北海道大学 HOKKAIDO UNIVERSITY 荒木研究室 ARAKI LABORATORY

Language Media laboratory, Hokkaido University

Michal Ptaszynski, Pawel Dybala, Radoslaw Komuda, Rafal Rzepka and Kenji Araki

Development of Emoticon Database for Affect Analysis in Japanese

ABSTRACT

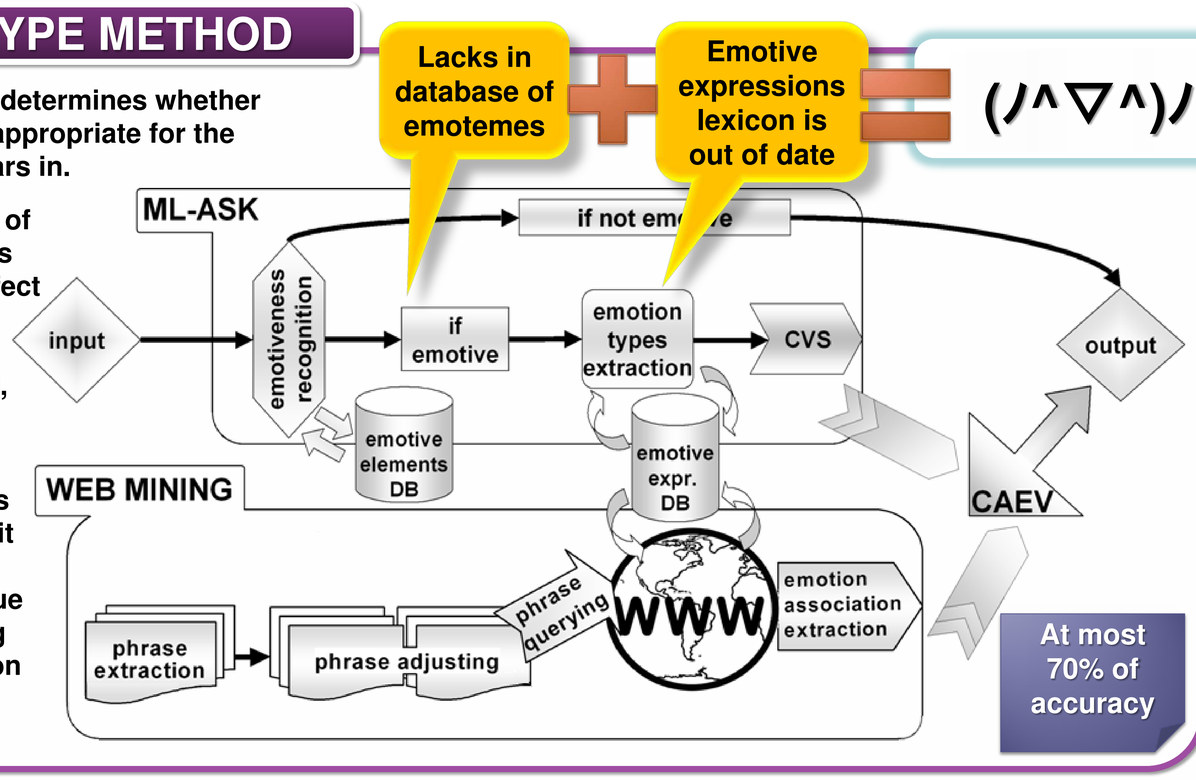
We present our work on creating a database of emoticons – face marks widely used to convey emotions in text-based online communication. The database is created by gathering emoticons from numerous dictionaries of face marks and online jargon. The inconsistencies in emotion classification provided by various dictionaries are solved by processing them with an affect analysis system developed previously. Having the emoticon database annotated automatically this way, we extract patterns from it patterns of semantic areas of emoticons, such as "eyes" and "mouths". Finally, we perform annotation of the semantic areas based on co-occurrence statistics and the theory of kinesics.

PROTOTYPE METHOD

The method [1] determines whether the emotion is appropriate for the context it appears in.

For recognition of emotions it uses Ptaszynski's affect analysis and annotation system, ML-Ask [2],

To verify the contextual appropriateness of those states it uses Shi's Web mining technique [3] for gathering emotive common sense from the Internet.



