17CF			

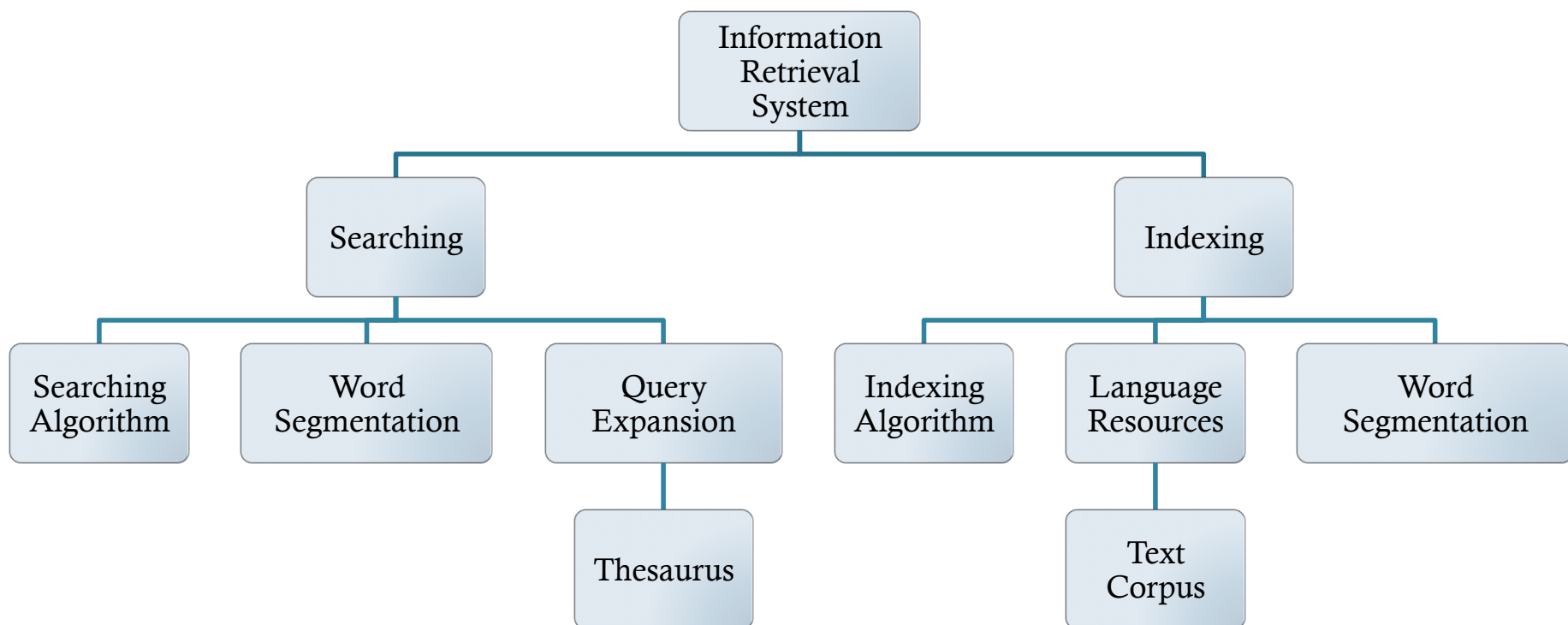
KHMER

ស ស ក ម ជ

ស ស ក ម ជ

OVERVIEW OF THE IR SYSTEM

- Building an IR system for the language like Khmer is a challenging task due to the limited number of studies in Khmer language processing, and the lack of Khmer language resource such as Text Corpus.



THE FUNDAMENTAL WORKS OF KHMER IR SYSTEM

- Three kind of fundamental works for Khmer IR system as well as Khmer NLP have been studied:
 - Khmer text corpus
 - The query expansion techniques for Khmer IR
 - The Khmer word segmentation.

BUILDING A KHMER TEXT CORPUS

- Objective: build a Khmer text corpus which is useful and beneficial to all types of research in Khmer language processing.

Text Collection

- Sources: Internet (websites and blogs).
- Method: Semi-automatic.

Preprocessing Tasks

- Cleaning: remove the unwanted elements such as photos, HTML elements and so on.
- Labeling: assign the information of the text.

Corpus Annotations

- Sentence: Position, ID and length.
- Word: Position, ID and length.
- POS: part-of-speech of the words.

Corpus Encoding

- eXtensible Corpus Encoding Standard (XCES*): an XML-based corpus encoding .

