上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Chinese Opinion Mining in JRLLT: Exploration of Fine-Grained OM Approaches

Tianfang Yao

Dept. of Computer Science and Engineering

Shanghai Jiao Tong University

Shanghai, China

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Outline

- Brief Introduction to JRLLT
- Text Annotation
- Preprocessing and Classification for Subjective Texts
- Recognition of Topics
- Analysis of Sentiment of Words and Sentences
- Identification of Topic-Sentiment Relations
- Conclusion

Brief Introduction to JRLLT

- UdS-SJTU Joint Research Lab for Language Technology (JRLLT)

The Department of Computational Linguistics and Phonetics

The Department of Computer Science and Engineering

Brief Introduction to JRLLT

- It was established on March 10, 2005.

Brief Introduction to JRLLT

- It is located in the Dept. of Computer Science and Engineering.

Brief Introduction to JRLLT

- Organization and Staffing
  - Heads of the Lab
    - Prof. Huanye Sheng
    - Prof. Dr. Hans Uszkoreit
  - Principal Researchers
    - Associate Prof. Dr. Fang Li
    - Dr. Feiyu Xu
    - Associate Prof. Dr. Tianfang Yao
  - Other Members
    - PhD and Master Candidates

Brief Introduction to JRLLT

- Mission of the Lab
  - provide a platform for defining and carrying out projects supported by public sources as well as private enterprises in China, Germany and elsewhere.
  - shorten the path of new technologies developed by UdS and SJTU from the lab into deployed systems and products suitable for the Chinese and international markets.

Brief Introduction to JRLLT

- Research Directions of the Lab
  - Language Technology Research
  - Web Mining Applications
  - Chinese Information Processing
  - Information Services

Brief Introduction to JRLLT

- Granted Projects
  - **NSFC (National Natural Science Foundation of China) Projects**
    - Research on Fine-grained Opinion Mining for Chinese Texts (2008-2010)
    - Threads and Topics Detection for News Events (2009–2011)
  - **Enterprise Projects**
    - Research and Application of Data Mining Technology under Big Data Environment (2014-2015)

Brief Introduction to JRLLT

- WISCOM Data Service Platform (2012-2013)
- Research on the Application of Time-space based Chinese Information Retrieval Model (2011–2012)
- Research on the Application of Cultural Entity and Relationship for National Geographical Names (2010–2011)
- Teenage Interests Analysis Based on Web Access Log (2007–2008)
- An Intelligent Platform for Information Retrieval (2004–2006)

Brief Introduction to JRLLT

- **NSFC Project:** Research on Fine-grained Opinion Mining for Chinese Texts (Grant No.: 60773087)
  - Subjectivity text is very different from objective text.
  - No existing annotation corpus is available.
  - Most of text opinion mining approaches are coarse, especially for Chinese texts.
  - There are still problems to be resolved, such as implicit topic.
  - Accuracy and robustness of opinion mining approaches are not ideal.

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

- Subjective Text Types (Liu, Yao et al., 2008)
  - According to the **standardization**, **domain** and **amount of text available**, we can category subjective texts.

Standardization Degree	**Text Form**
Standard	novel, essay, poetry, review, readers'letter, news note, interview report, scientific paper, regular book review / film review, poem comment
Fairly Standard	diary, composition, mail, correspondence
Non - Standard	blog, forum, BBS, reader review / film review, personal home page, chat log, post bar

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

Category	Text Form	
Domain	single domain	product forum, reader review, scientific paper
	multiple domain	synthetical forum, BBS, novel, essay, poetry, diary, composition, newspaper review, readers'letter, interview, blog
Amount of Text Available	more	forum, BBS, reader review / film review, blog, personal page / space, chat log, post bar
	general	diary, composition, newspaper review, readers'letter, interview, news chronicle, scientific paper, regular book review / film review, mail
	less	novel, essay, poem, poem comment, letter

Text Annotation

- Texts collected should contain a higher proportion of opinion sentences
- Text size should be large enough for experiment
- Text content difficulty should be controlled within the usability of current mining technology
- Text domain should be involved in a wide range

- Subjective Sentence Types (Huang, Yao et al., 2008)

	Topic	Sentiment
Sentence Structure (Opinion Elements)	single	single
	single	multiple
	multiple	single
	multiple	multiple

Text Annotation

	Topic	Sentiment
Sentence Structure (Opinion Elements)	explicit	explicit
	explicit	implicit
	implicit	explicit
	implicit	implicit

- Annotation Tags
  - Word Segmentation (Deparser)
  - POS Tags (National Standard GB/T 20532-2006)
  - Dependency Types (Deparser)
  - Opinion Elements (Self-Defined)

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

Level	Tags	Attribute	Description
Document	text	no./textdom/txttype	Text number/domain/type
Document	paragraph	pnum/ptype	Paragraph number/type
Sentence	sentence	stnnum	Sentence number
Sentence	claim	cnum/ctype	Claim number/type
Sub-sentence	holder	hnum/htype	Holder number/type
Sub-sentence	topic	tnum/ttype	Topic number/type
Sub-sentence	sentiment	stnum/stmtype/stmpolt/stmstrn	Sentiment number/type/polarity/strength
Sub-sentence	relation	rtype/tnum/hnum/stmnum/deprel/deppair	Relation type/number/dependence relation/dependence pair
Word	stmword	wnum/wpos/wpolt/wstrn/wfunc/wsem	Sentiment word number/POS/polarity/strength/function/semantic
Word	gword	wnum/wpos/wfunc/wsem	General word number/POS/function/semantic

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

- An Example for Annotated Sentence
  - For example：我不喜欢这车。价格太贵。

    I do not like this **car**. The price is too high.