I. DEFINITION OF EMOTICON

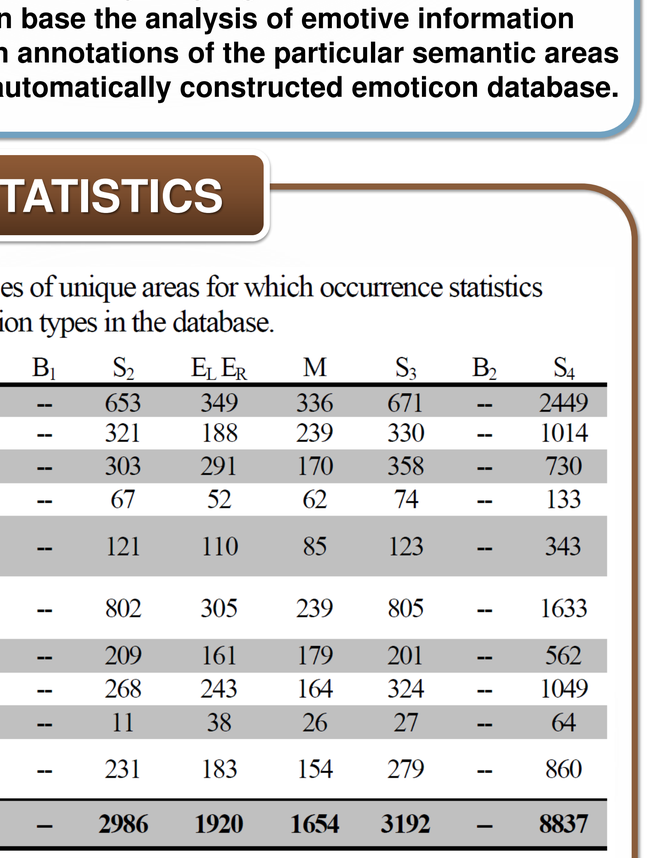
Emoticon - one-line string of symbols containing at least one set of semantic areas: "mouth" [M], "eyes" [E_L], [E_R], "emoticon borders" [B_1], [B_2], and "additional areas" [S_1] - [S_4]. Minimal emoticon set considered contains: "E_L,E_R", "E_L,M", "M,E_R", "S_1/S_2,E_L/M" or "M/E_R,S_3/S_4".

No. of sets	Emoticon	S_1	B_1	S_2	E_L	M	E_R	S_3	B_2	S_4	...
1	ヽ(。·ω·)ノ	ヽ	(	。	·	ω	·	N/A	)	ノ	
1	(ーー;)	N/A	(	N/A	ー	N/A	ー	;	)	N/A	
		SET 01								SET 02	
2	(^^)人(^^)	N/A	(	N/A	^	N/A	^	N/A	)	人	(^^)
2	☆- (●≧ ▿)人(▿ ≦ ●)- ☆	☆-	(	●	≧	▿	N/A	N/A	)	人	(▿ ≦ ●)- ☆
				SET 01		SET 02		SET 03		SET 04	
4	(▽ ´○)Ⅲ´★)ω´☆)∀ ´●)			(▽ ´○)		Ⅲ´★)		ω´☆)		∀ ´●)	

II. THEORY OF KINESICS

Kinesics - refers to all non-verbal behavior related to movement, such as postures, gestures and facial expressions (synonym of "body language"). Created by Birdwhistell in 1952-1970 [5, 6].

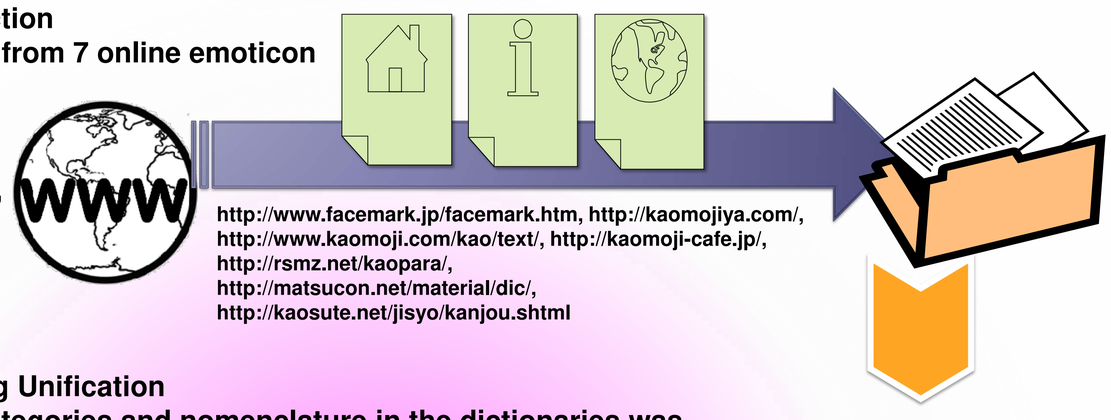
-Non-verbal behavior is used in everyday communication systematically and can be studied similarly to language.
-A minimal part distinguished in kinesics is a *kineme* - the smallest set of body moves containing a certain meaning, e.g. raising eyebrows, annotated by *kinegraphs*.



Emoticons: representations of body language in online text-based communication. → We can base the analysis of emotive information conveyed in emoticons on annotations of the particular semantic areas (kinemes) grouped in an automatically constructed emoticon database.

