Multi-lingual Opinion Analysis Applied to World News: A Case Study

David Kirk Evans, Noriko Kando
{devans,noriko.kando}@nii.ac.jp
National Institute of Informatics

Outline

- NTCIR Opinion Analysis Task
  - Corpus Introduction
    - Top Opinion Holders, Interesting words
  - Summarization corpus
    - Participant graphs

NTCIR Opinion Analysis Task

- Chinese, English, Japanese news documents over 30 topics
- Annotated by three annotators at sentence level
- Opinionated, opinion holder, polarity, relevant to topic

Corpus Annotation

- Three annotators per document
- ~ 20 docs per topic (EN, JA), 40 CH
- 1998~1999 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

Relatively little instruction given on the task, JA with most focus on annotator consistency

Corpus

Lang	Topics	Docs	Sents	Opin.	Rel.
CH	32	843	8,546	62% / 25%	39% / 16%
EN	28	439	12,525	30% / 7%	69% / 37%
JA	30	490	8,523	29% / 22%	64% / 49%

Lenient / Strict

This table differs from the one in the paper: the Chinese figures erroneously count all annotated instances – since there are three annotators per document, there are approximately 3x more reported documents and sentences. Japanese numbers included 4 sets released as training material.

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	0.4806
E	1-3	Opinionated	0.1704
E	2-3	Opinionated	0.2332
E	1-2	Relevant	0.5240
E	1-3	Relevant	0.0618
E	2-3	Relevant	0.5298
E	1-2	Polarity	0.5457
E	1-3	Polarity	0.2039
E	2-3	Polarity	0.2645
J	1-2	Opinionated	0.6541
J	1-3	Opinionated	0.5997
J	2-3	Opinionated	0.7681
J	1-2	Relevant	0.7176
J	1-3	Relevant	0.6966
J	2-3	Relevant	0.8394
J	1-2	Polarity	0.6919
J	1-3	Polarity	0.6367
J	2-3	Polarity	0.7875

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	0.4806
E	1-3	Opinionated	0.1704
E	2-3	Opinionated	0.2332
E	1-2	Relevant	0.5240
E	1-3	Relevant	0.0618
E	2-3	Relevant	0.5298
E	1-2	Polarity	0.5457
E	1-3	Polarity	0.2039
E	2-3	Polarity	0.2645
J	1-2	Opinionated	0.6541
J	1-3	Opinionated	0.5997
J	2-3	Opinionated	0.7681
J	1-2	Relevant	0.7176
J	1-3	Relevant	0.6966
J	2-3	Relevant	0.8394
J	1-2	Polarity	0.6919
J	1-3	Polarity	0.6367
J	2-3	Polarity	0.7875

Does content differ across languages?

- Who is expressing opinions?
- Are they positive or negative?
- What are important concepts expressed in opinions?
- Does the above differ by language?

Topic Examination

- Topic 010: "History Textbook Controversies, World War II"
- Examine polarity
- List top opinion holders, polarity
- Use mutual information / log likelihood measures to identify opinionated words

Topic 010 Information

	Docs	Sents	POS	NEU	NEG	Rel.
CH	41	1,641 (547)	198 (12%)	199 (12%)	528 (32%)	966 (59%)
EN	20	774 (258)	8 (1.0%)	57 (7.3%)	224 (28.9%)	359 (46.4%)
JA	20	2,358 (786)	149 (6.3%)	148 (6.3%)	319 (13.5%)	1269 (53.8%)

Annotated sentences and tags (not lenient or strict standard)

Opinion Holders

60	Author	1,32,17
21	S. Korea	0,11,10
20	Zhu Bangzao	0,18,2
14	History book group	2,5,7
7	S. Korean legislators	0,6,1
4	Jp. Ministry of Education	0,2,2

English

41	He	20,7, 14
24	Koizumi Junichiro	11,4,9
13	Ministry of Foreign Affairs	8,1,4
11	Hsieh, Chi-ta	1,7,3
9	Chu, Te-Lan	1,7,1
8	Lo, Fu-chen	4,0,4

Chinese

127	Author	28,42, 54
37	Korea	1,19,18
30	Hata Ikuhiko	11,4,15
26	Takamori Akinori	13,8,5
23	Tanaka Toshiaki	3,16,4
11	China	1,4,6

Japanese

Opinionated Terms

$$MI_{w,A} = log_2 \left(\frac{|A_w|}{|A|} \times \frac{|A| + |B|}{|A_w| + |B_w|} \right)$$

- Mutual Information
- Log Likelihood

$$G^2 = 2(alog(a) + blog(b) + clog(c) + dlog(d) \\ - (a+b)log(a+b) - (a+c)log(a+c) \\ - (b+d)log(b+d) - (c+d)log(c+d) \\ + (a+b+c+d)log(a+b+c+d))$$

A. Kilgarriff. Comparing corpora. International Journal of Corpus Linguistics, 6(1):97–133(37), 2001.
T. Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74, 1994.
K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics, pages 76–83, Vancouver, B.C., 1989. Association for Computational Linguistics.

