A Comparison between Microblog Corpus and Balanced Corpus from Linguistic and Sentimental Perspectives

Yi-jie Tang, Chang-Ye Li and Hsin-Hsi Chen
NLP Lab, CSIE, National Taiwan University
hhchen@csie.ntu.edu.tw, tangyj@nlg.csie.ntu.edu.tw

Abstract

Processing short messages is a challenging task in natural language processing. This paper analyzes the differences between Internet short messages and general articles by comparing the Plurk Corpus and the Sinica Balanced Corpus. Likelihood ratio and the *tóngyìcícílín* (同義詞詞林) thesaurus are adopted to analyze the lexical semantics of frequent terms in each corpus. Furthermore, the NTUSD sentiment dictionary is used to compare the sentiment distribution of the two corpora. The result is also applied to sentiment transition analysis.

Finding most critical words

$$lr_{AB}(w_i) = \log \frac{\frac{f_A(w_i)}{|A|}}{\frac{f_B(w_i)}{|B|}}$$

$lr_{AB}(w_i)$: Log of likelihood ratio of relative frequencies.
$f_A(w_i)$ and $f_B(w_i)$: Frequencies of w_i in A and B, respectively.
$|A|$ and $|B|$: Total words in A and B, respectively.

Analysis of Lexical Semantics

Top 10 high $lr_{Plurk\text{-}Sinica}$ words:

	1st	2nd	3rd	4th	5th
Chinese	一個	早安	噗	阿	晚安
English	*yíge* (indefinite article)	good morning	*pū* (a Plurk message)	*a* (sentence-final particle)	Good night or good evening
	6th	**7th**	**8th**	**9th**	**10th**
Chinese	超	囉	耶	睡	睡覺
English	super	*Lou* (sentence-final particle)	*Ye* (sentence-final particle)	sleep	sleep

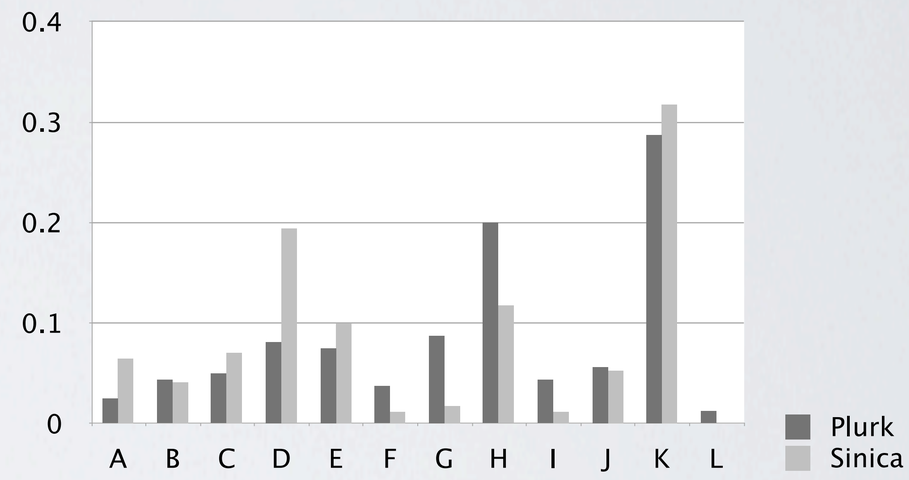
Top 10 high $lr_{Sinica\text{-}Plurk}$ words:

	1st	2nd	3rd	4th	5th
Chinese	其	表示	並	則	各
English	*qí* (3rd-peson possessive)	express; indicate	and; also	*zé* (conjunc-tion)	each
	6th	**7th**	**8th**	**9th**	**10th**
Chinese	由	所	每	及	於
English	from	*suǒ* (particle)	every	and	*yú* (preposi-tion)

Semantic Categories in Cilin

A. Person B. Object C. Time and space
D. Abstract thing E. Characteristics F. Motion
G. Mental activity H. Activity
I. Phenomenon and condition J. Relation
K. Auxiliary L. Greeting

Sense Distribution of Top 100 Words

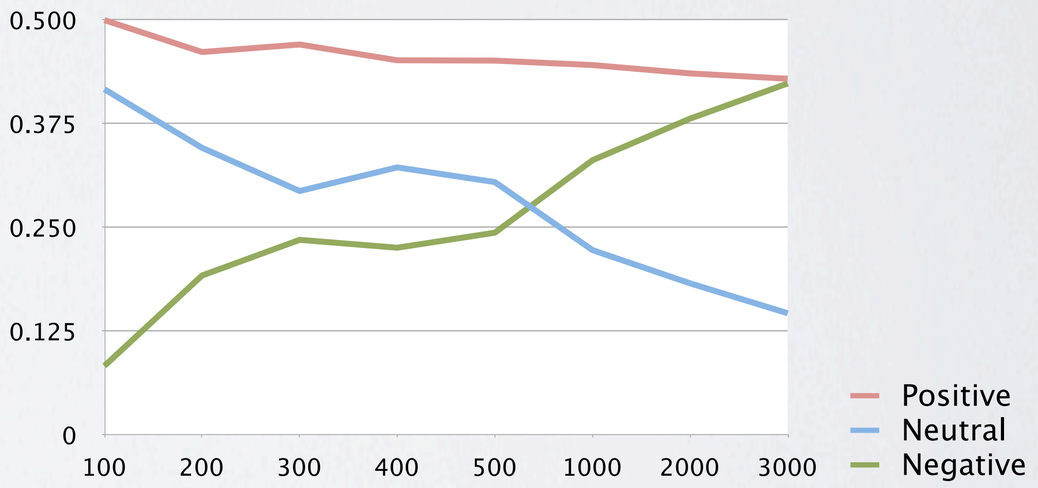


Literal Text vs. Spoken Language

(1) Monosyllabic or disyllabic
(2) Literal Chinese vs. Spoken Chinese
(3) Internet language usage

Sentiment Analysis

Ratio of sentiment words in top *n* frequent words in Plurk



Ratio of sentiment words in top *n* frequent words in Sinica