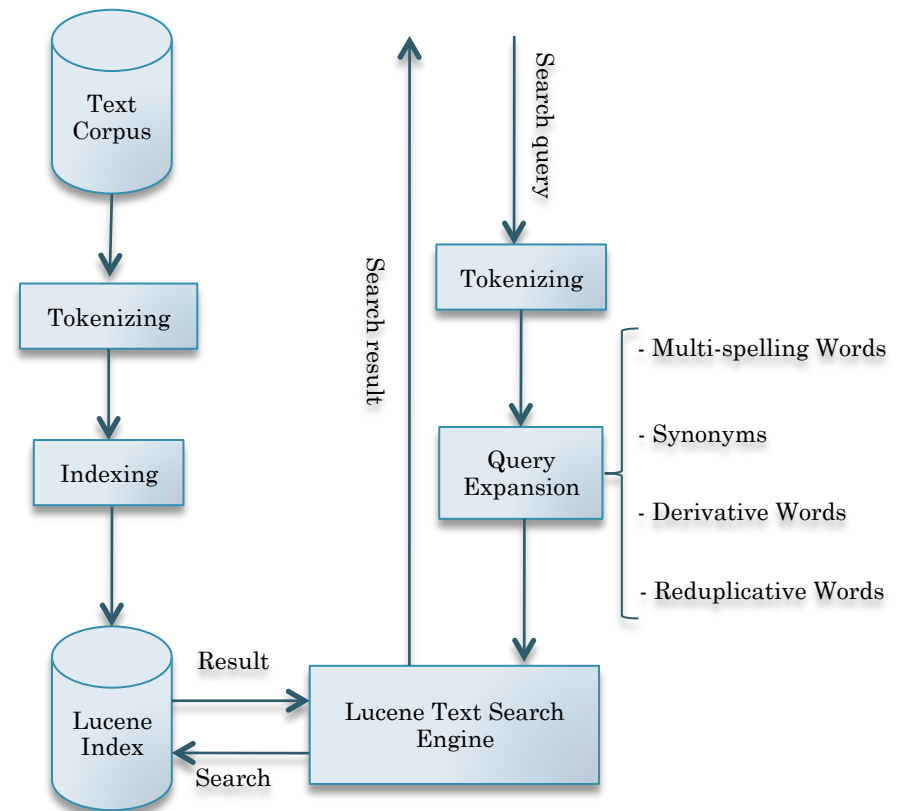
CURRENT CORPUS STATISTIC

- Corpus Statistics
 - **5906** articles in 12 different domains.
 - More than 3 millions words.
- The size of the corpus is relatively small at the moment, the expansion task is continuously undergoing.

Domain	# Article	# Sentence	# Word
Newspaper	5523	66397	2341249
Magazine	52	1335	42566
Medical	3	76	2047
Technology	15	607	16356
Cultural	33	1178	43640
Law	43	5146	101739
History	9	276	7778
Agriculture	29	1484	30813
Essay	8	304	8318
Story	108	5642	196256
Novel	78	12012	236250
Other	5	134	5522
Total	5906	94591	3000139

PROPOSED QUERY EXPANSION TECHNIQUES FOR KHMER IR

- Four types of QE technique based on the specific characteristics of Khmer language:
 - Spelling-variants
 - Synonyms
 - Derivative words
 - Reduplicative words
- A prototype of Khmer IR system was implemented. The system is based on:
 - Lucene*: a popular opened source full-text search framework.
 - Khmer word segmenter from PAN Cambodia Localization**.



* Apache Lucene: <http://lucene.apache.org>.

** K. W. Church, L. Robert, and L. Y. Mark. A Status Report on ACL/DCL. pages 84—91, 1991.


EXPERIMENTAL SET UP

- A Khmer text corpus, which consists of 954 articles, was used.
- The proposed prototype of Khmer IR was used for the evaluation.
- The Google web search engine was also used to evaluate the proposed QE.
- The text corpus was hosted in our laboratory web server in order that it can be indexed by Google.


EXPERIMENTAL PROCEDURE

- Four kinds of similar experiments we carried out for the four types of proposed QE techniques.

Input 10 original expandable queries for each type of experiments. Each query consists of at least an expandable word, and posses a specific topic.



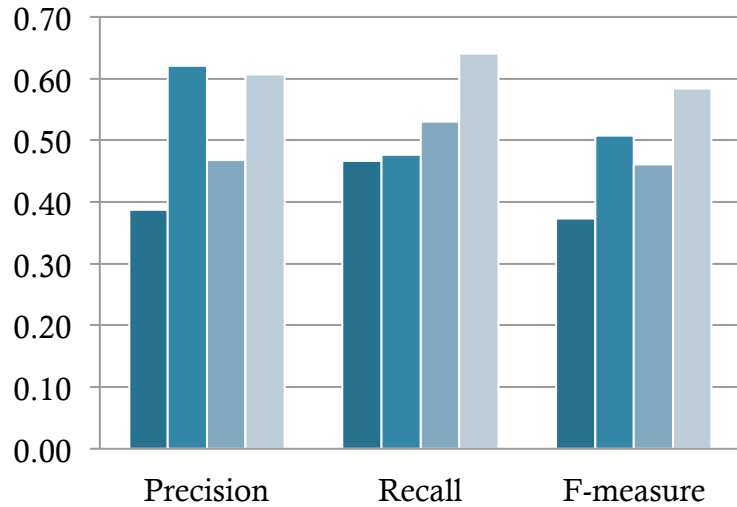
Re-input the expansion of the 10 original queries (manually expanded according to the query language of Lucene and Google) into both systems.



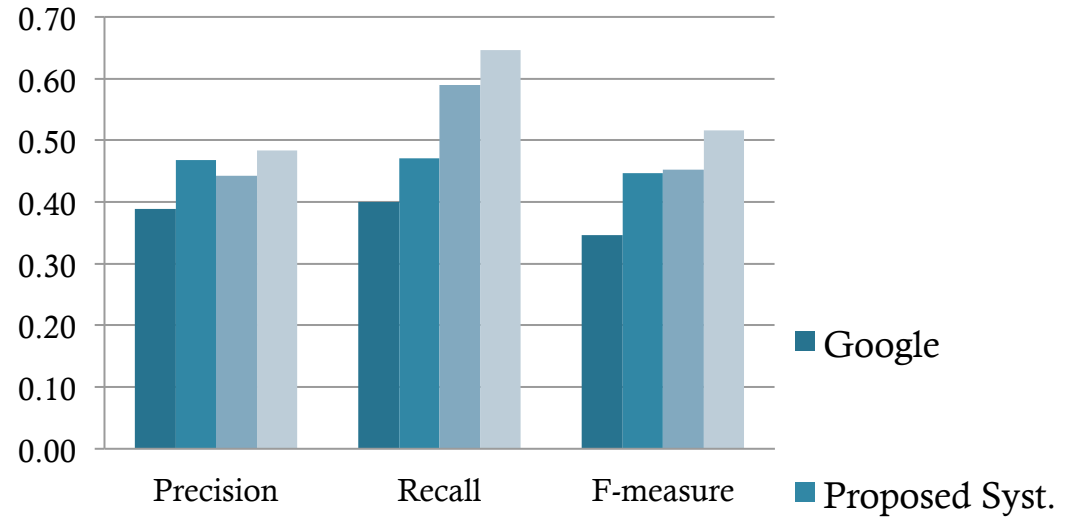
Calculate the Precisions, Recalls & F-measure of both systems.