    - **topic**
    - **sentiment**

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

```
<claim cnum="1" ctype="1t+2s">
  <sentence stnnum="1">
    <holder hnum="1" htype="coref">
      <gword wnum="1" wpos="r" wfunc="null" wsem="FirstPerson">
        我(I)
      </gword>
      <relation type="h-t" hnum="1" tnum="1" deprel="null" deppair="null">
      </relation>
    </holder>
    <sentiment stmnum="1" stmtype="privat" stmpolt="negative" stmstrn="-1">
      <gword wnum="2-3" wpos="d" wfunc="privat" wsem="neg">
        不(do not)
      </gword>
      <stmword wnum="3" wpos="v" wpolt="positive" wstrn="+2" wfunc="core" wsem="FondOf">
        喜欢(like)
      </stmword>
    </sentiment>
    <other>
      <gword wnum="4" wpos="r" wfunc="null" wsem="aValue|kind|special">这(this)</gword>
    </other>
    <topic tnum="1" ttype="substance">
      <gword wnum="5" wpos="n" wfunc="core" wsem="LandVehicle">
        车(car)
      </gword>
      <relation type="t-s" tnum="1" stmnum="1" deprel="VOB" deppair="s:t">
      </relation>
    </topic>
    <other>
      <punctuation pnum="6" pos="wp">。 </punctuation>
    </other>
  </sentence>
```

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

```
<sentence stnnum="2">
  <topic tnum="1.1" ttype="property">
    <gword wnum="1" wpos="n" wfunc="core" wsem="attribute|price|artifact|commercial">
      价钱(The price)
    </gword>
    <relation type="t-s" tnum="1.1" stmnum="2" deprel="SBV" deppair="s:t">
    </relation>
  </topic>
  <sentiment stmnum="2" stmtype="inten" stmpolt="negative" stmstrn="-4">
    <gword wnum="2-3" wpos="d" wfunc="inten" wsem="aValue|degree|very">
      太(too)
    </gword>
    <stmword wnum="3" wpos="a" wpolt="negative" wstrn="-2" wfunc="core" wsem="aValue|price|expensive|undesired">
      贵(high)
    </stmword>
  </sentiment>
  <other>
    <punctuation pnum="4" pos="wp">。</punctuation>
  </other>
</sentence>
</claim>
```

Text Annotation

- Annotation Tool (Wu, 2008)
  - Improve the efficiency of annotation
  - Improve the flexibility of annotation
  - Beneficial to the extension of tag set
  - Automatically check the consistency of annotation
  - Improve the quality of annotation
  - Steps

Text Annotation

Text Annotation

显示结果

文件(F)　标注(M)

保存　上一步　新任务

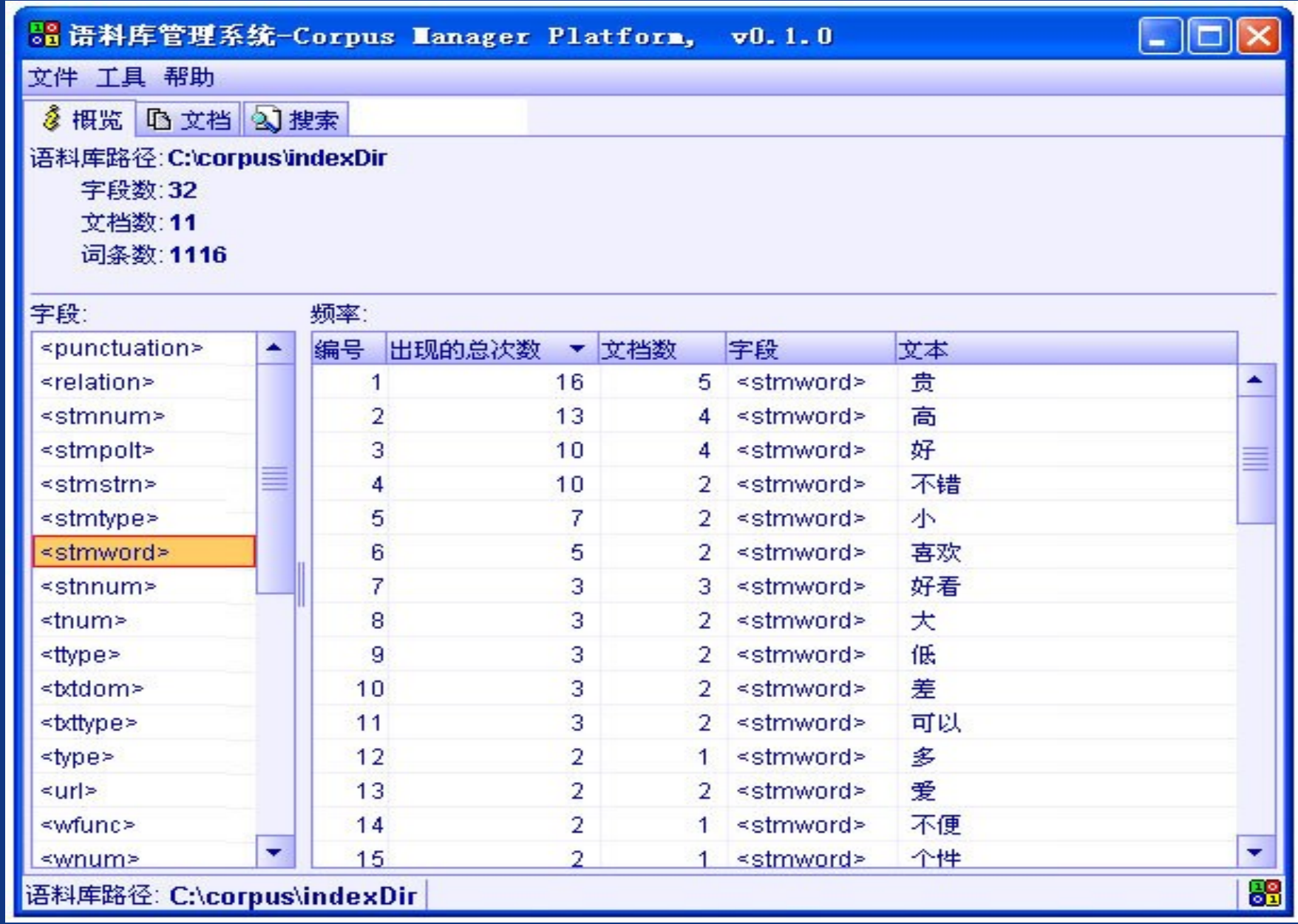
```
<text no.="1" txtdom="car" txttype="single">
<paragraph pnum="1" ptype="multiple">
<claim cnum="1" ctype="1s">
<sentence stnnum="1">
<other>
<gword wnum="1" wpos="r" wfunc="null" wsem="ThirdPerson|male">他
</gword>
</other>
<other>
<gword wnum="2" wpos="d" wfunc="privat" wsem="neg">不</gword>
</other>
<other>
<gword wnum="3" wpos="v" wfunc="null" wsem="be">是</gword>
```

