NTCIR-6 Opinion Task Overview

Hsin-Hsi Chen, David Kirk Evans, Yohei Seki

Opinion Analysis

- Given a sentence:
  - Does it express an opinion?
    - Polarity? (Positive, Negative, Neutral)
    - Who expresses the opinion?
  - Is it relevant to the document set topic?

Opinion Analysis

- Aomori Gov. Morio Kimura on Tuesday banned a ship from carrying highly radioactive waste into a port here, voicing concern that Tokyo may try to turn this tiny fishing village into a permanent nuclear dumping site.
- 2/3 Agree: opinionated
- 2/3 Agree: relevant to topic

  Give information regarding protests against nuclear power.

Opinion Analysis

- Aomori Gov. Morio Kimura on Tuesday banned a ship from carrying highly radioactive waste into a port here, voicing concern that Tokyo may try to turn this tiny fishing village into a permanent nuclear dumping site.
- 2/3 Agree: opinionated
- 2/3 Agree: relevant to topic

  Give information regarding protests against nuclear power.

Corpus Annotation

- Three annotators per document
- ~ 20 docs per topic (EN, JA), 40 CH
- 1998~2001 data
- CH annotators students, JA news-related, EN translators & teachers

Feature	Value	Req'd?
Opinionated	YES, NO	Yes
Opinion Holder	String, multiple per sentence possible	Yes
Relevant	YES, NO	No
Polarity	Positive, Neutral, Negative	No

Corpus Sources

- Japanese: 1998-2001 Yomiuri, Mainichi newspapers
- Chinese: 1998-2001 United Daily News, China Times, China Times Express, Commercial Times, Central and Daily News
- English: 1998-2001 Mainichi Daily News, Korea Times, Xinghua

Participation

- CH - 5 groups
- EN - 6 groups
- JA - 3 groups

Annotator Training

- JA: 1 topic for training, basic instructions for opinionated / relevant / polarity. 2 adjudication meetings. Checked 2 topics for disagreement, revised answers for higher agreement
- EN: 1 topic for training, 1 adjudication meeting, same guidelines as JA
- CH: 1 hour meeting with annotators, explained special cases, could ask questions for confusing cases, but did not dictate "answer"

Some Guidelines

- General beliefs, "common sense knowledge" are not opinions
- Expressions of future plans are not opinions
- JA: Rules about how to write opinion holders (title, position, affiliation, etc.)

Training Data

- 4 sample topics for Chinese and Japanese
- 1 sample topic for English
  - Reference to MPQA opinion corpus

Annotator Agreement

Lang	Min	Max	Avg.
CH	.0537	.4065	.2328
EN	.1673	.4799	.2943
JA	.5997	.7681	.6740

Cohen's Kappa

Annotator Agreement

- EN, JA have consistent annotators
- CH uses 3 annotators from pool of 7 (per-topic agreement)
- JA high agreement

Lang	Pair	Task	Kappa
E	1-2	Opinionated	0.4799
E	1-3	Opinionated	0.1673
E	2-3	Opinionated	0.2357
E	1-2	Relevant	0.2666
E	1-3	Relevant	0.4763
E	2-3	Relevant	0.4143
E	1-2	Polarity	0.4298
E	1-3	Polarity	0.1710
E	2-3	Polarity	0.2247
J	1-2	Opinionated	0.6499
J	1-3	Opinionated	0.6107
J	2-3	Opinionated	0.7919
J	1-2	Relevant	0.4130
J	1-3	Relevant	0.3676
J	2-3	Relevant	0.8576
J	1-2	Polarity	0.5736
J	1-3	Polarity	0.5341
J	2-3	Polarity	0.7734

Annotator Agreement

- EN, JA have consistent annotators
- CH uses 3 annotators from pool of 7 (per-topic agreement)
- JA high agreement
- EN #3 difficult!