RESULTS

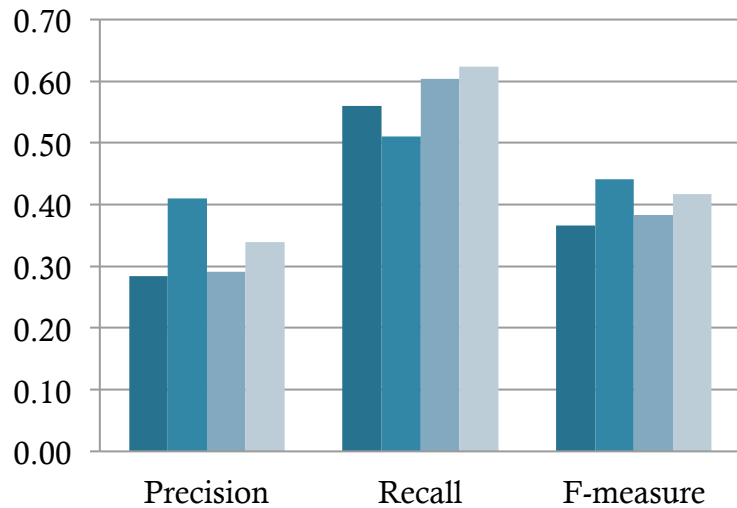
Spelling Variants



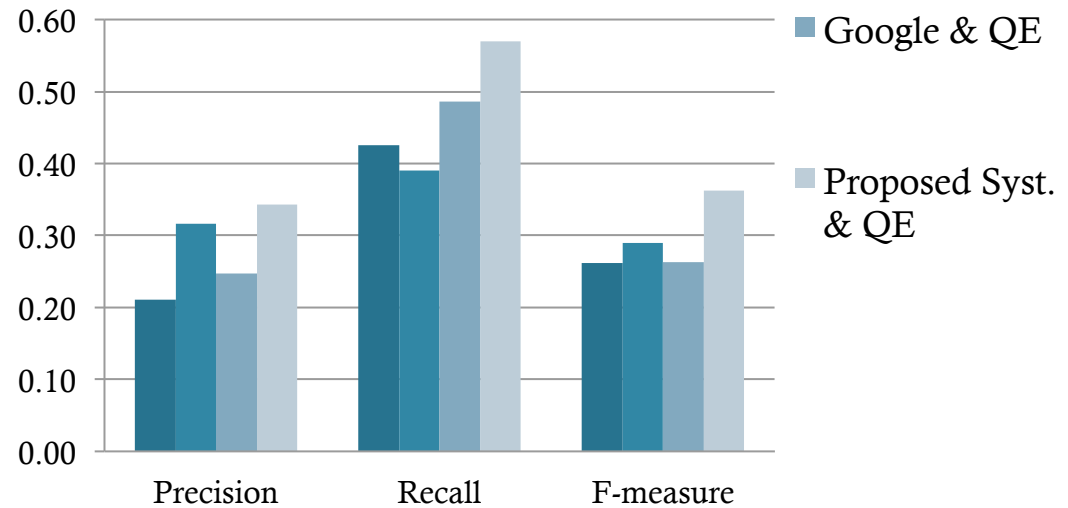
Synonyms



Derivative Words



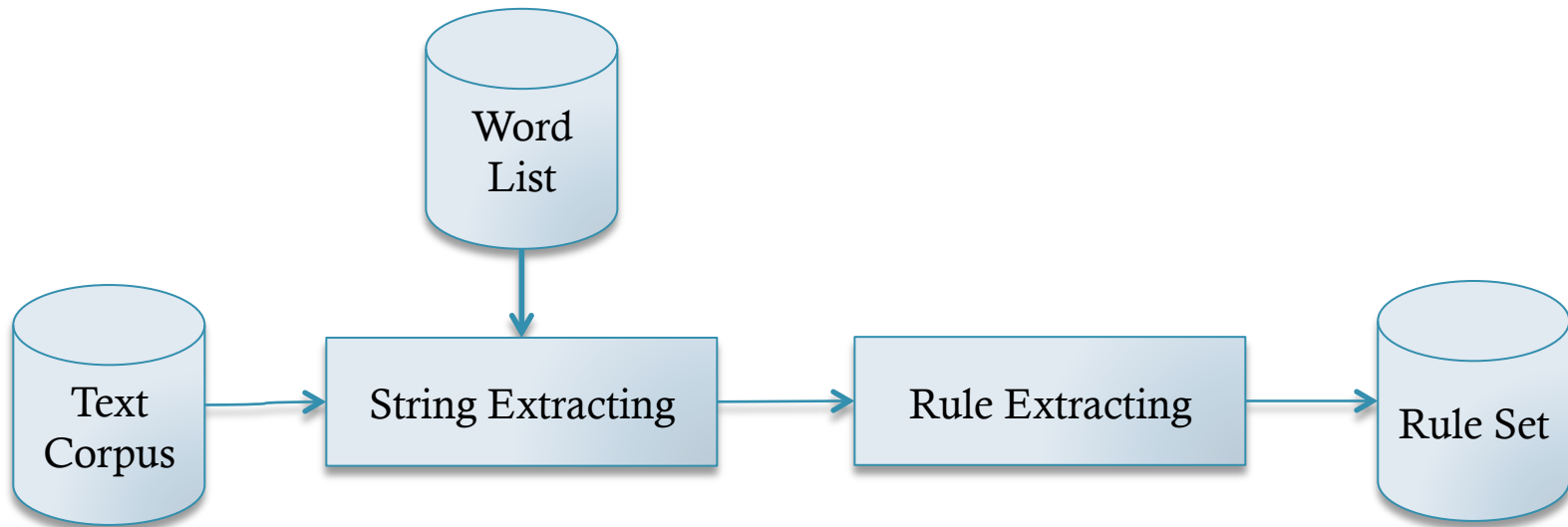
Reduplicative Words



A TRAINABLE RULE-BASED APPROACH FOR KHMER WORD SEGMENTATION

- A trainable rule-based approach using text corpus. Two main tasks were carried out:
 1. **Rule Learning:** create a rule set based on the text corpus.
 2. **Word Extraction:** extract words based on the obtained rule set and the statistical measurements.
- Issue in word segmentation:
 - Try to discover the out-of-vocabulary words: compound words, proper names , acronym and etc.

RULE LEARNING



- 5000 documents in the corpus were used.
- Extracting Strings: using the longest matching algorithm.

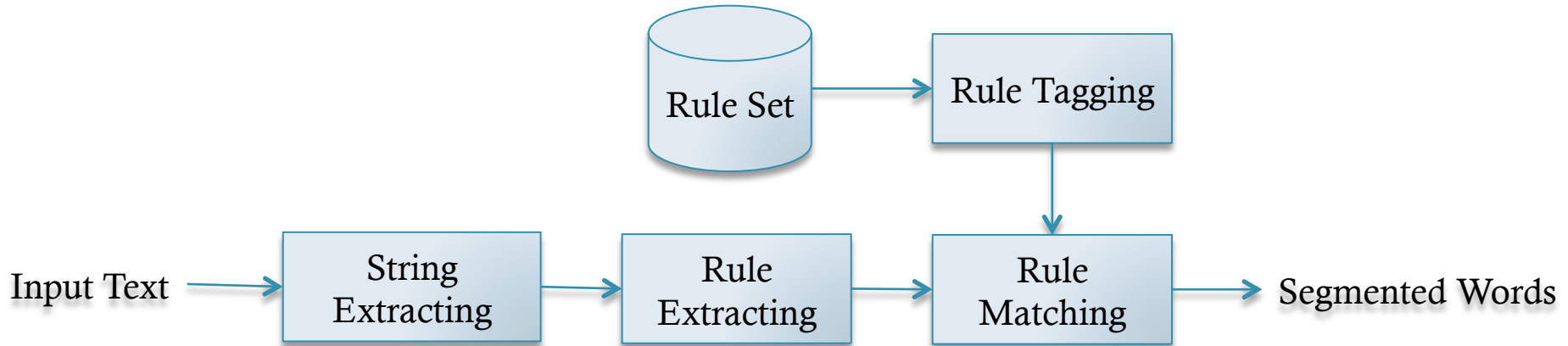
$abcdef\dots = \begin{cases} abc & \text{- if 'abc' is found in the dictionary.} \\ a & \text{- if no string started by 'a' is found in} \\ & \text{the dictionary.} \end{cases}$

- **Extracting Rules:**

- Using the SEQUITUR algorithm*.
- Each rule follows the equation: $R_i \leftarrow XY$ where X and Y is a string or a rule.

* C. Nevill-Manning and I. Witten. Identifying Hierarchical Structure in Sequences. *Journal of Artificial Intelligence Research*, 7:67--82, 1997.

