A CORPUS-BASED APPROACH TO FINGERPRINTING STYLISTIC FEATURES OF CLASSICAL CHINESE POETRY: A CASE STUDY OF LIU YONG AND SU SHI

Alex Chengyu Fang[*] **Wan-yin Li**
City University of Hong Kong

Jing Cao
Zhongnan University of Economics and Law, Wuhan

ABSTRACT

In this article,[†] we describe an experiment that is aimed at the use of ontological knowledge to identify the stylistic features of classical Chinese poetry.[1] In particular, this article addresses the task of automatic authorship attribution of classical Chinese poems. This work is motivated by the understanding that the creative language use by different poets can be characterised through their creative use of imageries which can be captured through ontological annotation. A corpus of lyric songs written by Liu Yong and Su Shi in the Song Dynasty[2] is used, which is word segmented and ontologically annotated. Different feature sets are constructed that represent all the possible combinations of word tokens and their ontological annotations. Machine learning techniques are applied and SVM used to evaluate the performance of the different feature sets. Empirical results show that word tokens alone can be used to achieve

[*] Corresponding author: alex.fang@cityu.edu.hk

[†] This current paper is a revised paper from the Roundtable Conference on Linguistic Corpus and Corpus Linguistics in the Chinese Context, held in May 2011 in the Hong Kong Institute of Education. A substantially different version of the paper has been published after the conference as "In search of poetic discourse of classical Chinese poetry: An imagery-based stylistic analysis of Liu Yong and Su Shi" by Fang, A.C., W.-Y. Li and J. Cao in *Chinese Language and Discourse* 2(2):232-249.

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 63-81

2409-2878/2015/25-0004

an accuracy of 87% in the task of authorship attribution between Liu Yong and Su Shi. More interestingly, ontological knowledge is shown to produce significant performance gains when combined with word tokens. This observation is reinforced by the fact that most of the feature sets with ontological annotation outperform the use of bare word tokens as features. Specifically, our empirical experiment shows that word tokens combined with ontological annotations achieve an overall accuracy of 89%, expressed in F-value, for the task of authorship attribution between Liu Yong and Su Shi.

KEYWORDS

Syntax **O**ntology **I**magery **M**achine learning **P**oetic style

基於語料庫的古詩詞文學風格辨識：柳永及蘇軾詩詞範例研究

方稱宇　李昀燕　曹競

香港城市大學　中南財經政法大學，武漢

提要

本文描述了基於本體知識而設計的一系列古典詩詞文學風格識別的實驗，並著重於有關古體詩詞著作權歸屬的鑑定。文章立意於在詩詞創作中，不同的作者傾向於應用別具一格並具有個人風格的意象創作詞語，而這些詞語可由已標註的本體知識庫追溯到相關意象。本文所採用的語料庫包含了宋代代表詞作家柳永和蘇軾的詩詞，並且已做了詞語切分及本體知識標註。實驗採用了機器學習技術中的SVM算法，對所有已切分詞彙及相關本體的不同組合特徵進行了反覆測試。實驗結果顯示了單純使用詞彙組合為特徵對柳永和蘇軾作品的著作權歸屬鑑定可達到87%的精確度。實驗結果進一步表明，若結合相關詞彙的本體知識，精確度方面則有明顯的提高：當使用由詞彙及相關本體知識所構成的組合特徵集進行測試時，實驗總體F-value則高達89%，從而以實證結果肯定了本體知識的使用對於著作權歸屬鑑定的實際貢獻。

關鍵詞

語法　本體知識　意象　機器學習　詩詞文學風格