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Text Annotation

- Annotation Corpus (Song, Liu et al., 2008)
  - Domain: Automobile Forum Texts
  - Size: About 800 Sentences
  - Corpus Management System
  - Statistics（General Information, Subjective Information）
  - Retrieval（Keyword, Context）
  - Others

Text Annotation



Preprocessing and Classification

- Network Informal Language (NIL)
  - The subjective text is an important processed object of opinion mining, but sometimes there are many informal expressions in a subjective text.
  - The authors of the subjective texts have the personal expression habits which are not restricted to a formal grammar, so Network Informal Language (NIL) emerges.

Preprocessing and Classification

- NIL Words and Expression Forms (Zhang and Yao, 2008; Yao and Zhang, 2009)

Words Formation	Number	Example
Abbreviation of English or Chinese Pinyin	127	"**GF**" = 女朋友 (Girl Friend)
Homophonic of Chinese Words	36	"斑竹"(bamboo) = "版主"(web moderator)
Transliteration and Foreign expression	38	"粉丝" (noddles) = "歌迷" (fans)
Partial Tone in Chinese	148	"稀饭"(porridge) = "喜欢" (like)
Numbers	39	"**94**" = "就是"(precisely)
Mixture of the above Forms	54	"**3q**" = "谢谢" (thank you)

Note: The statistics for the above words formation comes from 4,315 web pages.

Preprocessing and Classification

- Categories of NIL
  - Typical Informal Chinese Word
    - Those that contain letters, numbers and mixed abbreviated forms represent the words of standard language.
  - Fuzzy Informal Chinese Word
    - Those are standard form of vocabulary literally, but in fact they express the meaning of non-standard words.

Preprocessing and Classification

- Experiments for Recognition of NIL
  - Recognition for Typical Informal Chinese Words
    - Rule based Sequential Covering Algorithm (that is, it is a rule based classification algorithm)

Data Set	Rec.	Prec.	F-Mea.
Typical NIL Data Set	0.682	0.871	0.765

Note: Training Set: 5791 Sentences; Test Set: 3722 Sentences.

  - Recognition for Fuzzy Informal Chinese Words
    - NB based Classification Model

Preprocessing and Classification

– SVM based Classification Model

Feature Combination	Naive Bayes (NB)			Support Vector Machine (SVM)		
	Rec.	Prec.	F-Mea.	Rec.	Prec.	F-Mea.
F_a+F_b	0.531	0.662	0.589	0.476	0.711	0.570
F_a+F_c	0.708	0.677	0.692	0.635	0.783	0.701
F_a+F_d	0.675	0.813	0.738	0.832	0.806	0.819
$F_a+F_c+F_e$	0.734	0.810	0.770	0.794	0.765	0.779
$F_a+F_c+F_d$	0.847	**0.926**	**0.885**	**0.887**	0.821	0.853

Note: Ten-fold Cross Validation - 9513 Sentences.

a. The typical NIL words; b. The words which express opinions, suggestions or contain sentiment; c. The first and second person; d. Irregular punctuations; e. The punctuations with sentiment.

Preprocessing and Classification

- Subjective Text
  - It gives non-factual description.
  - It contains the opinions, emotions and attitudes from individual, group, or organization.
- Classification Approaches (Liu, Liu et al., 2008)
  - Adopt rules, SVM and NB model to classify texts respectively
  - Synthesize results according to the weight of texts (The more feature is, the higher weight is)

Preprocessing and Classification

- The feature used for rule-based and SVM classification model (The NB classification model uses unigram feature (That is, high frequency words and sentiment words))

s established on March 10, 2005.

Preprocessing and Classification

Feature	Specification
personal pronoun	the first, second and third personal pronoun
suggestion word	such as suggestion verb
NIL word and phrase	they come from Web
special punctuation	such as exclamatory mark, question mark
non-accurate digit	accurate digit is used for the feature of objective text
sentiment word	positive and negative words from HowNet
interjection	include interjection and modal particle
symbol string	this feature is used for human's free writing

Preprocessing and Classification

- COAE 2008 (Chinese Opinion Analysis Evaluation)
  - Classify 39,976 texts into subjective and objective texts, then select 4,000 subjective texts with maximum confidence.
  - Our method: adopt rule-based classification method to assign weight score (according to the number of feature) to subjective texts; use SVM classification model to obtain 18,862 subjective texts and 21,114

Preprocessing and Classification

objective texts; use NB classification model to get 15,734 subjective texts and 24,242 objective texts. From the above three methods' result, we take 4,000 texts respectively, remove redundant texts, and remain 6,219 texts. Then use rule-based method to filter final 4,000 texts with higher weight.

- Test and Result:
  - SJTUCSTask4Run1: The method has been mentioned above.