WORD EXTRACTION



- Similar to the Rule Learning: String Extraction & Rule Extraction.
- Rule Tagging:
 - Each rule is tagged to be word based on the statistical measurements.
- The rules that matched to the rules after tagging will be extracted as words in the rule matching process.

RULE TAGGING

- Rule: $R_i \leftarrow XY$ where X and Y is a string or a rule.
- Two types of statistical measurements were used in the tagging process:
 - The Entropies*: Left Entropy and Right Entropy.

$$LE(R) = - \sum_{\forall x \in A} P(xR | R) \log_2 P(xR | R) \quad \text{and} \quad RE(R) = - \sum_{\forall y \in A} P(Ry | R) \log_2 P(Ry | R)$$

- Where R is the considered rule, A is the alphabet, x and y is any string co-occurred before and after R .

- The collocation measurements are used to measure the strength of two variables are are likely collocated rather than appeared by chance.
 - Mutual Information (MI)**:
$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$
 - Mutual Dependency (MD)***:
$$D(x, y) = I(x, y) - I(xy) = \log_2 \frac{P^2(xy)}{P(x)P(y)}$$
 - Log-Frequency Mutual Dependency (LFMD)***: $D_{LF} = D(x, y) + \log_2 P(xy)$
 - The Chi-square Test.

* C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379--423, 1948.