Opinionated Terms

Log Likelihood	Mutual Information
textbook	invaders
history	denigration
Japanese	blurs
textbooks	biased
facts	Stage
Japan	Rally
Asian	Netizens
draft	Cyber
descriptions	militarists
distorted	tragedies

Log Likelihood	Mutual Information
教科書 textbooks	忠実 faithful
歴史 history	不見識 rash
検定 offic. approval	おかしい strange
修正 revision	いこ　(unfair?)
ない (negation)	郁 culture progress
韓国 Korea	許容 permission
1	欺瞞 deception
記述 description	東京都立大
つくる	山住 Yamazumi Pn
美化 glorification	危惧 misgivings

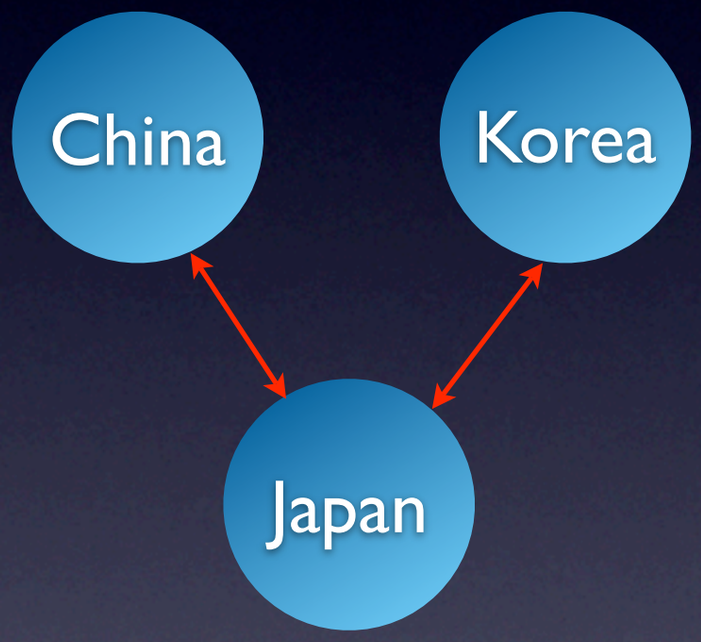
Cross-language Summarization

- Simple approach to characterize documents: term counts, entity counts, interesting words (M.I., L.L.)
- Easily apply to multiple languages
- Analysis results translate more easily
- Opinionated summaries: extract sentences with many opinionated terms

Summarization Corpus

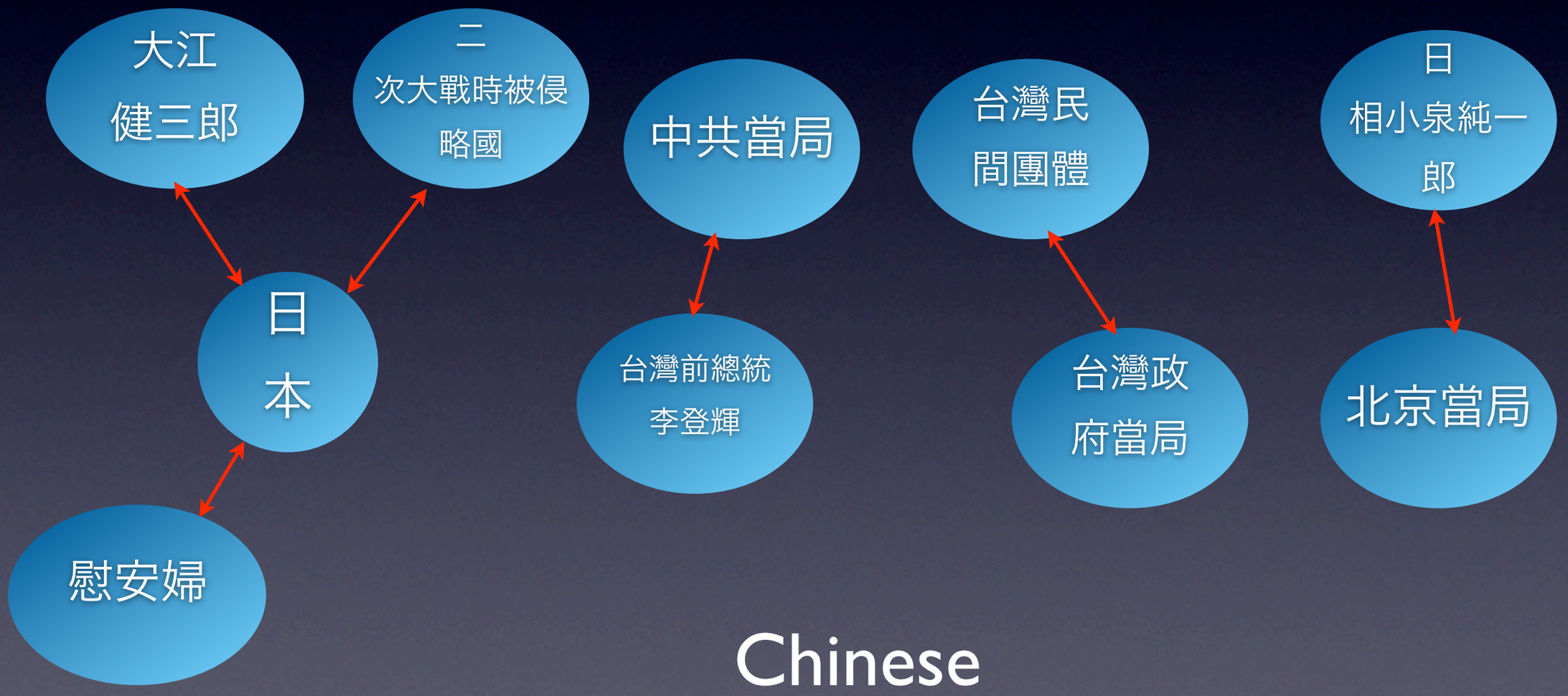
- Same data set used for Opinion Analysis
- C, E*, J three summaries (800文字, 400 words) for each topic
- Summary, conflict sentences, participants