Lang	Pair	Task	Kappa
E	1-2	Opinionated	0.4799
E	1-3	Opinionated	0.1673
E	2-3	Opinionated	0.2357
E	1-2	Relevant	0.2666
E	1-3	Relevant	0.4763
E	2-3	Relevant	0.4143
E	1-2	Polarity	0.4298
E	1-3	Polarity	0.1710
E	2-3	Polarity	0.2247
J	1-2	Opinionated	0.6499
J	1-3	Opinionated	0.6107
J	2-3	Opinionated	0.7919
J	1-2	Relevant	0.4130
J	1-3	Relevant	0.3676
J	2-3	Relevant	0.8576
J	1-2	Polarity	0.5736
J	1-3	Polarity	0.5341
J	2-3	Polarity	0.7734

Average Chinese Agreement Per-topic

0.400	0.377	0.232	0.221
0.406	0.104	0.235	0.070
0.225	0.122	0.236	0.235
0.194	0.271	0.134	0.311
0.160	0.274	0.316	0.093
0.261	0.195	0.366	0.142
0.363	0.269	0.366	0.128
0.320	0.054	0.200	0.152

0.054 ~ .406

Average English Agreement Per-topic

0.390	0.207	0.319	0.155
0.315	0.413	0.321	0.323
0.438	0.438	0.457	0.142
0.264	0.284	0.192	0.225
0.233	0.446	0.202	0.171
0.245	0.395	0.218	0.154
0.222	0.219	0.401	0.094

0.094 ~ 0.438

Annotator 1-2 English Agreement Per-topic

0.5174	0.4608	0.3574	0.4058
0.4374	0.7079	0.5235	0.5205
0.4565	0.4308	0.6163	0.4274
0.5505	0.4771	0.3903	0.5620
0.3630	0.5884	0.4664	0.3673
0.4287	0.5101	0.4427	0.3760
0.4803	0.5532	0.5075	0.2825

0.2825 ~ 0.7079

Average Japanese Agreement Per-topic

0.4547	0.6199	0.5871	0.9061
0.7607	0.4935	0.5512	0.5766
0.4561	0.6508	0.7442	0.7121
0.6413	0.4640	0.7922	0.6046
0.7163	0.7462	0.6728	0.7035
0.7764	0.7489	0.7176	0.7046
0.3795			

0.3795 ~ .9061

Corpus

Lang	Topics	Docs	Sents	Opin.	Rel.
CH	28	843	8,546	62% / 25%	39% / 16%
EN	28	439	8,417	23% / 5%	27% / 11%
JA	30	490	15,279	29% / 22%	64% / 49%

Lenient / Strict

Evaluation Metrics

- Precision, Recall, F-Measure over opinionated, relevant, polarity
- Semi-automatic evaluation of opinion holders (precision, recall, f-measure)
- Multiple approaches developed

Polarity Differences (Strict System POS)

Annotation				Behavior
POS	NEU	NEG	NOT	
3	0	0	0	LWK+, DKE+, YS+
2	0	1	0	LWK skip, DKE-, YS-
0	0	0	3	LWK-, DKE-, YS-
0	0	1	2	LWK skip, DKE-, YS-

Polarity Differences (Lenient System POS)

Annotation				Behavior
POS	NEU	NEG	NOT	
3	0	0	0	LWK+, DKE+, YS+
2	0	1	0	LWK +, DKE+⅔, YS+
0	0	0	3	LWK-, DKE-, YS-
0	0	1	2	LWK-, DKE-, YS-
1	0	2	0	LWK-, DKE+⅓, YS-
1	1	0	1	LWK+, DKE+⅓, YS prec. down, recall no change

Evaluation Differences

- JA DKE-YS Lenient: .014 / Strict: .002
- JA LWK-YS Lenient: .108 / Strict: .110
- EN DKE-YS Lenient: .578 / Strict: .677
- EN DKE-LWK Lenient: 1.08 / Strict: .607
- CH DKE-LWK Lenient: .204 / Strict: .047
- CH DKE-YS Lenient: .174 / Strict: .126
- CH LWK-YS Lenient: .03 / Strict: .079

Chinese Rank Differences

DKE	LWK	YS
UMCP-1	UMCP-1	UMCP-1
UMCP-2	UMCP-2	UMCP-2
NTU	Gate-1	Gate-1
Gate-1	NTU	NTU
Gate-2	Gate-2	Gate-2
CUHK	CUHK	CUHK
ISCAS	ISCAS	ISCAS

Lenient Opinionated F-measure

Chinese Rank Differences

DKE	LWK	YS
GATE-2	GATE-2	GATE-2
CUHK	CUHK	CUHK
NTU	NTU	GATE-1
GATE-1	GATE-1	UMCP-1
UMCP-1	UMCP-1	NTU
UMCP-2	UMCP-2	UMCP-2
ISCAS	ISCAS	ISCAS