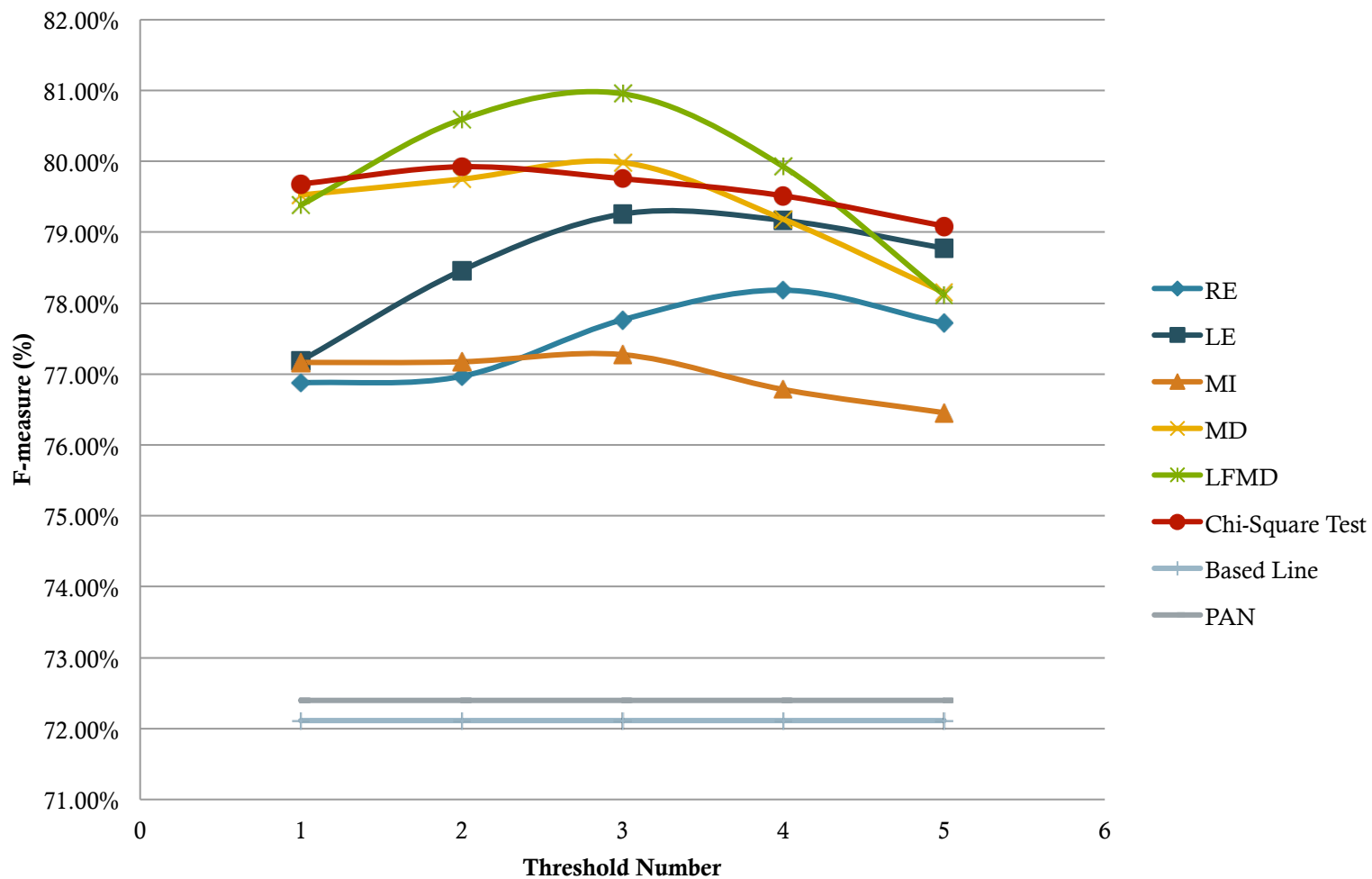
** K. W. Church, L. Robert, and L. Y. Mark. A Status Report on ACL/DCL. pages 84—91, 1991.