Preprocessing and Classification

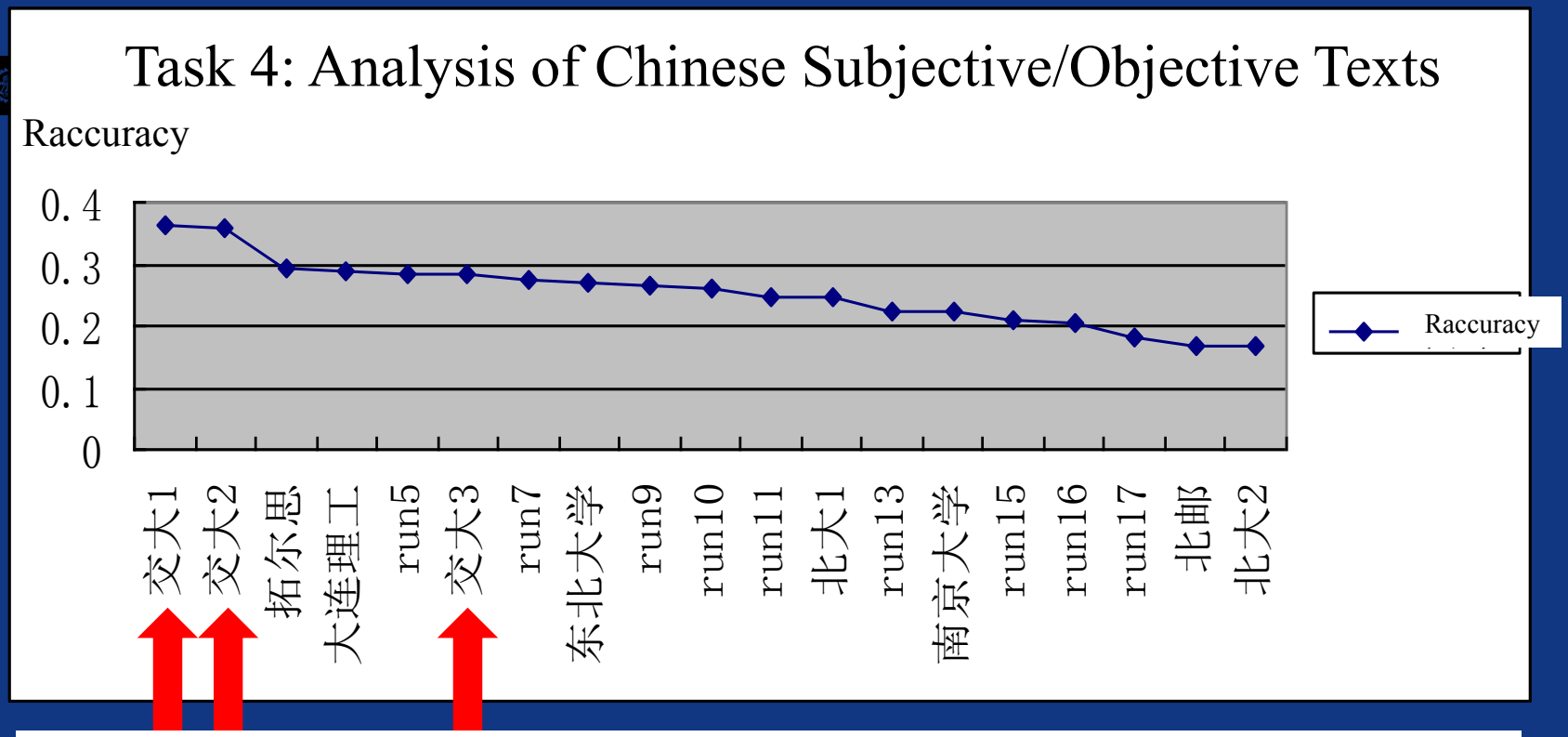
- SJTUCSTask4Run2: Only use rule-based classification method to select former 4000 texts.
- SJTUCSTask4Run3: Similar to the the method of SJTUCSTask4Run1, but add the pu
  long text (be
  score than s

Raccuracy, Acc10 and Acc1000 mean that the former 500, 10 and 1000 texts are selected for human evaluation respectively.

accuracy_by_lexicon means that 4000 texts are automatically evaluated by machine according to the lexicon

Evaluation Results / 评 / 果 / Evaluation Metrics	Raccuracy	Acc10	Acc1000	accuracy_by_lexicon
SJTUCSTask4Run1	0.363	0.9	0.698	0.8405
SJTUCSTask4Run2	0.2838	1	0.541	0.82875
SJTUCSTask4Run3	0.3596	0.9	0.697	0.83025
Median	0.2488	0.4	0.387	0.84
Best	0.363	1	0.698	0.95075

Preprocessing and Classification



Our three results (Totally, 13 organizations including universities, institutes and companies participated in the evaluation)

Recognition of Topics

- Recognition in Specific Domain (Automobile) (Yin and Yao, 2008)
  - Utilize annotated corpus, we do lexical analysis and syntactic dependency analysis.
  - In addition, we use the relations between sentiment words and topics as well as ontology to recognize non-identified potential topics.
  - The experimental result is shown as follows:

Recognition of Topics

	Sentence Size	Number of Topics	Recall	Precision
Syntactical Analysis based Approach	450	1080	0.792	0.74
Sentiment Word based Improved Approach	450	1080	**0.852**	**0.802**

Recognition of Topics

- Recognition in Multiple Domains (Car, Laptop, Mobile Phone, Digital Camera)
  - Adopt SVM classification model and POS & WSD (Word Semantic Disambiguity) features.
  - Utilize COAE 2008 training data, we use five-fold cross validation method to train data for SVM model.
  - For WSD features, we select different feature combination, that is, WSD1, WSD2, WSD3 are the feature combinations from coarse-grain to fine-grain

Recognition of Topics

features.

- The experimental result is shown as follows:

Recognition of Topics

Domains	Features	Precision	Recall	F-measure	Improvements
Cars	PoS(Baseline1)	0.0958	0.8414	0.1716	-
	Baseline1+WSD1	0.1185	0.8660	0.2080	**21.21%**
	Baseline1+WSD2	0.1489	0.8592	0.2531	**47.49%**
	Baseline1+WSD3	0.1840	0.8518	0.3017	**75.81%**
	PoS+Word(Baseline2)	0.2223	0.6908	0.3336	-
	Baseline2+WSD1	0.2325	0.7584	0.3536	**5.99%**
	Baseline2+WSD2	0.2364	0.7947	0.3622	**8.59%**
	Baseline2+WSD3	0.2576	0.7811	0.3852	**15.46%**
Laptop Computers	PoS(Baseline1)	0.1502	0.8680	0.2559	-
	Baseline1+WSD1	0.1882	0.8599	0.3084	**20.49%**
	Baseline1+WSD2	0.2096	0.8537	0.3360	**31.29%**
	Baseline1+WSD3	0.2454	0.8409	0.3789	**48.06%**
	PoS+Word(Baseline2)	0.2834	0.7364	0.4086	-
	Baseline2+WSD1	0.3007	0.7773	0.4328	**5.92%**
	Baseline2+WSD2	0.3093	0.8070	0.4465	**9.27%**
	Baseline2+WSD3	0.3374	0.7956	0.4728	**15.72%**
Mobile Phones	PoS(Baseline1)	0.1781	0.8726	0.2954	-
	Baseline1+WSD1	0.2101	0.8709	0.3379	**14.38%**
	Baseline1+WSD2	0.2290	0.8775	0.3625	**22.70%**
	Baseline1+WSD3	0.2640	0.8666	0.4040	**36.74%**
	PoS+Word(Baseline2)	0.2894	0.7913	0.4224	-
	Baseline2+WSD1	0.3009	0.8230	0.4392	**3.98%**
	Baseline2+WSD2	0.3012	0.8395	0.4420	**4.63%**
	Baseline2+WSD3	0.3225	0.8277	0.4629	**9.58%**
Digital Cameras	PoS(Baseline1)	0.1668	0.8521	0.2784	-
	Baseline1+WSD1	0.1997	0.8633	0.3235	**16.22%**
	Baseline1+WSD2	0.2197	0.8753	0.3505	**25.90%**
	Baseline1+WSD3	0.2591	0.8735	0.3987	**43.20%**
	PoS+Word(Baseline2)	0.2882	0.7773	0.4187	-
	Baseline2+WSD1	0.3028	0.8213	0.4408	**5.28%**
	Baseline2+WSD2	0.3075	0.8422	0.4488	**7.20%**
	Baseline2+WSD3	0.3277	0.8319	0.4686	**11.92%**

Analysis of Sentiment of Words and Sentences

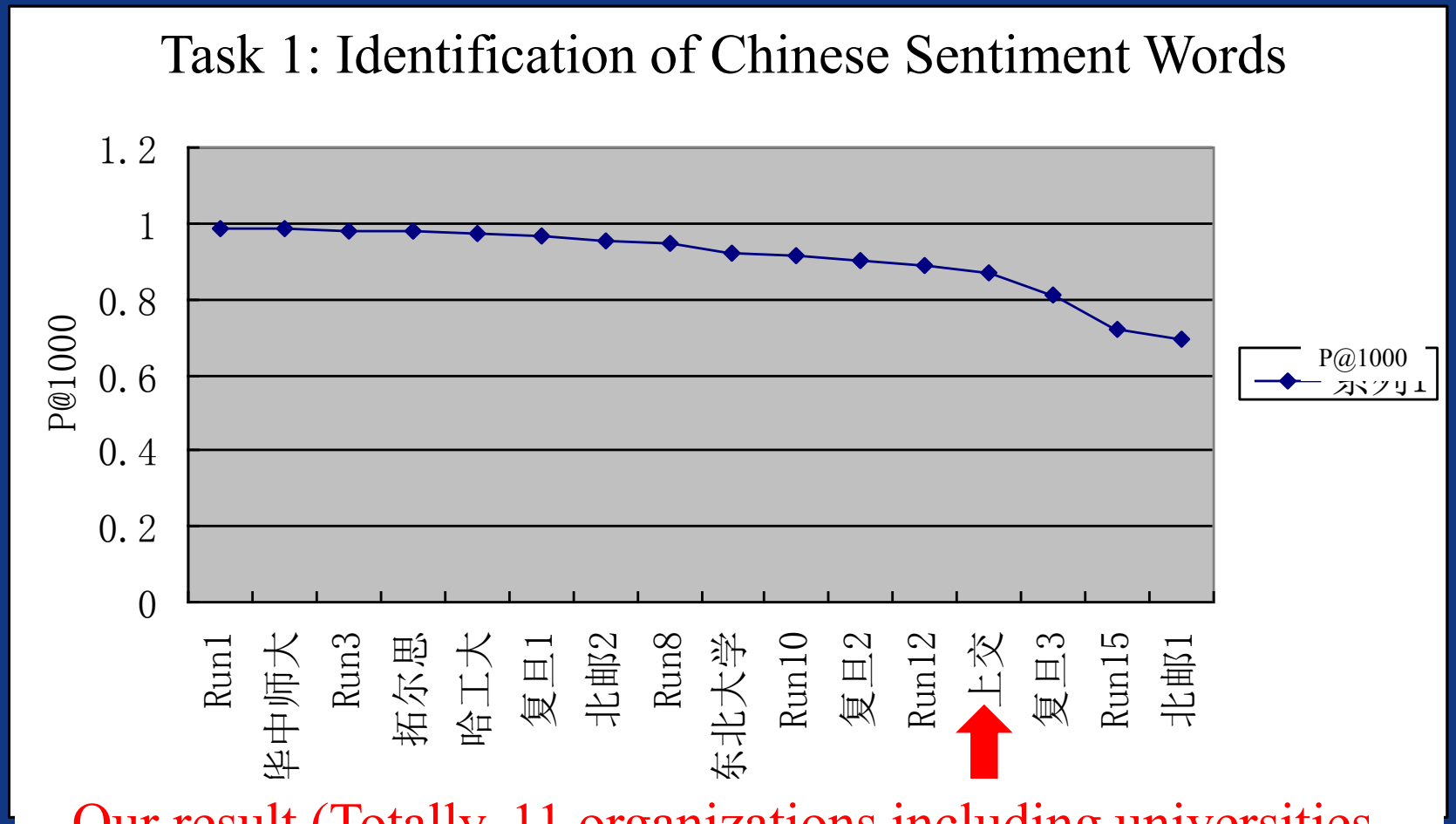
- Sentiment Word and Polarity Identification (Liu, Liu et al., 2008)
  - Identification Approach
    - With the help of DeParser, build parsing trees;
    - For verbs, adjectives, adverbs, nouns, Idioms:
      - ✓ Check whether there is the dependence relation between constituents related to the above word classes (POS)

Analysis of Sentiment of Words and Sentences

- ✓ Find topics which corresponds to sentiment words
- ✓ If there exists the above sentiment word in sentiment lexicon, according to the type of sentiment words, decide its polarity
- ✓ If there is unknown word, record it for artificial judging later

