Topic-Oriented Information Detection and Scoring

Saike HE

Joint work with Xiaolong ZHENG,
Changli ZHANG, and Lei WANG

ISI Team
Institute of Automation
Chinese Academy of Sciences

Road Map

- Introduction
- Hybrid approach for TOIDS
- Experimental Results
- Conclusions

Road Map

- Introduction
- Hybrid approach for TOIDS
- Experimental Results
- Conclusions

Background

- Information Detection (TOIDS) is a critical task in Intelligence and Security Informatics (ISI)

Background

- Information Detection (TOIDS) is a critical task in Intelligence and Security Informatics (ISI)
- Traditional Solutions

Background

- Information Detection (TOIDS) is a critical task in Intelligence and Security Informatics (ISI)

- Traditional Solutions
  - Dictionary-Based Methods
  - Statistical and Machine Learning Methods

Background

- Information Detection (TOIDS) is a critical task in Intelligence and Security Informatics (ISI)

- Traditional Solutions
  - Dictionary-Based Methods
  - Statistical and Machine Learning Methods

- Drawbacks

Background

- Information Detection (TOIDS) is a critical task in Intelligence and Security Informatics (ISI)

- Traditional Solutions
  - Dictionary-Based Methods
  - Statistical and Machine Learning Methods

- Drawbacks
  - Influenced by the coverage of the lexicon
  - Paleness in domain adaption

Motivation

- Word combination help filter relevant documents with higher accuracy

Motivation

- Word combination help filter relevant documents with higher accuracy
  - Improve precision rate

Motivation

- Word combination help filter relevant documents with higher accuracy
  - Improve precision rate

- Characteristic words

Motivation

- Word combination help filter relevant documents with higher accuracy
  - Improve precision rate

- Characteristic words
  - Related words
  - Unrelated words

Motivation

- Word combination help filter relevant documents with higher accuracy
  - Improve precision rate
- Characteristic words
  - Related words
  - Unrelated words
- Domain adaptation problem

