2.2 Personal Language Habit

Same words often have different meanings when used by persons with different language habit. For an individual I, to characterize I's opinion more accurately and avoid the interference from I's personal language habit, we normalize I's each language feature f with I's personal language habit value $H_f(I)$. $H_f(I)$ is measured by Formula (1), where $f(I, I_i)$ is the f's value of the languages said by I to another individual I_i, and C is the set of all individuals who communicate with I.

$$H_f(I) = \frac{1}{|C|}\sum_{I_i \in C} f(I, I_i) \quad (1)$$

Then $f'(I, I_i)$ is the normalized value of $f(I, I_i)$ according to I's personal language habit, which can be calculated with Formula (2):

$$f'(I, I_i) = \frac{f(I, I_i) - H_f(I)}{H_f(I)} \quad (2)$$

3. EXPERIMENTS

We utilize the Enron email dataset, which contains 0.5M mails among employees. We retained only those interrelationships where at least 15 emails were sent in each direction. The filtered set contains 1078 interrelationships between 647 individuals.

For each directed pair of individuals <I_i, I_j>, we calculated four linguistic features with the emails' content sent from I_i to I_j, to characterize I_i*'s* opinion on his interrelationship with I_j:

(1) "Frequency": average emails count per day from I_i to I_j.

(2) "Length": average words count per email from I_i to I_j.

(3) "Quality": average perplexity score per sentence in the emails from I_i to I_j. Higher perplexity score means lower language quality. The perplexity score is calculated by the SRI language modeling toolkit.

(4) "Sentiment": average sentiment score per word in the emails from I_i to I_j. X-Similarity sentiment dictionary is used to score the positive, negative and neuter words as 1, -1 and 0, respectively.

Firstly, for 1078 pairs of individuals, we calculated the paired T-statistic score between their pair-wise language feature values of exchange emails content. In Table 1, the 1st, 2nd & 3rd column show the results of original pair-wise language features, the pair-wise personal language habit (Formula (1)) and the normalized pair-wise language features (Formula (2)), respectively. For each column, the t value and the ration between the average pair-wise difference and average value of each feature are given.

Table 1. Pair-wise language features differences.

Feature	**t value / (average difference)/(average value)**		
	Original	**Habit**	**Normalized(Opinion)**
Frequency	13/0.6487	17/0.6316	22/0.8823
Length	12/0.7636	16/0.7039	25/0.8059
Quality	12/0.5311	12/0.5006	24/0.8323
Sentiment	23/0.4545	23/0.4141	23/0.7931

In Table 1, in each column, the pair-wise differences of each language feature is significant (>95%). This significance is also verified by the remarkable ration degree between the average pair-wise difference and average value of each feature. These experimental results indicate that on this dataset, the pair-wise interactive language features are significantly asymmetric (column 1). This asymmetry is a joint effect of the asymmetry of individuals' subjective opinions (column 3) and the asymmetry of their personal language habits (column 2). Besides, the opinions' asymmetry is more significant than that of the habits.

Secondly, for each individual, we also calculate the Pearson correlation score between his language styles and those of his counterparts. Higher correlation indicates that one tends to adapt his language use to different conversation partners, and thus reduce the asymmetry degree. In Table 2, for each feature, we calculate the average correlation scores of the individuals who have the top/last 10 average/deviation feature values, respectively. It can be observed that on each feature, the individuals with top10 average/deviation of the language feature values obtain better correlation scores than those having the last10 average/deviation. We suppose that higher average score on sentiment and quality indicate more positive personality, and higher deviation score on all features tend to indicate more flexible personality. With this hypothesis, results in Table 2 indicate that the individuals who have more positive and flexible personality (Top 10 individuals) tend to adapt to the counterparts' style in the communication (higher correlation) and thus lead to less asymmetry of the subjective opinions on the interrelationship. This experimental result inspires us to predicate the asymmetry of the subjective asymmetry with the help of individuals' personality characters.

Table 2. Pair-wise language feature correlations for different personality (Avg indicates Postive & Dev indicates Flexible).

Feature	Top 10 on Avg.	Last 10 on Avg.	Top 10 on Dev.	Last 10 on Dev.
Frequency	.5621	.4031	.6112	.4452
Length	.1037	-.0538	.0015	-.042
Quality	.1239	.0806	.2477	.0709
Sentiment	.0588	-.0205	.1037	-.0227

4. CONCLUSIONS AND FUTURE WORK

Our experimental investigation makes some attempts to provide suggestive evidences for the subjective asymmetry of interrelationships, and potentially leads to a promising direction to model the social relationship asymmetrically from the pair-wise subjective opinions with interactive languages. To obtain more convincing result, we need to investigate more diversified datasets from social media. And we will also try to build asymmetric model of interrelationships integrating subjective language features and objective features.

5. ACKNOWLEDGMENTS

This work is supported by the National High-tech R&D Program of China (2015AA015403), Chinese National Program on Key Basic Research Project (2013CB329304), Major Project of National Social Science Fund of China (14ZDB153) and Tianjin Younger Natural Science Foundation (14JCQNJC00400).

6. REFERENCES

[1] Habermas, J. 1981. *The Theory of Communicative Action*. Beacon Press.

[2] Holmes, J. 2013. *An introduction to sociolinguistics (4 edition)*. Routledge, London and New York.

[3] Sapir, E. 1929. *The Status of Linguistics as a Science*. Linguistic Society of America.

[4] West, R., et al. 2014. Exploiting Social Network Structure for Person-to-Person Sentiment Analysis. *Transactions of the Association for Computational Linguistics*.