



A Political News Corpus in Chinese for Opinion Analysis

Benjamin K. Tsou Bin Lu

**Language Information Sciences Research Centre,
City University of Hong Kong**



Introduction

- Opinion analysis
 - Opinions incorporated in factual news reports represent a common phenomenon
- Expression-level corpus
 - MPQA corpus of 10,000 sentences with words and phrases annotated in context (Wiebe et al.).
- Sentence-level corpus
 - Opinion analysis corpus used at NTCIR-6 and NTCIR-7 (Chinese, Japanese and English).
- Document-level corpus (un-annotated)
 - Movie reviews (Pang et al.)



Introduction (cont'd)

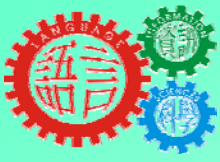
- A novel annotation scheme: three levels
 - 1) Expression, 2) sentence, 3) document
- A Chinese election news corpus
 - Using proposed annotation scheme.
 - Elections:

2004 US presidential election
2007 HK chief executive election
2008 US presidential election
- Agreement study shows
 - good consistency among different annotators on the three levels.



Annotation scheme

- Expression level annotation
 - Salient Polar **Word** (*Word*)
 - Salient Polar **Chunk / Phrase** (*Chunk*)
- Sentence level annotation
 - Salient opinionated **sentences**
- Document level annotation
 - Focus person
 - Focus event



Expression level annotation

- Identify and annotate opinion-bearing words and chunks (or phrases) in context.
- Word (Salient Polar Word)
 - an inherently positive or negative word
- Chunk (Salient Polar Chunk)
 - a polar expression more than a word
 - three types
 - Collocations
 - 陳先生豎起拇指大贊曾蔭權 (Mr. Chen gave thumbs up to and praised Donald Tsang)
 - Context-dependent expression
 - 有經驗 (experienced), 好/壞的經驗 (good/bad experience)
 - Polar words with contextual valence shifter
 - 很成功 (very successful)



Expression level annotation (cont'd)

- Annotate salient opinion expressions using a common frame (similar to that of NTCIR-6/7), including
 - *expression itself*
 - *polarity*
 - *intensity of the polarity*
 - *opinion holder*
 - *opinion target*



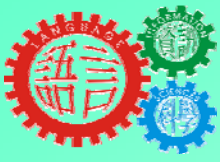
Sentence level annotation

- Identify salient opinionated sentences, and annotate them with the following features:
 - *opinion holder*
 - *opinion target*
 - *polarity*
 - *intensity of the polarity*



Document level annotation

- Identify and annotate **focus person(s)** and **focus event(s)** in news reports with *polarity* and *intensity of the polarity*.
- **Focus person**
 - the candidate(s) or highly related person(s) in the given elections
 - 2008 US presidential election
 - Barack Obama, John McCain, Joe Biden, Sarah Palin, George W. Bush, Hillary Clinton, etc.
 - 2004 US presidential election
 - Bush, Kerry, etc.
- **Focus event**
 - major event(s) discussed in news reports
 - E.g. the first presidential debate between two candidates.



Data source 1

- LIVAC synchronous corpus (<http://www.livac.org>)
 - News related to the three elections
- More than 10 annotators

Election title	#doc	#sentence
2004 US presidential election	~600	~12K
2007 HK chief executive election	~1,000	~18K
2008 US presidential election	~200	~3K
Total	~1.8K	~33K



Data source 2

- Other political personalities
 - Deng Xiaoping
 - Tung Chee Hwa
 - Koizumi Junichiro
 - Chen Shui-bian
 - etc.