Topic 010 Summary
Participants



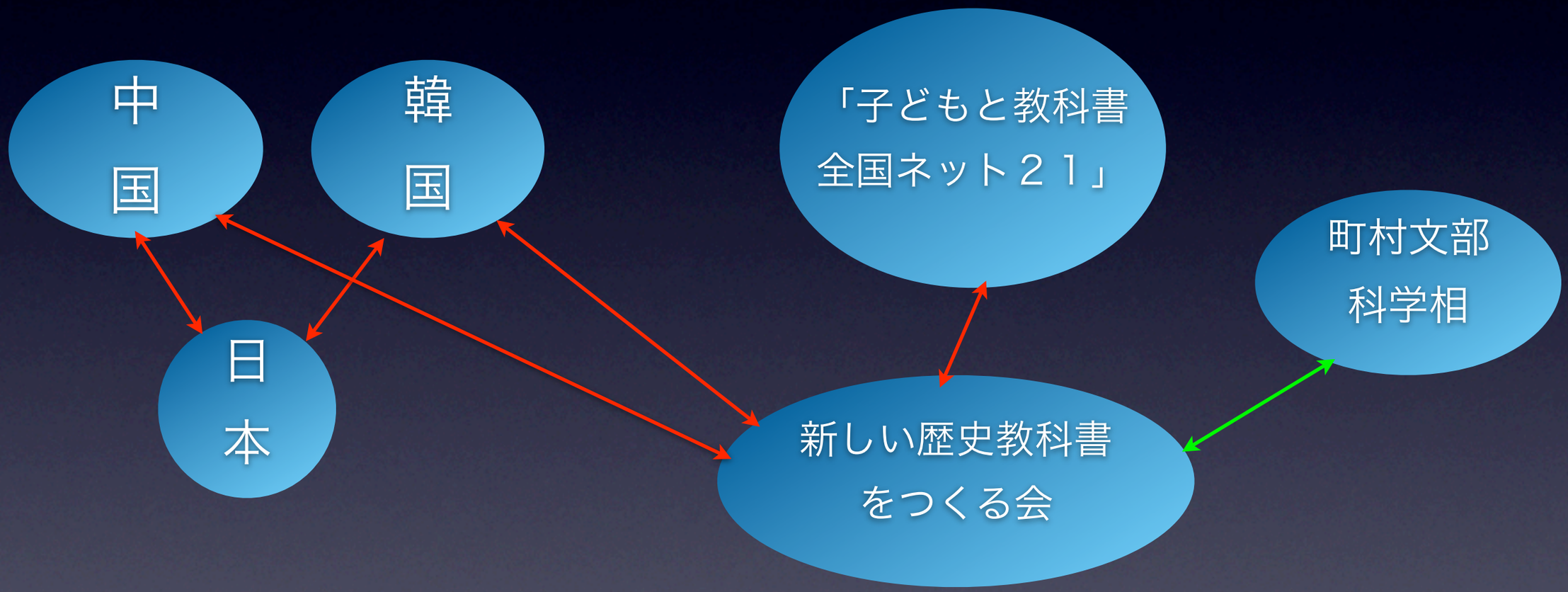
English

Topic 010 Summary Participants



Chinese

Topic 010 Summary
Participants



Japanese

Conclusions

- NTCIR Opinion Analysis Task
  - Cross-lingual, comparable corpus, opinion annotations
- LL, MI Statistical measures to identify interesting words in opinionated text
- Cross-lingual Summarization using simple counts and metrics, easily translated NEs, analysis results
- http://research.nii.ac.jp/ntcir/ntcir-ws6/