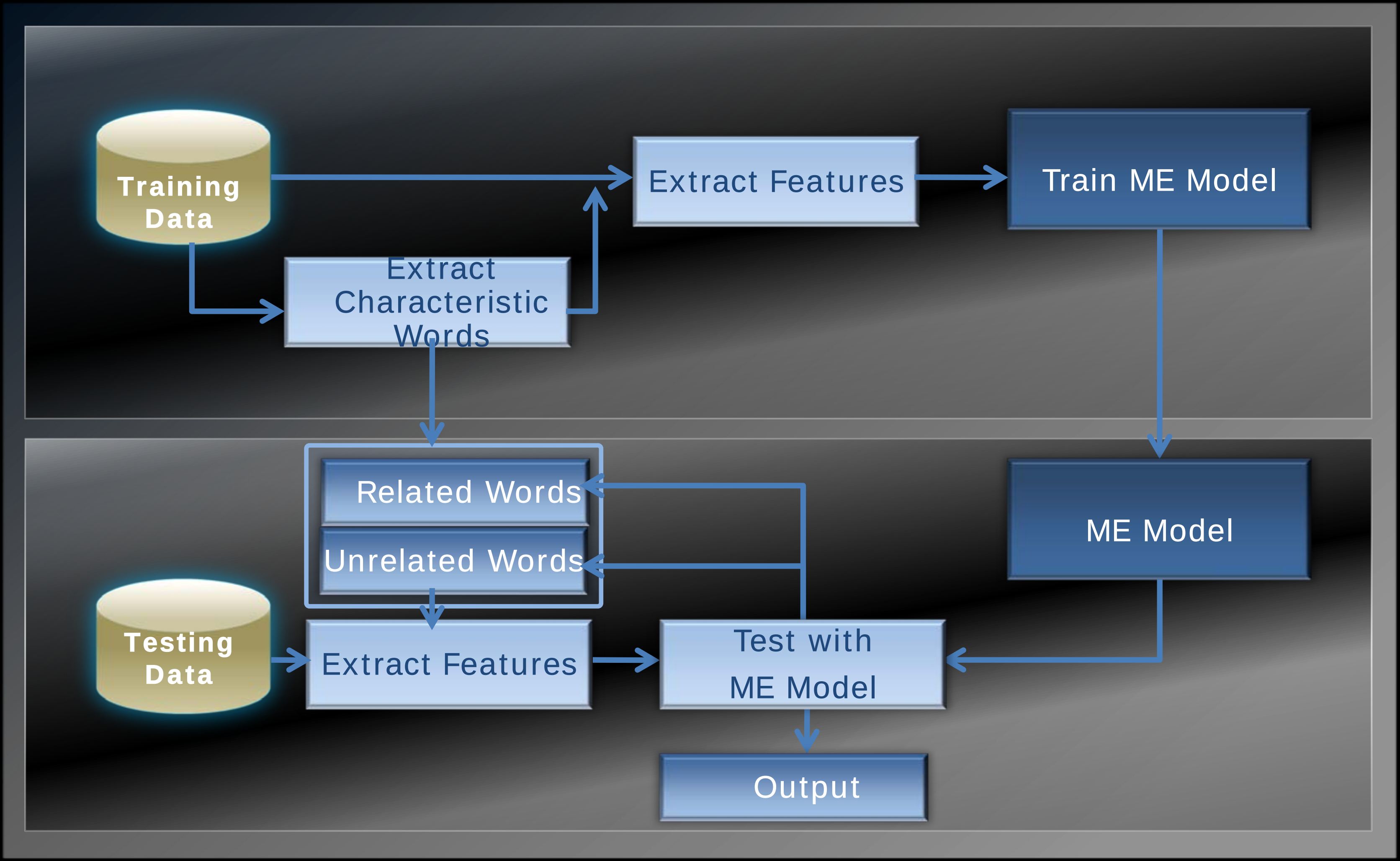
Motivation

- Word combination help filter relevant documents with higher accuracy
  - Improve precision rate
- Characteristic words
  - Related words
  - Unrelated words
- Domain adaptation problem
  - Self learning