*** A. Thanopoulos, N. Fakotakis and G. Kokkinakis. Comparative Evaluation of Collocation Extraction Metrics

EXPERIMENTAL SETUP

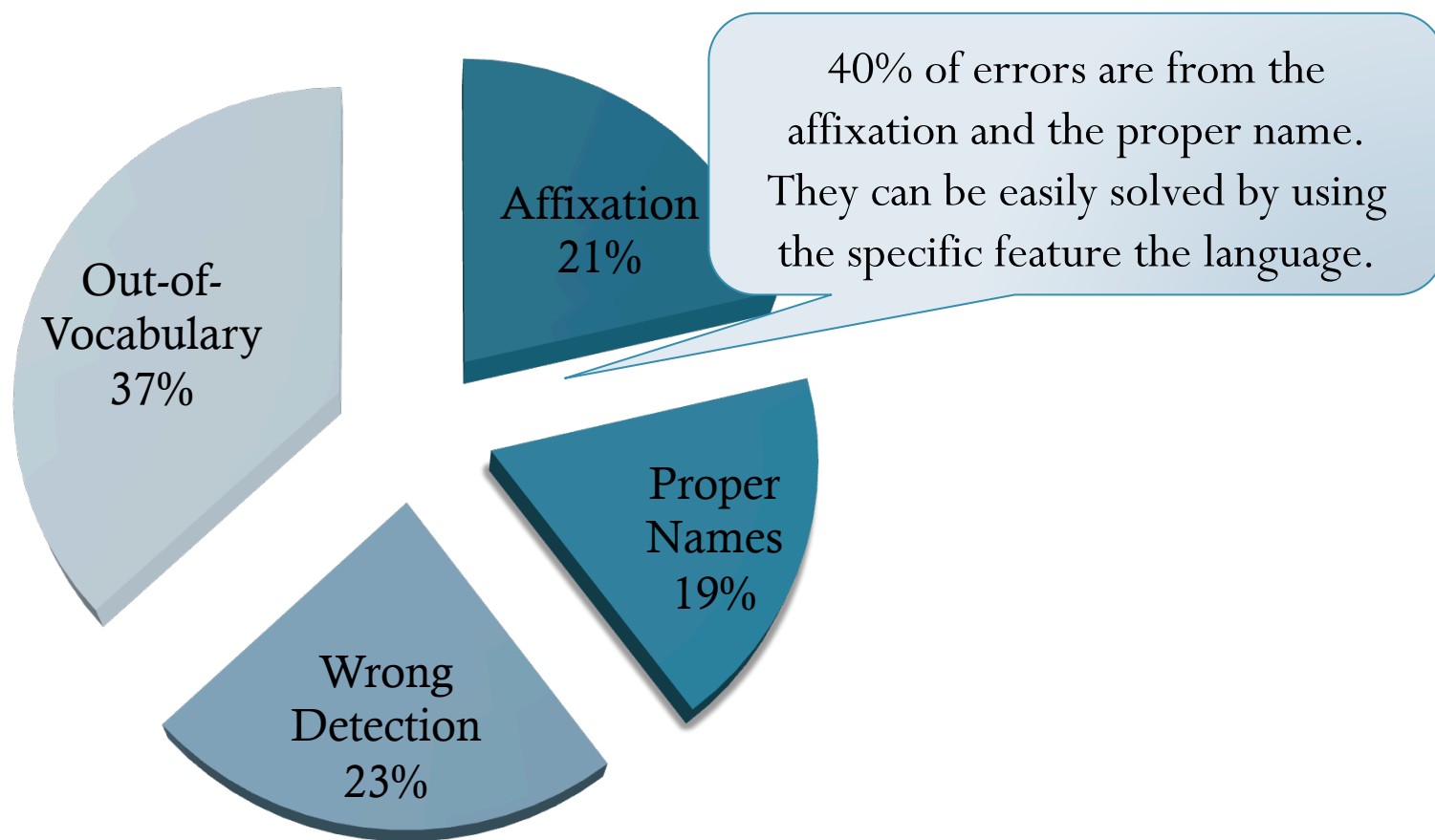
- Test Data: about 6000 words with 20% of out-of-vocabulary words.
- Experiments were conducted for each type of statistical measurements.
- For each type statistical measurement, 5 best selected thresholds were evaluated.
- Precision and Recall were calculated.
- Compare to the current state-of-the-art of Khmer word segmentation from PAN.

RESULTS



RESULT DISCUSSION

- In the case of LFMD with the threshold = -25



CONCLUSION

- Three studies have been investigated: Khmer Corpus, Query Expansion for Khmer IR and Khmer Word Segmentation.
 - We have built a Khmer text corpus which will be a great contribution to the future research of Khmer language processing.
- The four proposed QE techniques showed the improvement of the proposed Khmer IR system as well as Google.
- A new approach for Khmer Word Segmentation was proposed, the results has shown the outperformance of the proposed approach over the current state-of-the-art of Khmer Word Segmentation.

THANK YOU VERY MUCH!

SEQUITUR ALGORITHM

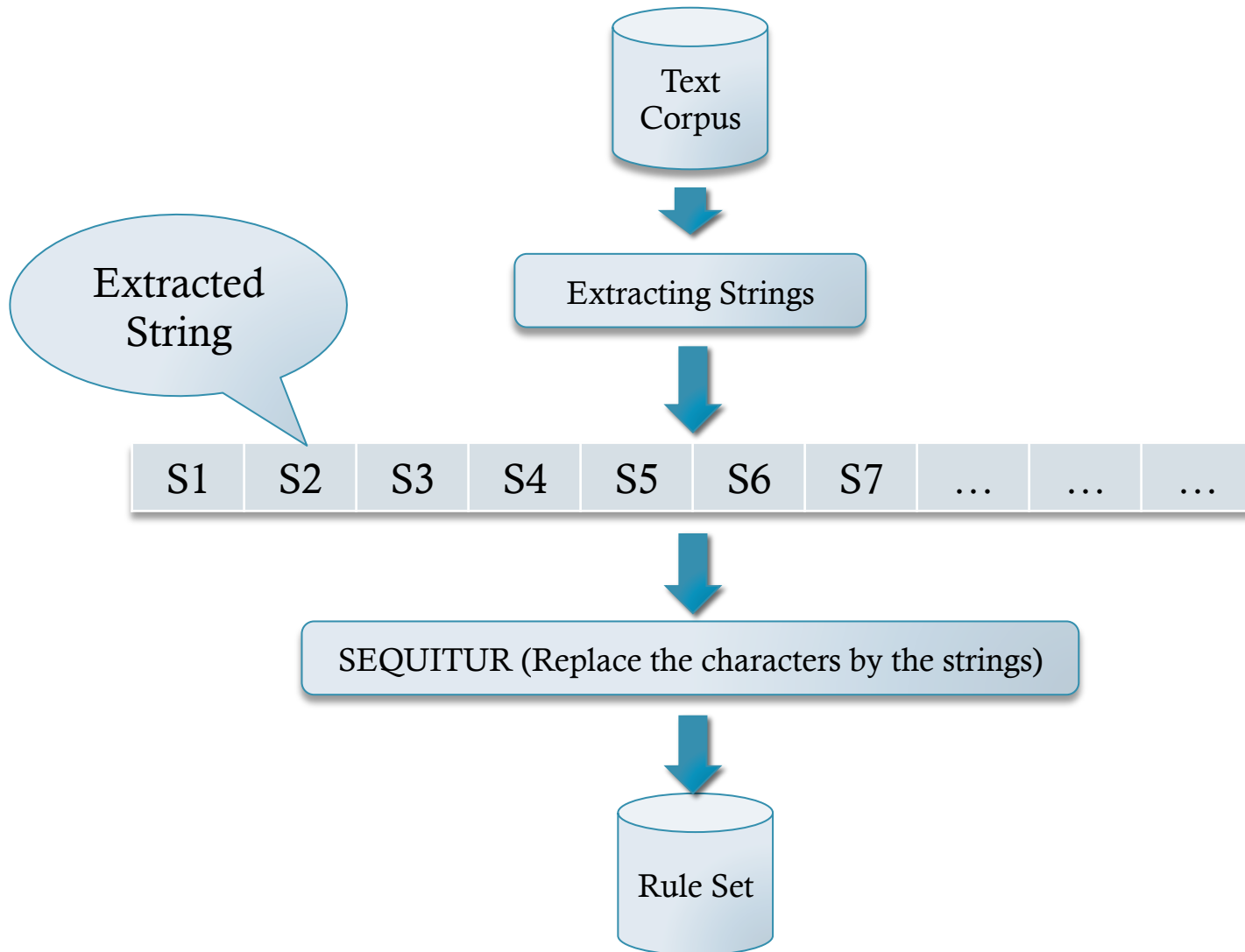
- The SEQUITUR scans through the text and detects the repeated sequence of 2 strings which is appeared more than once. The repeated sequence is replaces by a rule. This action is repeated until there is no repeated sequence found in the text.