Strict Opinionated F-measure

EN Rank Differences

DKE	LWK	YS
GATE-1	GATE-1	GATE-2
GATE-2	GATE-2	GATE-1
ICU-IR	ICU-IR	ICU-IR
Cornell	NII	NII
NII	IIT-1	IIT-1
IIT-1	Cornell	TUT-1
TUT-1	TUT-1	TUT-2
TUT-2	TUT-2	IIT-2
IIT-2	IIT-2	Cornell

Lenient Opinionated F-measure

EN Rank Differences

DKE	LWK	YS
ICU-IR	ICU-IR	ICU-IR
NII	NII	NII
GATE-1	GATE-1	GATE-1
GATE-2	GATE-2	GATE-2
Cornell	IIT-1	IIT-1
IIT-1	Cornell	Cornell
TUT-1	TUT-1	TUT-1
TUT-2	TUT-2	TUT-2
IIT-2	IIT-2	IIT-2

Strict Opinionated F-measure

JA Rank Differences

DKE	LWK	YS
TUT-1	TUT-1	TUT-1
EHBN-1	EHBN-1	EHBN-1
EHBN-2	EHBN-2	EHBN-2
NICT-1	NICT-1	NICT-1
NICT-2	NICT-2	NICT-2

Lenient or Strict Opinionated F-measure

Results DKE Lenient

	Group	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
Chinese	CUHK	.819	.520	.636	.797	.828	.813	.480	.431	.454
	NTU	.667	.888	.762	.636	1.0	.777	.286	.538	.374
	UMCP	.645	.976	.777	.683	.519	.590	.256	.549	.349
	ISCAS	.590	.664	.625	-	-	-	.170	.271	.209
	GATE	.747	.591	.660	-	-	-	-	-	-
	IIT	.325	.588	.419	-	-	-	.120	.287	.169
Japanese	TUT	.310	.575	.403	.392	.597	.473	.088	.215	.125
English	Cornell	.317	.651	.427	-	-	-	.073	.197	.107
	NII	.325	.624	.427	.510	.322	.395	.077	.194	.110
	GATE	.324	.905	.477	.286	.632	.393	-	-	-
	ICU-KR	.396	.524	.451	.409	.263	.320	.151	.264	.192
	EHBN	.531	.453	.488	-	-	-	-	-	-
	NICT	.671	.315	.429	.598	.669	.632	.298	.149	.199
	TUT	.552	.609	.589	.630	.645	.638	.274	.322	.296

Results DKE Lenient

	Group	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
Chinese	Average	P 0.6936 R 0.7278 F 0.6920			P 0.7053 R 0.7823 F 0.7266			P 0.2980 R 0.4473 F 0.3465		
English	Average	P 0.3328 R 0.6445 F 0.4340			P 0.3993 R 0.4535 F 0.3953			P 0.1018 R 0.2314 F 0.1406		
Japanese	Average	P 0.5846 R 0.4690 F 0.5020			P 0.6140 R 0.6570 F 0.6350			P 0.2860 R 0.2355 F 0.2475		

Results DKE Strict

	Group	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
Chinese	CUHK	.340	.575	.428	.468	.901	.616	.197	.595	.296
	NTU	.265	.922	.412	.342	1.0	.509	.108	.666	.186
	UMCP	.245	.988	.393	.404	.570	.473	.086	.615	.150
	ISCAS	.221	.662	.331	-	-	-	.059	.314	.099
	GATE	.253	.979	.402	-	-	-	-	-	-
	IIT	.070	.578	.125	-	-	-	.027	.322	.049
	TUT	.065	.553	.117	.171	.605	.266	.016	.195	.029
English	Cornell	.069	.662	.125	-	-	-	.010	.135	.018
	NII	.073	.642	.131	.242	.355	.287	.014	.185	.027
	GATE	.070	.940	.130	.112	.579	.188	-	-	-
	ICU-KR	.102	.616	.175	.177	.266	.213	.034	.301	.061
Japanese	EHBN	.414	.479	.444	-	-	-	-	-	-
	NICT	.545	.348	.425	.470	.693	.560	.168	.150	.158
	TUT	.414	.620	.497	.505	.681	.580	.161	.339	.218