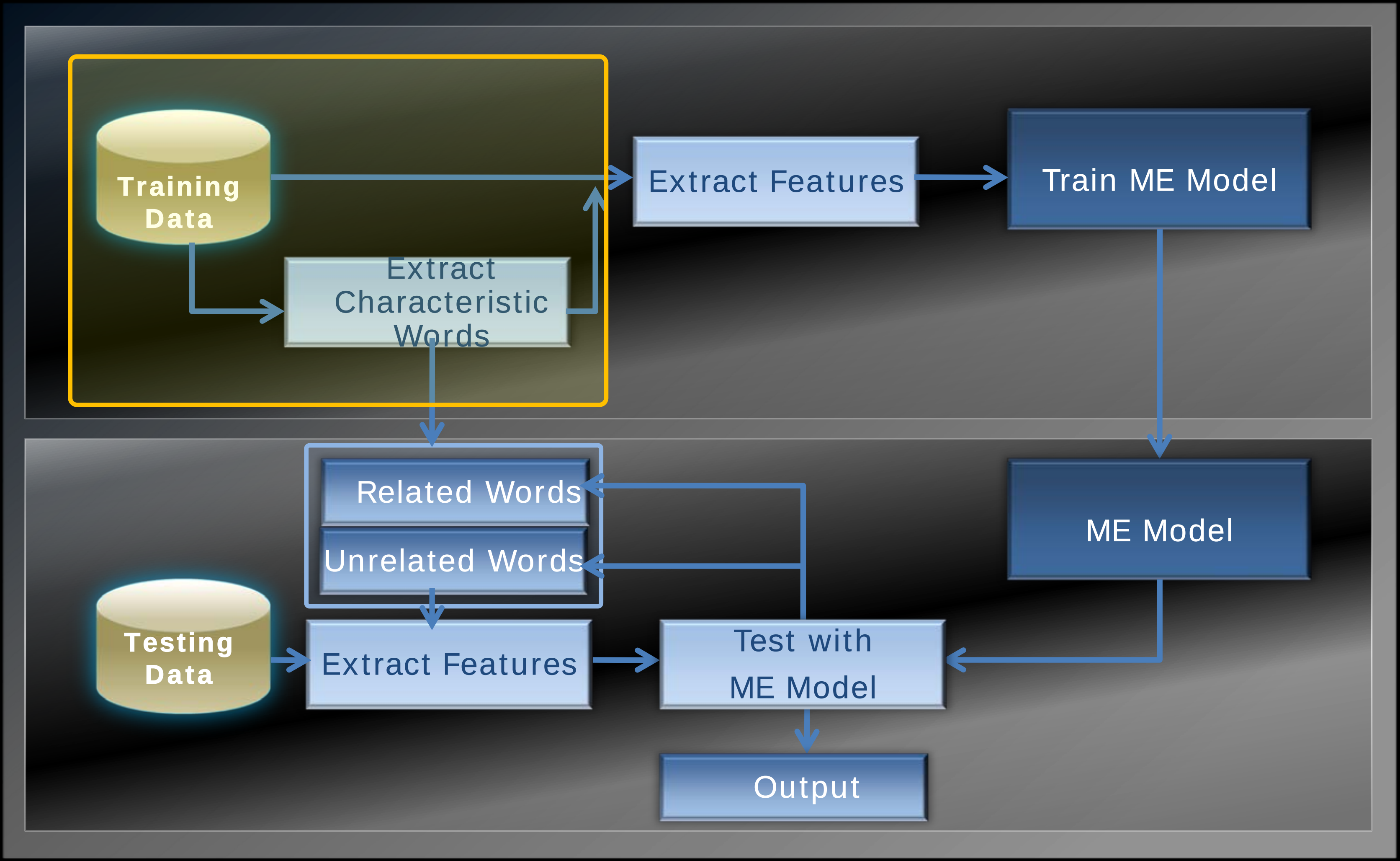
Road Map

- Introduction
- Hybrid approach for TOIDS
- Experimental Results
- Conclusions

Flow chart of TOIDS system



Flow chart of TOIDS system



Z-Score Algorithm

	Topic related	Rest	
ω	a	b	a + b
not ω	c	d	c + d
	a + c	b+ d	n=a+b+c+d

Z-Score Algorithm

	Topic related	Rest	
ω	a	b	a + b
not ω	c	d	c + d
	a + c	b+ d	n=a+b+c+d

$$\mathrm{Zscore}(\omega) = \frac{a - n!\Pr(\omega)}{\sqrt{n!\Pr(\omega) \cdot (1 - \Pr(\omega))}}$$

Z-Score Algorithm

	Topic related	Rest	
ω	a	b	a + b
not ω	c	d	c + d
	a + c	b+ d	n=a+b+c+d

$$\mathrm{Zscore}(\omega) = \frac{a - n!\Pr(\omega)}{\sqrt{n!\Pr(\omega)\cdot(1 - \Pr(\omega))}}$$

- where

$$\Pr(\omega) = (a + b)/n$$

$$n! = a + c$$

Z-Score Example

	Topic related	Rest	
“Bomb”	561	241	802
-“Bomb”	69,324	55,100	124,424
n!= a + c	69,885	55,341	125,226