- Example:

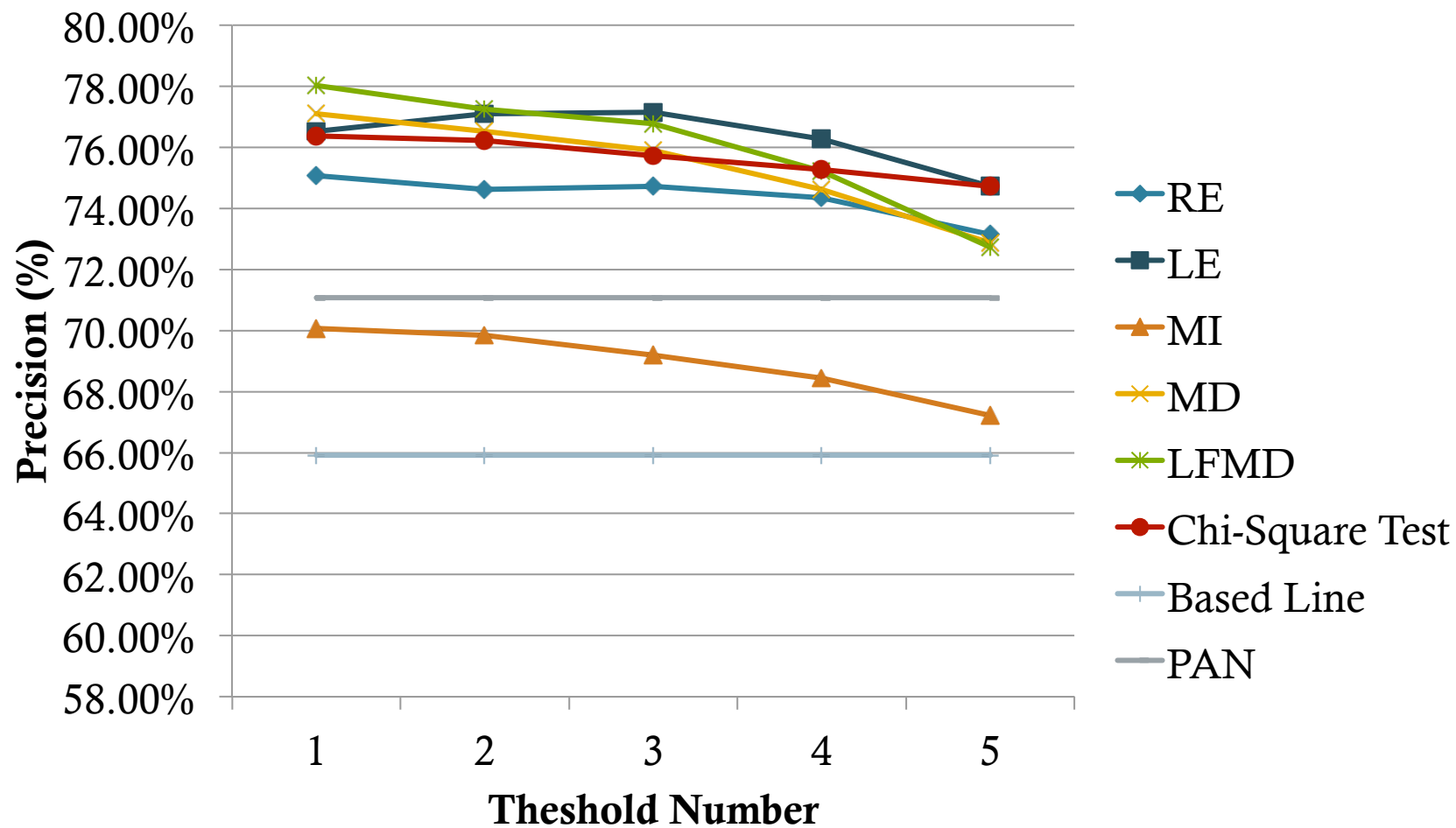
- “abcdbcabcd”

