Construction of a Chinese Corpus for the Analysis of the Emotionality of Metaphorical Expressions

Dongyu Zhang, Hongfei Lin, Liang Yang, Shaowu Zhang, Bo Xu

ABSTRACT

We construct a corpus on metaphor, with 5,605 manually annotated sentences in Chinese. We present an annotation scheme that contains annotations of linguistic metaphors, emotional categories and intensity. The annotation agreement analyses for multiple annotators are described. We also use the corpus to explore and analyze the emotionality of metaphors. To the best of our knowledge, this is the first relatively large metaphor corpus with an annotation of emotions in Chinese.

ANNOTATION SCHEME

Data Collection. We selected data from a wide range of sources including books, journals, movie scripts, and networks. In addition, we focused on sources with rich emotional information such as microblogs.

Annotation Model. We annotated metaphorical sentences with target and source domain vocabulary, emotion categories and intensity, metaphor categories (verb/noun), data sources, and metaphor devices such as "like," "as," etc. as "indicators". The annotation model is:

Table 1: Information on Corpus Sources

Sources	Characters	Words	Sentences
Books	258,9723	182,046	9,6182
Journals	52,0743	39,7065	2,1640
Scripts	168,236	108,184	11,852
Networks	124,6329	87,2210	9,5153
Total	4,525,031	1,559,505	224,827

(Target, Source, EmotionCategory, Intensity, MetaphorCategory, [indicator], DataSource)

Annotation Process. Given detailed guidelines and a formal training lesson, seven annotators used cross-validation methods for annotation.

Difficulties. 1) Different opinions on the distinction between literal and metaphorical sense. 2)The inter-annotator agreement on the choice of emotion category.

Quality Monitoring and Control Methods. 1) Entry interface. We provided an interface that allowed us to enter information precisely and quickly. 2) Error correction. We used emotional lexicon ontology as a support tool to correct human errors.

RELIABILITY

The agreement on the identification of source target domain words was $\kappa=0.82$ (kappa score); the choice of emotion category scored $\kappa=0.68$; classification of emotion intensity was $\kappa=0.58$.

ANALYSIS

We annotated 82% sentences out of total metaphorical sentences as containing emotions. The most frequent emotion is love. Our results indicate emotions are related to some particular source and target words.

Fig 1. The number of sentences of each emotion category
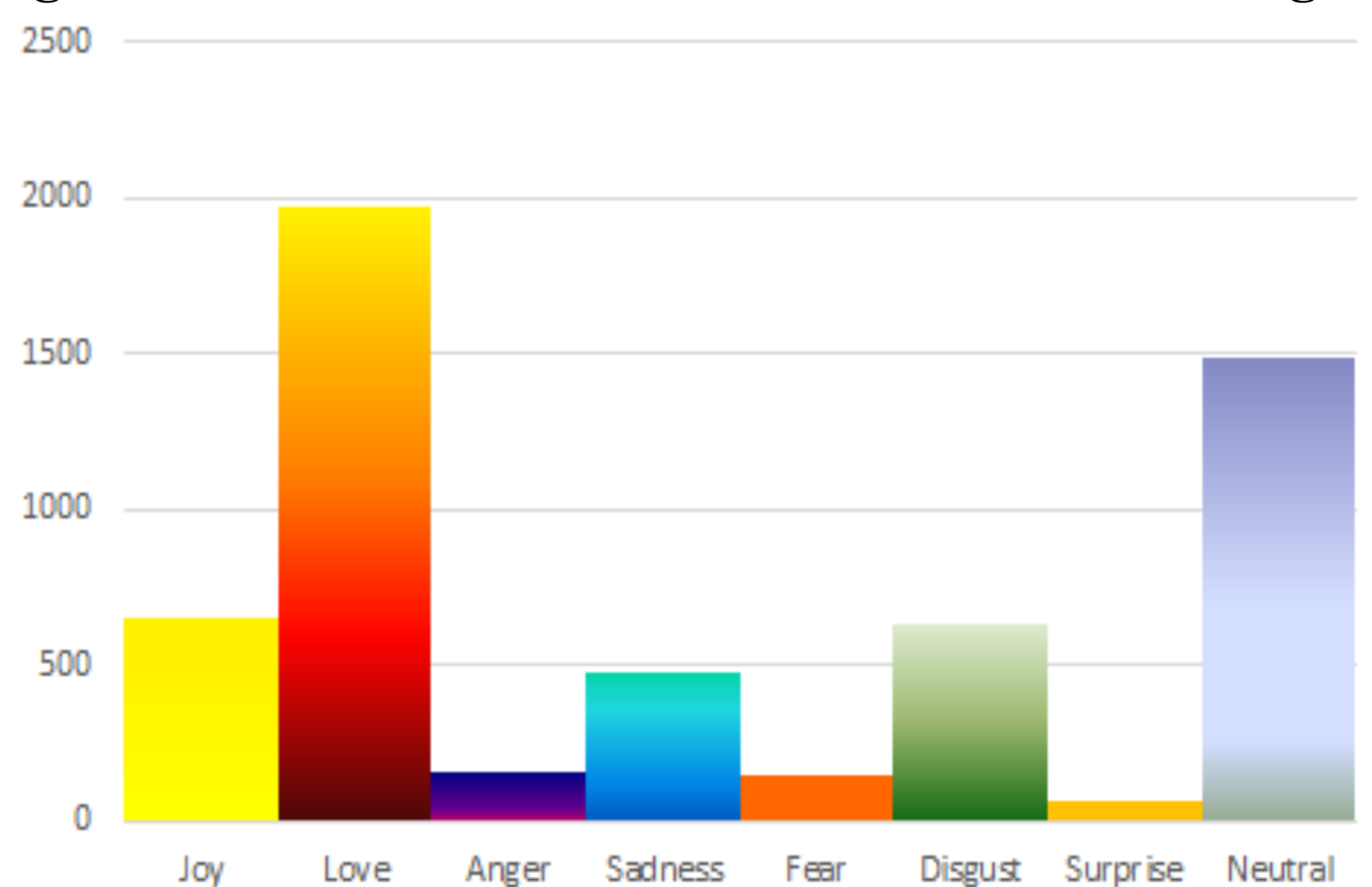
CONCLUSION

With 5,605 diverse instances and 101,616 Chinese characters of metaphor, our corpus provides an important resource with relatively fine-grained sorting and annotation with both metaphors and emotion. We annotated for a range of emotions, rather than just positive and negative, as is often done. Furthermore, we hope research using bilingual resources will be conducted on the datasets we have released.