Z-Score Example

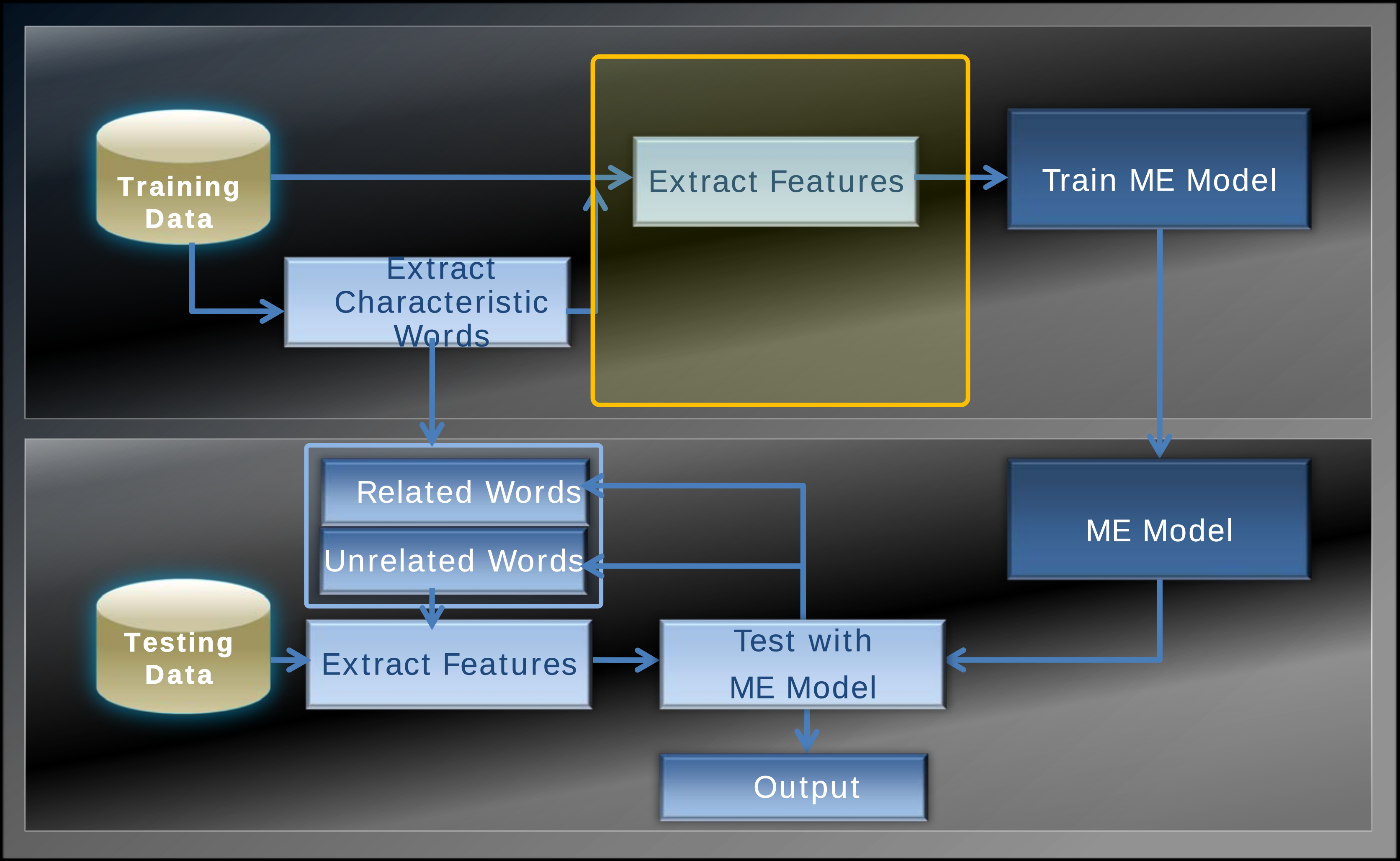
	Topic related	Rest	
“Bomb”	561	241	802
-“Bomb”	69,324	55,100	124,424
n!= a + c	69,885	55,341	125,226

$$Zscore(\omega) = \frac{a - n!Pr(\omega)}{\sqrt{n!Pr(\omega) \cdot (1 - Pr(\omega))}}$$

$$= \frac{561 - 69885 * 802/125226}{\sqrt{69885 * 802/125226 * (1 - 802/125226)}}$$