symbol number	the string so far	resulting grammar	remarks
1	a	$S \rightarrow a$	
2	ab	$S \rightarrow ab$	
3	abc	$S \rightarrow abc$	
4	abcd	$S \rightarrow abcd$	
5	abcdb	$S \rightarrow abcdb$	
6	abcdbc	$S \rightarrow abcdbc$	bc appears twice
		$S \rightarrow aAdA$ $A \rightarrow bc$	enforce digram uniqueness
7	abcdbca	$S \rightarrow aAdAa$ $A \rightarrow bc$	
8	abcdbcab	$S \rightarrow aAdAab$ $A \rightarrow bc$	
9	abcdbcabc	$S \rightarrow aAdAabc$ $A \rightarrow bc$	bc appears twice
		$S \rightarrow aAdAaA$ $A \rightarrow bc$	enforce digram uniqueness. aA appears twice
		$S \rightarrow BdAB$ $A \rightarrow bc$ $B \rightarrow aA$	enforce digram uniqueness
10	abcdbcabcd	$S \rightarrow BdABd$ $A \rightarrow bc$ $B \rightarrow aA$	Bd appears twice
		$S \rightarrow CAC$ $A \rightarrow bc$ $B \rightarrow aA$ $C \rightarrow Bd$	enforce digram uniqueness. B is only used once

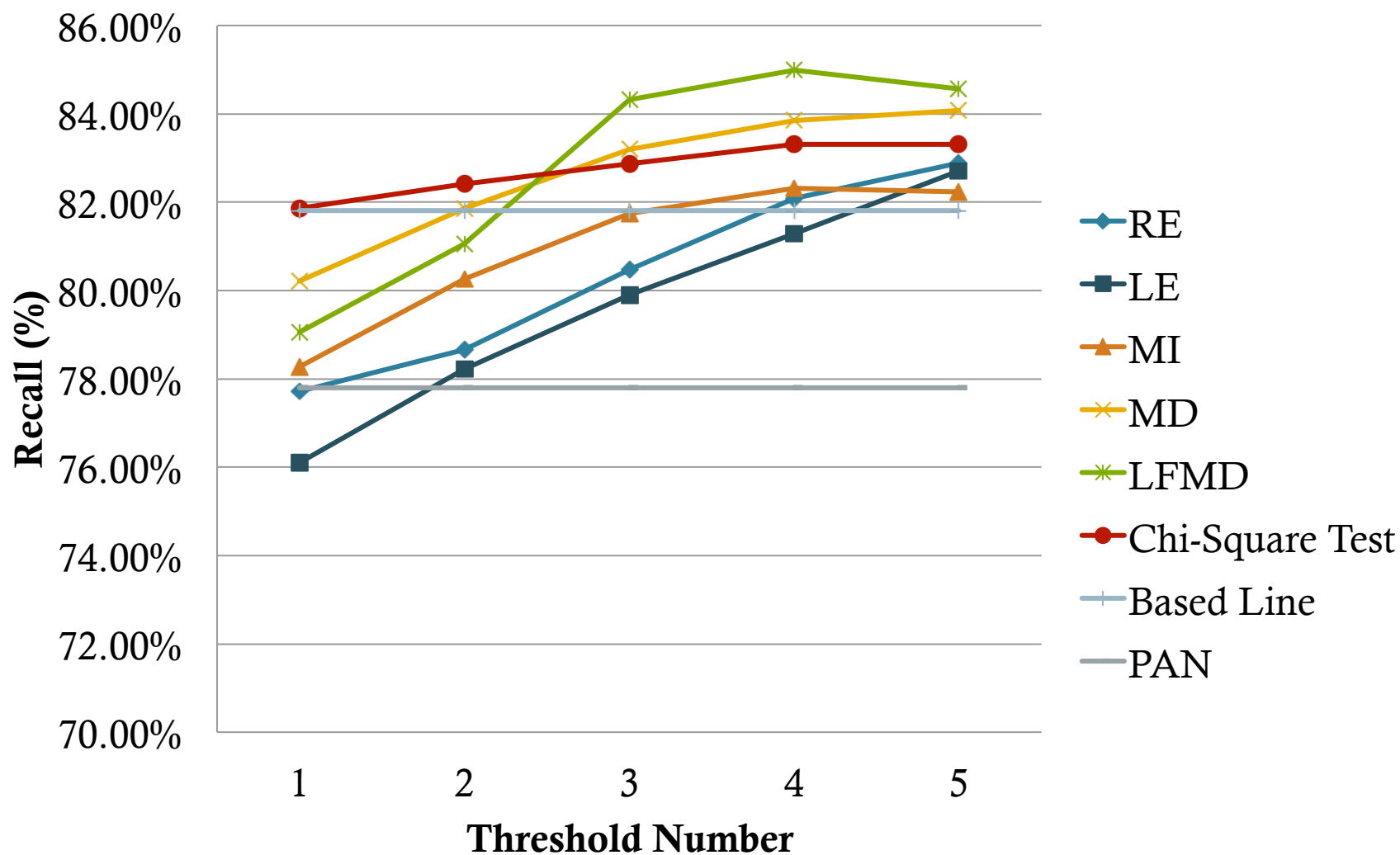
HOW TO EXTRACT RULE FROM THE EXTRACTED STRINGS?



PRECISION RESULTS



RECALL RESULTS



F-MEASURE RESULTS