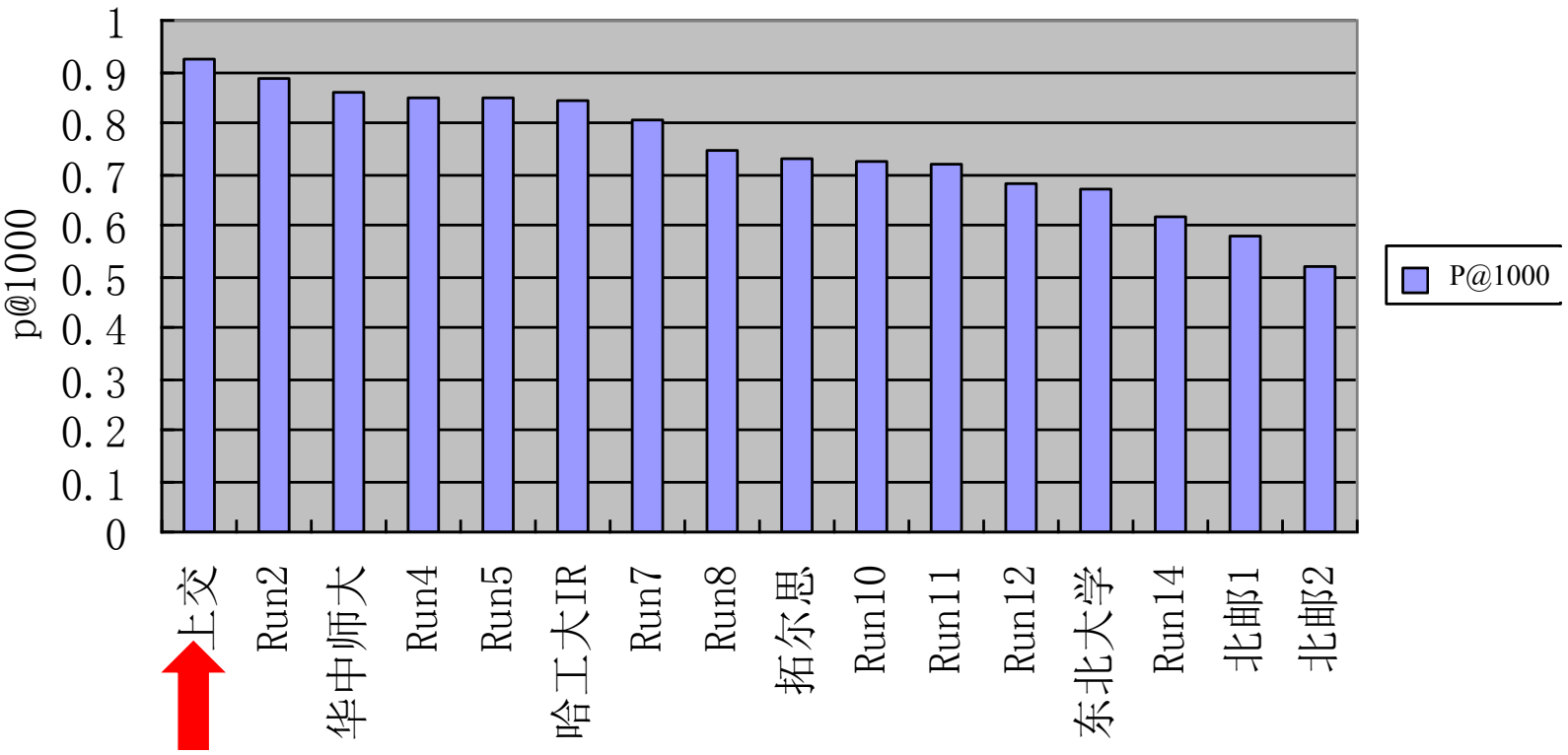
➢ COAE 2008: Test and Result

Analysis of Sentiment of Words and Sentences



Our result (Totally, 11 organizations including universities, institutes and companies participated in the evaluation)

Analysis of Sentiment of Words and Sentences



Our result (Totally, 11 organizations including universities, institutes and companies participated in the evaluation)

Analysis of Sentiment of Words and Sentences

- Sentiment Sentence Analysis (Yao and Li, 2009)
  - Use SVM, ME and CRF classification models to classify the sentiment of sentences.
  - Internal features: lexical features and POS features
  - External features: dependence features and context features
  - Feature template = Internal features + External features

Analysis of Sentiment of Words and Sentences

- The experimental results for automobile reviews, mobile phone reviews and mixed reviews:

		SVM	ME	CRF
Mixed Reviews	Precision	0.719	0.712	0.735
	Recall	0.74	0.738	0.789
	F-measure	0.729	0.725	**0.761**
Automobile Reviews	Precision	0.807	0.821	0.887
	Recall	0.816	0.832	0.868
	F-measure	0.811	0.826	**0.877**
Mobile Phone Reviews	Precision	0.699	0.714	0.723
	Recall	0.723	0.739	0.77
	F-measure	0.711	0.726	**0.745**

Analysis of Sentiment of Words and Sentences

- Obviously, the performance for sentiment sentence classification using CRF classification model is better than that using other two classification models.

Identification of Topic-Sentiment Relations

- Direct Relation (DR) / Indirect Relation (IR)
  - DR definition:
    - The 1st DR: A topic and the corresponding sentiment exist in a same sentence (or clause), and the distance between them is short.
    - The 2nd DR: A topic and the corresponding sentiment exist in contiguous clauses (i.e., not in same clause). The word between the clauses is conjunction (i.e., coordinative clauses)

Identification of Topic-Sentiment Relations

- IR definition:
  - A topic and the corresponding sentiment exist in different clauses, and there is no direct modified relation (i.e., it is possible that the relation between them is found through semantic analysis or derivation)
- According to our statistics on the annotation corpus for above DR and IR relations, they have the proportion of 96.3%.

Identification of Topic-Sentiment Relations

- DR Identification Approach (Chen, Liu, and Yao, 2009)
  - During the topic extraction, we choose the tf/idf algorithm according to the effects.
  - For the sentiment extraction, we use the sentiment-word dictionary to find the sentiment words.
  - For the relation extraction, we extracted the pairs of topics and sentiments as the candidate set, and then employed the closest distance matching model

Identification of Topic-Sentiment Relations

(CDMM) and SVM model to do that.