$$= 5.3787$$

Flow chart of TOIDS system



Training Data

Extract Features

Train ME Model

Extract Characteristic Words

Related Words

Unrelated Words

ME Model

Testing Data

Extract Features

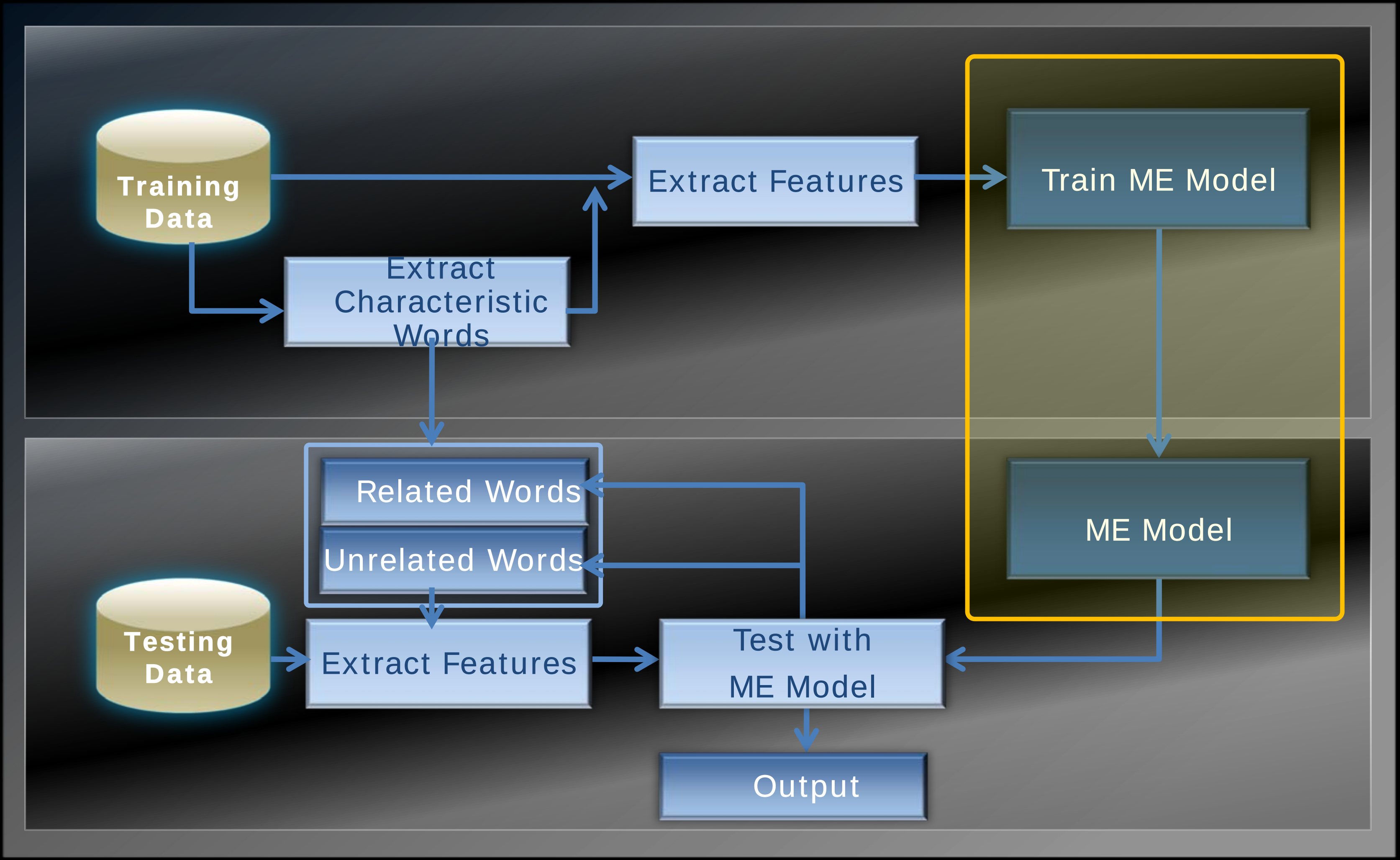
Test with ME Model

Output

Features

Features	Type	Description
n-gram word	Related Unrelated	n-gram for related words n-gram for unrelated words
n-gram POS	Related Unrelated	n-gram for related POS tags n-gram for unrelated POS tags
word number	Related Unrelated	related word number in current sentence unrelated word number in current sentence
major POS	Related Unrelated	POS tag correspond to highest Z-Score value POS tag correspond to lowest Z-Score value

Flow chart of TOIDS system



Statistical Model

- Use the relevance of each sentence to derive a document’s relevance

Statistical Model

- Use the relevance of each sentence to derive a document’s relevance

- To derive the relevance of document d, we introduce a variable Rel_score(d)