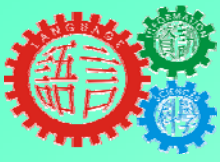


Agreement study

- **Annotators:** A & J & S
- **Data:** 56 documents (956 sentences)
- **Metrics:** Kappa & Agr (Wiebe et al. 2005)

$$agr(a||b) = \frac{|A \text{ matching } B|}{|A|}$$

- **Agreement on THREE levels**
 - Expression, sentence & document



Agreement on the **EXPRESSION** level

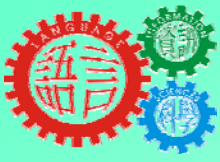
Word	agr(a b)	agr(b a)	Average
A & J	0.87	0.47	
A & S	0.78	0.52	
J & S	0.69	0.86	
			0.70

Chunk	agr(a b)	agr(b a)	Average
A & J	0.53	0.17	
A & S	0.50	0.18	
J & S	0.54	0.58	
			0.42

- **Wiebe et al.'s MPQA corpus (LRE 2005)**
- **Annotators:** A & M & S
- **Data:** 13 documents with a total of 210 sentences

Table III. Inter-annotator agreement: expressive subjective elements

<i>a</i>	<i>b</i>	<i>agr(a b)</i>	<i>agr(b a)</i>	Average
A	M	0.76	0.72	
A	S	0.68	0.81	
M	S	0.59	0.74	
				0.72



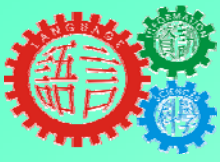
Agreement on the **SENTENCE** level

- Salient opinionated sentence recognition

	Kappa	Agree
A & J	0.50	0.82
A & S	0.56	0.95
J & S	0.81	0.84
Average	0.62	0.87

Wiebe's MPQA Corpus

	All Sentences		Borderline Removed		
	κ	agree	κ	agree	% removed
A & M	0.75	0.89	0.87	0.95	11
A & S	0.84	0.94	0.92	0.97	8
M & S	0.72	0.88	0.83	0.93	13



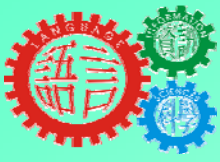
Agreement on the **SENTENCE** level

- Salient opinionated sentence recognition

	Kappa	Agree
A & J	0.50	0.82
A & S	0.56	0.95
J & S	0.81	0.84
Average	0.62	0.87

The NTCIR-6 opinion corpus
Kappa Summary

Language	Minimum	Maximum	Average
CH Opinionated	0.0537	0.4065	0.2328
JA Opinionated	0.5997	0.7681	0.6740
EN Opinionated	0.1704	0.4806	0.2947



Agreement on the **DOCUMENT** level

a) Focus Person

focus person	Agr(a b)	agr(b a)	Average
A & J	0.76	0.85	
A & S	0.70	0.82	
J & S	0.88	0.92	
			0.82

b) Focus Event

focus event	Agr(a b)	agr(b a)	Average
A & J	0.61	0.61	
A & S	0.55	0.55	
J & S	0.75	0.75	
			0.64



Future enhancement: Shallow parsing, etc.

- Bush dislikes democrats.
- Democrats dislikes Bush.



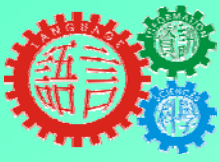
Conclusion remarks

- A novel annotation scheme: three levels
 - 1) Expression, 2) sentence, 3) document
- An annotated election news corpus
 - Using the proposed annotation scheme.
- The agreement study shows
 - Good consistency among different annotators on three levels.



Future work

- To enhance multi-level and fine-grained annotation of this corpus for NLP applications.
- To investigate how the corpus could be used in the evaluation of Chinese opinion analysis.
- To make it public to research community in future.



References

- Pang B., Lee L., and Vaithyanathan S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*, pp.79–86.
- Seki Y., Evans D.K., Ku L.W., Chen H.H., Kando N., and Lin C.-Y. 2007. Overview of opinion analysis pilot task at NTCIR-6. *Proc. of the Sixth NTCIR Workshop*. May 2007, Japan.
- Tsou B.K.Y., Tsoi W.F., Lai T.B.Y., Hu J., and Chan S.W.K. 2000. LIVAC, A Chinese Synchronous Corpus, and Some Applications. *Proceedings of the ICCLC International Conference on Chinese Language Computing*, Chicago. pp. 233–238.
- Tsou B.K.Y., Yuen W.M.R., Kwong O.Y., Lai T.B.Y., Wong W.L. 2005. Polarity classification of celebrity coverage in the Chinese press. In *Proceeding of the 2005 International Conference on Intelligence Analysis*. Virginia, USA.
- Wiebe J., Wilson T., Cardie C. 2005. Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.



Thanks!

Q & A