– CDMM model:

- ✓ Depending on the closest distance matching criterion, find the corresponding topic (words or phrase) with the shortest distance for each sentiment word.
- ✓ For the sentiment word whose topic is not found, it is matched by long-distance mode across punctuations until backtrack to the last

Identification of Topic-Sentiment Relations

topic. We suppose they are matched.

- ✓ If we do not find the corresponding sentiment word for any topic, give it up.

– SVM model:

- ✓ Lexical features: topic word; sentiment word; the left word of topic and sentiment word; the right word of topic and sentiment word.
- ✓ POS features: the left word of topic and sentiment word; the right word of topic and

Identification of Topic-Sentiment Relations

sentiment word.

- ✓ Semantic features: topic words, such as, product name, product attribute etc.
- ✓ Position features: the position of topic words and sentiment words; the number of words between topic words and sentiment words; the number of other topic words between topic words and sentiment words; the number of other sentiment words between topic

Identification of Topic-Sentiment Relations

words and sentiment words

- DR Experiment:
  - Corpus: 220 texts, each text contains 6-10 sentences, totally, 1500 sentences. They come from laptop, mobile phone, automobile and digital camera domains (each domain contains 55 texts). Among them, 200 texts are used for training corpus; 20 texts are used for test corpus.
  - In training corpus, we annotated 1200 positive

Identification of Topic-Sentiment Relations

relations and 1600 negative relations.

- The first experiment (Comparison for the different scales of annotated positive / negative relations in training sets): model1 - positive relations: 800; negative relations: 1600. model2 - positive relations: 1200; negative relations: 1600. The result shows that the performance from the model2 is better than one of the model1. It is important to balance the proportion for positive

Identification of Topic-Sentiment Relations

and negative training relations.

Text	Model1 Precision	Model1 Recall	Model1 F-measure	Model2 Precision	Model2 Recall	Model2 F-measure
digital camera 1	0.7778	0.3043	0.4375	0.8	0.5217	**0.6316**
digital camera 2	1	0.3333	0.5000	1	0.5556	**0.7143**
digital camera 3	1	0.5	0.6667	1	0.5	0.6667
digital camera 4	0.6667	0.4444	0.5333	0.6923	0.5	**0.5806**
digital camera 5	0.4286	0.25	0.3158	0.5	0.4167	**0.4546**

- The second experiment (Comparison for the performance of relation identification for different domains): use SVM and model2.

Identification of Topic-Sentiment Relations

Text	Precision	Recall	F-measure
digital camera 1	0.8	0.5217	0.6316
digital camera 2	1	0.5556	0.7143
digital camera 3	1	0.5	0.6667
digital camera 4	0.6923	0.5	0.5806
digital camera 5	0.5	0.4167	0.4546
automobile 1	1	0.4	0.5714
automobile 2	0.8333	0.7143	0.7692
automobile 3	0.8889	0.8	0.8421
automobile 4	1	0.4	0.5714
automobile 5	0.8571	0.6667	0.7500
laptop 1	1	0.75	0.8571
laptop 2	1	0.5	0.6667
laptop 3	1	0.5	0.6667
laptop 4	0.7778	0.6364	0.7000
laptop 5	1	0.6667	0.8000
mobile phone 1	0.5	0.3333	0.4000
mobile phone 2	0.75	0.5	0.6000
mobile phone 3	1	0.5	0.6667
mobile phone 4	1	0.5	0.6667
mobile phone 5	0.75	0.5	0.6000

Identification of Topic-Sentiment Relations

The result shows that the performance for relation identification from different domain is different. Among them, the performance for relation identification from automobile and laptop is better than other two domains. The possible reason is the number of technical terms is different.

– The third experiment (Comparison for the different classification model): compare the

Identification of Topic-Sentiment Relations

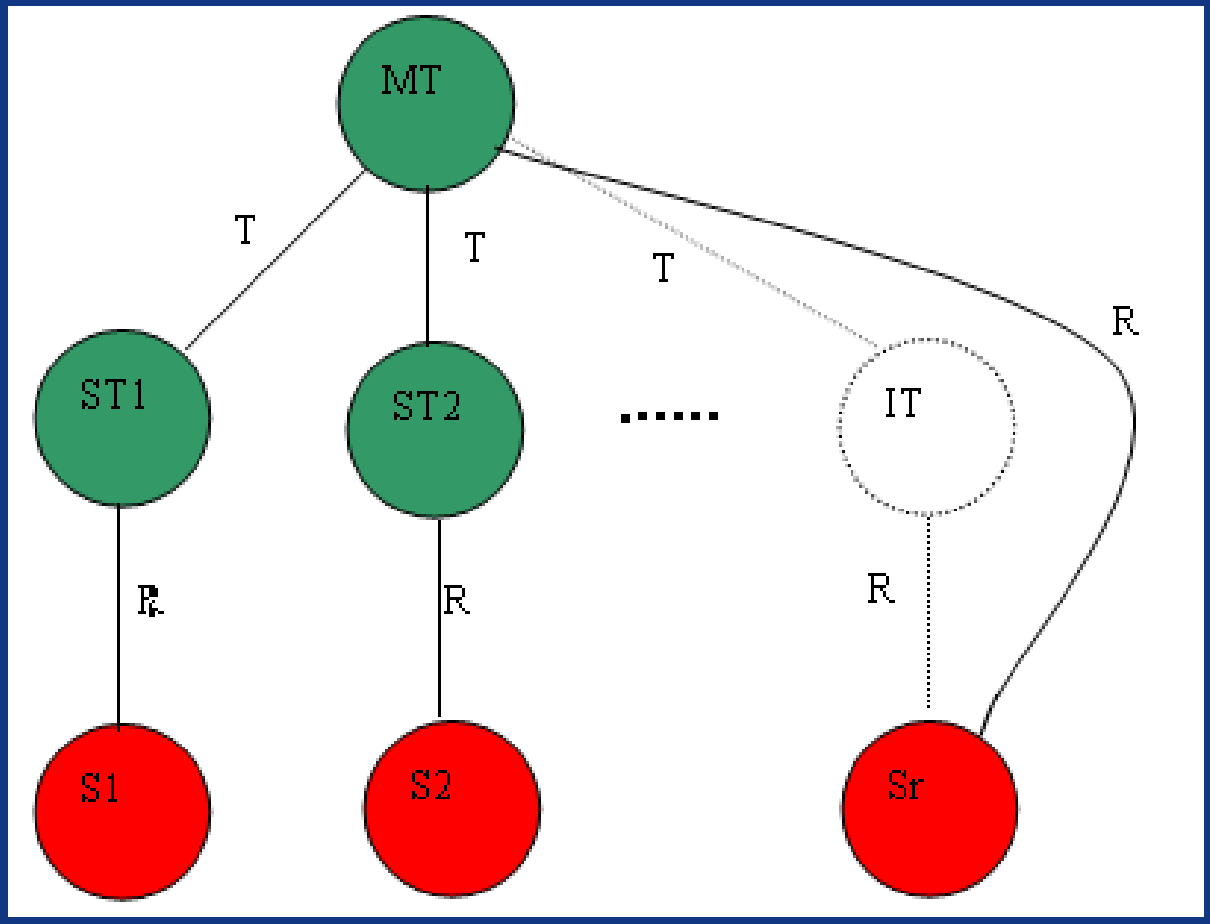
CDMM model and SVM model in digital camera domain. The result shows that the performance of SVM model is obviously superior to one of CDMM model.