Statistical Model

- Use the relevance of each sentence to derive a document’s relevance
- To derive the relevance of document d, we introduce a variable Rel_score(d)

$$\mathrm{Rel}\_score(d) = \#\mathrm{Rel}\_Sentence / \#Sentence$$

Statistical Model

- Use the relevance of each sentence to derive a document's relevance

- To derive the relevance of document d, we introduce a variable Rel_score(d)

$$\mathrm{Rel}\_score(d)=\#Rel\_Sentence/\#Sentence$$

- Maximum Entropy Model(MEM)

Statistical Model

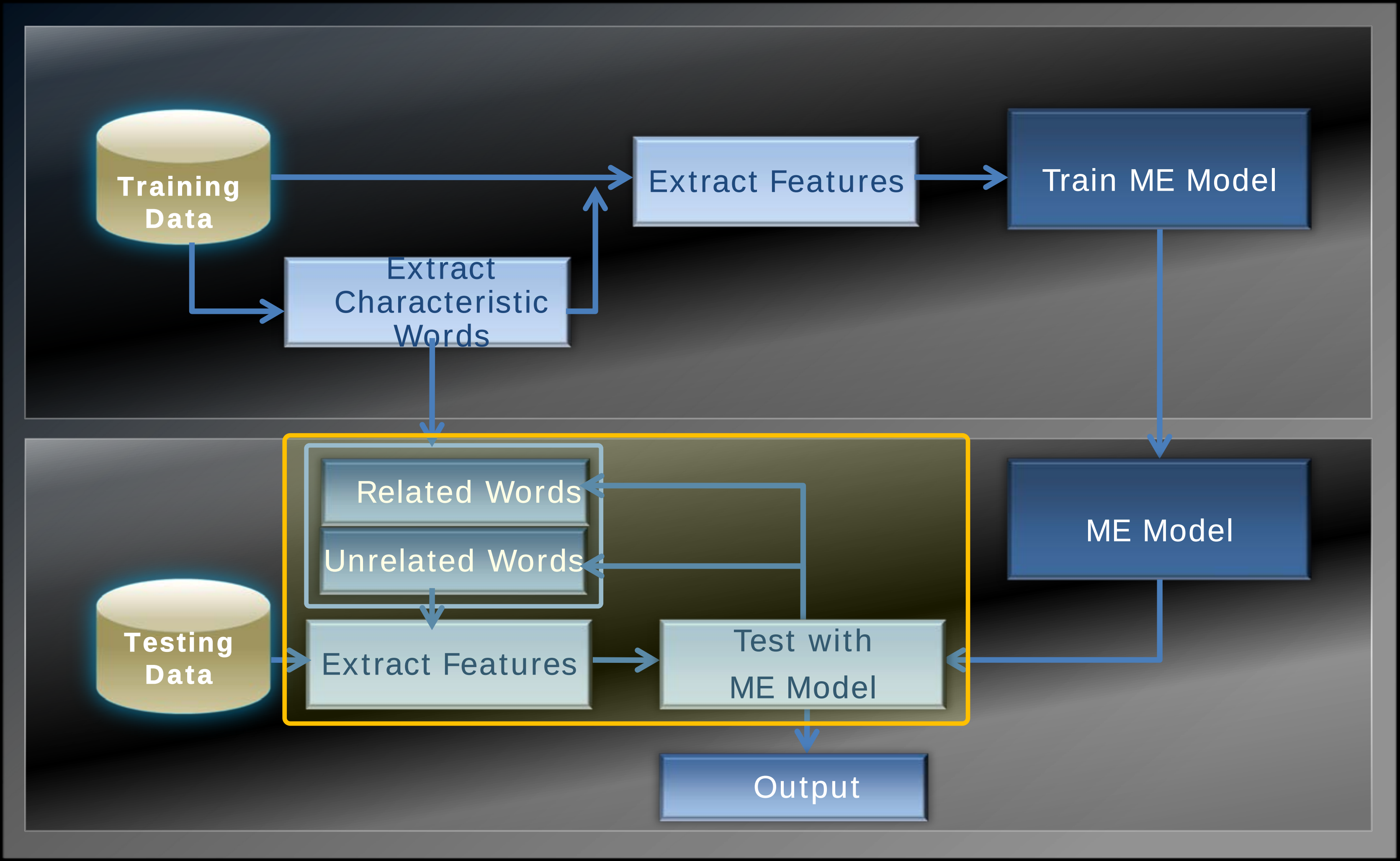
- Use the relevance of each sentence to derive a document's relevance