III. DATABASE OF EMOTICONS

1. Resource Collection
Extract emoticons from 7 online emoticon dictionaries:
Face-mark Party,
Kaomojiya,
Kaomoji-toshokan,
Kaomoji-café,
Kaomoji Paradise,
Kaomojisyo
Kaomoji Station



http://www.facemark.jp/facemark.htm, http://kaomojiya.com/, http://www.kaomoji.com/kao/text/, http://kaomoji-cafe.jp/, http://rsmz.net/kaopara/, http://matsucon.net/material/dic/, http://kaosute.net/jisyo/kanjou.shtml

2 Database Naming Unification
- The number of categories and nomenclature in the dictionaries was not unified.
-Processed all category names with ML-Ask [2] according to coherent classification of emotions based on Nakamura [4].
-Extracted 11,416 emoticons (10,137 unique ones, 89%).
-Group by emotion type

3 Extraction of Semantic Areas appearing in unique emoticons.

4 Annotation of Semantic Areas
- Occurrence frequency of the area in the emotion type database was calculated for every triplet, eye pair and mouth.
- Automatic annotation according to the probability of emotion expression (occurrence frequency).

```
Input: ・゜・(/Д`;)・゜・
Find match in raw emoticon database: ・゜・(/Д`;)・゜・
    If no match, localize ELMER triplet in the ELMER triplet
    database: /Д`
    If no triplet found, look for any ELMER combination;
    If no combination matched, find any ELER or M from separate
    semantic area database: /;`, Д
Localize emoticon borders B1,B2 : (,)
Localize additional areas S1,S2,S3,S4 : ・゜・,;,・゜・
Determine the emoticon structure: S1: ・゜・, B1:(, S2:N/A, ELER:
/`, M: Д, S3:;, B2:), S4: ・゜・
Look for next emoticon;
```

IV. DATABASE STATISTICS

Table 1. Distribution of all types of unique areas for which occurrence statistics was calculated across all emotion types in the database.

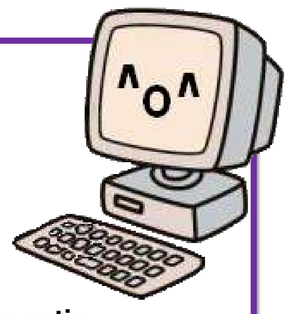
areas	E_LME_R	S_1	B_1	S_2	E_L E_R	M	S_3	B_2	S_4
joy, delight	1298	1469	--	653	349	336	671	--	2449
anger	741	525	--	321	188	239	330	--	1014
sadness,	702	350	--	303	291	170	358	--	730
fear	124	72	--	67	52	62	74	--	133
shame, shyness	315	169	--	121	110	85	123	--	343
liking, fondness	1079	1092	--	802	305	239	805	--	1633
dislike	527	337	--	209	161	179	201	--	562
excitement	670	700	--	268	243	164	324	--	1049
relief	81	50	--	11	38	26	27	--	64
surprise, amazement	648	405	--	231	183	154	279	--	860
overall	**6185**	**5169**	**--**	**2986**	**1920**	**1654**	**3192**	**--**	**8837**

Database Coverage
Unique raw emoticons = 10,137
Unique E_LME_R triplets = 6,185
unique pairs of eyes E_L,E_R = 1,920
Unique mouths M = 1,654
All possible combinations of triplets E_L,E_R × M = 3,175,680 possibilities

Unique raw emoticons (10,137) = 0.3% of full coverage
Unique triplets (6,185) = 0.2% of full coverage.

Without our approach (kinesics) we loose 97% of possible coverage!

CONCLUSIONS



We presented a description of a database of emoticons to be used in further research on affect analysis. The database contains over ten thousand of unique emoticons collected from the Internet. These emoticons are automatically distributed into emotion classes. The emoticons are divided into semantic areas (mouths or eyes). The division into semantic areas is based on Birdwhistell's [5,6] theory of kinesics.

The database will be used in emoticon analysis system. The database contains over ten thousand of raw emoticons and over 3 million of possible combinations covering most of the possibilities.
Planned evaluation: 1) emoticon detection in a sentence; 2) emoticon extraction from a sentence; 3) division of emoticon into semantic areas; 4) emotion classification of emoticons.

REFERENCES

1.M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki, "Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States." In Proceedings of IJCAI-09, Pasadena, CA, USA, pp. 1469-1474, 2009.
2.M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki, "Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum-," In Proc. of PACLING-09, Japan, 2009.
3.W. Shi, R. Rzepka and K. Araki, "Emotive Information Discovery from User Textual Input Using Causal Associations from the Internet," FIT-08, pp. 267-268, 2008.
4.A. Nakamura, Kanjo hyogen jiten, Tokyodo, 1993.
5.R. L. Birdwhistell, Introduction to kinesics: an annotation system for analysis of body motion and gesture. University of Kentucky Press, 1952.
6.R. L. Birdwhistell, Kinesics and Context. University of Pennsylvania Press, Philadelphia, 1970.