Results DKE Strict

	Group	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
Chinese	Average	P 0.2648 R 0.8252 F 0.3932			P 0.4047 R 0.8237 F 0.5327			P 0.1125 R 0.5474 F 0.1828		
English	Average	P 0.0748 R 0.6652 F 0.1338			P 0.1755 R 0.4513 F 0.2385			P 0.0202 R 0.2276 F 0.0368		
Japanese	Average	P 0.4577 R 0.4823 F 0.4553			P 0.4875 R 0.6870 F 0.5700			P 0.1645 R 0.2445 F 0.1880		

Holder evaluation

- Semi-automatic evaluation
- Match system extracted holders to annotator holder list, automate the process in some way
- Time consuming, only first priority run evaluated

Opinion Holder Results

	Group	Lenient			Strict		
		P	R	F	P	R	F
Chinese	CUHK	0.647	0.754	0.697	0.707	0.785	0.744
	NTU	0.652	0.172	0.272	0.661	0.177	0.279
	UMCP	0.241	0.410	0.303	0.293	0.438	0.351
	ISCAS	0.458	0.405	0.430	0.470	0.406	0.436
	GATE	0.427	0.154	0.227	0.419	0.156	0.227
	IIT	0.198	0.409	0.266	0.054	0.461	0.097
	TUT	0.117	0.218	0.153	0.029	0.241	0.051
English	Cornell	0.163	0.346	0.222	0.041	0.392	0.074
	NII	0.066	0.166	0.094	0.018	0.169	0.032
	GATE	0.121	0.349	0.180	0.029	0.398	0.055
Japanese	ICU-KR	0.303	0.404	0.346	0.085	0.515	0.146
	EHBN	0.138	0.085	0.105	0.079	0.094	0.086
	NICT	0.238	0.102	0.143	0.133	0.110	0.120
	TUT	0.226	0.224	0.225	0.131	0.251	0.172

Opinion Holder Results

	Group	Lenient			Strict		
		P	R	F	P	R	F
Chinese	CUHK	0.647	0.754	0.697	0.707	0.785	0.744
	NTU	0.652	0.172	0.272	0.661	0.177	0.279
	UMCP	0.241	0.410	0.303	0.293	0.438	0.351
	ISCAS	0.458	0.405	0.430	0.470	0.406	0.436
	GATE	0.427	0.154	0.227	0.419	0.156	0.227
	IIT	0.198	0.409	0.266	0.054	0.461	0.097
English	TUT	0.117	0.218	0.153	0.029	0.241	0.051
	Cornell	0.163	0.346	0.222	0.041	0.392	0.074
	NII	0.066	0.166	0.094	0.018	0.169	0.032
	GATE	0.121	0.349	0.180	0.029	0.398	0.055
	ICU-KR	0.303	0.404	0.346	0.085	0.515	0.146
Japanese	EHBN	0.138	0.085	0.105	0.079	0.094	0.086
	NICT	0.238	0.102	0.143	0.133	0.110	0.120
	TUT	0.226	0.224	0.225	0.131	0.251	0.172

Opinion Holder Results

	Group	Lenient			Strict		
		P	R	F	P	R	F
Chinese	CUHK	P 0.485 R 0.379 F 0.386			P 0.510 R 0.392 F 0.407		
	NTU						
	UMCP						
	ISCAS						
	GATE						
English	IIT	P 0.161 R 0.315 F 0.210			P 0.043 R 0.363 F 0.076		
	TUT						
	Cornell						
	NII						
	GATE						
	ICU-KR						
Japanese	EHBN	P 0.201 R 0.137 F 0.158			P 0.114 R 0.152 F 0.126		
	NICT						
	TUT						

Discussion

- Relevance < Opinionated < Polarity < Holder
- CH, EN, JA corpora have different annotator agreement: training issue or data issue?
- How to evaluate when annotators do not agree?
- What do the results MEAN?

Difficulties

- Annotator agreement (address with better experiment design?)
- Imprecise task formulation
- Meaningful Cross-language comparison?
- Data format inconsistencies
- Lack of shared tools for evaluation / processing / etc.

Future Work

- Increase group participation in multiple languages (only TUT, GATE this year)
- What is upper bound on annotator performance
- Can we compare across languages?
- Applications!

NTCIR-7

- Multilingual Opinion Analysis Task (MOAT) or Multilingual Evaluation of Opinions on the Web (MEOW)?
  - Opinionated sentence detection
  - Polarity (POS, NEG, NEU)
  - Opinion holder
  - Opinion target [optional]
  - Topic relevance [optional]