- To derive the relevance of document d, we introduce a variable Rel_score(d)

$$\mathrm{Rel\_}score(d) = \#\mathrm{Rel\_}Sentence/\#Sentence$$

- Maximum Entropy Model(MEM)
  - Scoring based on Rel_score
  - 5 degree levels

Flow chart of TOIDS system



Self Learning

- Augment characteristic word lexicon based on prediction results

Self Learning

- Augment characteristic word lexicon based on prediction results
  - Words that occur in related sentences yet not found in unrelated word lexicon are added to related word lexicon

Self Learning

- Augment characteristic word lexicon based on prediction results
  - Words that occur in related sentences yet not found in unrelated word lexicon are added to related word lexicon
  - Words that occur in unrelated sentences yet not found in related word lexicon are added to unrelated word lexicon

Road Map

- Introduction
- Hybrid approach for TOIDS
- Experimental Results
- Conclusions

Experiment Setting

- Corpora

Experiment Setting

- Corpora
  - About 5000 webpage documents from 10 websites

Experiment Setting

- Corpora
  - About 5000 webpage documents from 10 websites
  - Documents are crawled from websites directly rather than retrieval with specified key words

Experiment Setting

- Corpora
  - About 5000 webpage documents from 10 websites
  - Documents are crawled from websites directly rather than retrieval with specified key words
  - Manually labeled

Experiment Setting

- Corpora
  - About 5000 webpage documents from 10 websites
  - Documents are crawled from websites directly rather than retrieval with specified key words
  - Manually labeled
  - About 20 percent of topic related documents

System comparison under different configurations

Round	Precision	Recall	F-Score
ME + SingleRelWordMatch	62.38	84.36	75.49
ME + RelatedWordCom	79.84	82.75	81.76
ME + CharacteristicWordCom	82.30	82.81	82.64
ME + CharacteristicWordCom + Self-Learning	83.49	83.02	83.18

Experiment Setting

- Domain Adaptation

Experiment Setting

- Domain Adaptation
  - Training: one sub-collection related to transportation (500 documents)