Text	CDMM Precision	CDMM Recall	CDMM F-measure	SVM Precision	SVM Recall	SVM F-measure
digital camera 1	0.4063	0.5652	0.4728	0.8	0.5217	**0.6316**
digital camera 2	0.5455	0.6667	0.6000	1	0.5556	**0.7143**
digital camera 3	0.2222	0.3333	0.2666	1	0.5	**0.6667**
digital camera 4	0.25	0.3889	0.3044	0.6923	0.5	**0.5806**
digital camera 5	0.3182	0.5833	0.4118	0.5	0.4167	**0.4546**

Identification of Topic-Sentiment Relations

- IR Identification (Chen and Yao, 2010)
  - An example: 这部车价格有些贵，发动机很酷，总体来说不错，挺漂亮的。(The car has a high price, a cool engine, a good entirety and looks beautiful.)
  - In this example, for the sentiment word beautiful, there is no explicit topic. We say there is an implicit topic for the sentiment word beautiful. How to find the implicit topic?

Identification of Topic-Sentiment Relations



Identification of Topic-Sentiment Relations

- We utilize an ontology to find a <Concept, Sub-Concept> relation. In this example, the concept is car, so one of the sub-concepts is 外形/外观/外表...(apperance). Because it satisfies the collocation between the word beautiful and the word apperance, so apperance is one of the implicit topics for beautiful. Thus, we can transform a indirect relation between topic and sentiment into direct relation.
- The algorithm is described as follows:

Identification of Topic-Sentiment Relations

Input: sentence S, ontology dictionary O, collocation database C

Output: opinion-element relation set R

Step1: R1 = DependencyParsing(S); R.append(R1)

Step2: R2 = CheckParallel(S); R.append(R2)

Step3: SentimentRemain = IsRemaining(S, R1, R2)

    If Empty (SentimentRemain): **return** R

Step4: Foreach sentiment s in SentimentRemain:

        CandidateImplicitTopic = Lookup(s, C)

        If Empty (CandidateImplicitTopic): **continue**

        Else: Foreach topic t in CandidateImplictTopic:

            Upperleveltopic = **NULL**

            flag = IsPathExistsInOntology(t, upperleveltopic, O)

            If flag is true:

                R.Add(<upperlevel, s>)

                C.AddPotential(<upperlevel, s >

Step5: **return** R

Identification of Topic-Sentiment Relations

- IR Experiment
  - Corpus: we labeled 1000 sentences on the car domains, all the sentences are from Internet product reviews, among which there are 1447 direct relations and 232 indirect relations. In order to prove the generality of our approach, we also make use the COAE2008 Task3's corpus which includes four domains. After crossing checked and verified, about 6994 correct and

Identification of Topic-Sentiment Relations

- disambiguous opinion-element relations. In addition, we choose Google and Baidu as main search engines to collect 7600 collocation pairs.
  - We define baseline1 and baseline2 by using closest-pair and dependency parsing respectively, we also add the Maximum Entropy experiment by J. Zhang for comparison.

Identification of Topic-Sentiment Relations

Method	Precision	Recall	F-measure
Baseline1(Closest -pair)	0.516	0.7385	0.6072
Baseline2(Parsing)	0.7253	0.8599	0.7863
Baseline3(ME)	0.811	0.7534	0.7809
Our method	0.7392	0.9337	**0.8247**

Text Domain	Precision	Recall	F-measure
Automobile	0.722	0.8875	**0.796**
Laptop	0.7024	0.8232	0.7576
Digital Camera	0.7387	0.834	0.783
Mobile Phone	0.7231	0.8522	0.7823
Average	0.7212	0.849	0.7795

Identification of Topic-Sentiment Relations

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Conclusion

- Opinion Mining for Chinese subjective texts is a novel and challenging research direction.
- The concept of Chinese opinioned subjective texts is defined and their text category architecture and sentence type are illustrated.
- A fine-grained annotation approach is proposed and a corresponding corpus for the research is developed.
- Automatic classification and Network Informal Language in the preprocessing are investigated

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Conclusion

- using an approach combining statistical, machine learning and rule extraction.
- A topic identification approach for Chinese opinioned sentences is proposed, which can be used for domain-specific and domain-independent corpus.
- The analysis of the sentiments for Chinese opinioned sentences is studied using the information of sentiment words and statistical probability model.

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Conclusion

- The identification of direct and indirect relations between topics and sentiments for Chinese opinioned sentences is researched using an approach combining syntactical analysis and linguistic rule.
- A demo system of fine-grained opinion mining is designed and implemented.
- The part of the outcomes are evaluated under a canonical mode in COAE 2008, which obtained advantageous evaluation results.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Thank you for your attention!