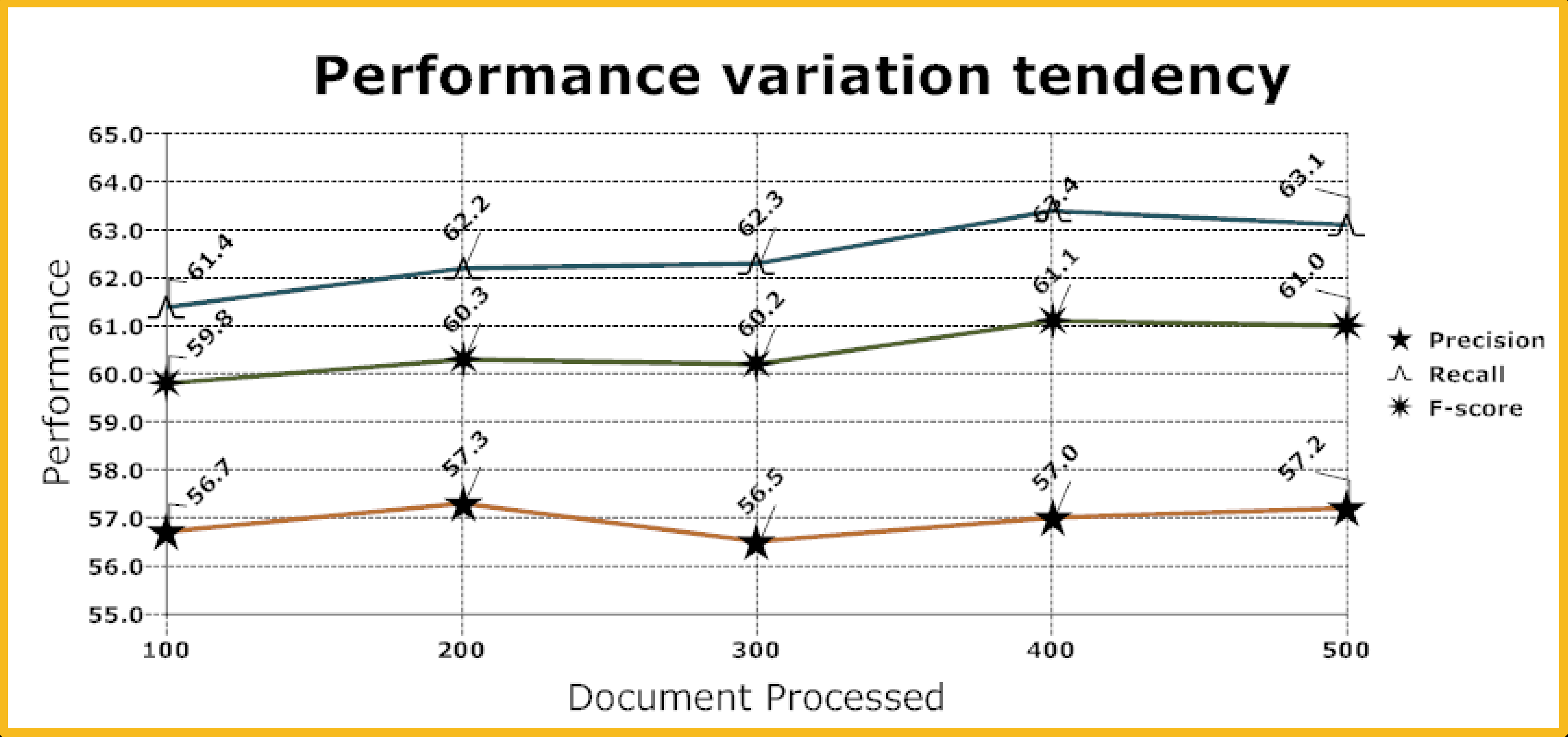
Experiment Setting

- Domain Adaptation
  - Training: one sub-collection related to transportation (500 documents)
  - Testing: one concerning criminal incidents (400 documents)

Experiment Setting

- Domain Adaptation
  - Training: one sub-collection related to transportation (500 documents)
  - Testing: one concerning criminal incidents (400 documents)
  - Measurement: whenever one hundred new documents were classified, F-score is recalculated over all testing documents processed till the current time

Performance variation tenden



Scoring results from TOIDS

Title	Rel_score
一颗子弹 马英九连战矛盾面临 … One Bullet The contradiction between Ma Ying-jeou and Lien Chan faces …	5
男子为讨68.6万贷款持刀 … For debt collection of 686,000, men armed with knives …	4
河南交通厅长董永安落马… Transport Minister in Henan province Dong Yongan collapses …	5
被囚俄罗斯寡头能把2亿… Jailed Russian oligarch uses 200,000,000 …	3

Road Map

- Introduction
- Hybrid approach for TOIDS
- Experimental Results
- Conclusions

Future work

Future work

1 Use quantified Z-score of characteristic words to judge a sentence’s relevance

Future work

1 Use quantified Z-score of characteristic words to judge a sentence's relevance

2 Distinguish the importance of sentences occurring at different positions

Future work

1 Use quantified Z-score of characteristic words to judge a sentence’s relevance

2 Distinguish the importance of sentences occurring at different positions

3 Implement mutual enhancement mechanism between sentence and document

Future work

1. Use quantified Z-score of characteristic words to judge a sentence's relevance
2. Distinguish the importance of sentences occurring at different positions
3. Implement mutual enhancement mechanism between sentence and document
4. Enable mistakenly extracted characteristic words eliminated automatically

Thank you!

Thank you!

Questions?