Evaluation of Ontology Knowledge in Chinese Classical Poetry Classification

Alex Chengyu Fang and Wanyin Li

Department of Chinese, Translation and Linguistics
City University of Hong Kong, Hong Kong SAR
{acfang, claireli}@cityu.edu.hk

Abstract

This paper describes preliminary research in the use of ontological knowledge for the task of automatically classifying classical Chinese poetry (CCCP) according to authorship. Based on a collection of poems written by Liu Yong (987–1053 AD) and Su Shi (1037– 1101 AD), which have been analyzed according to a taxonomy of ontological entities at the lexical level, the research looks into the issue of whether characteristic features can be automatically extracted as important stylistic differences between the two poets. This paper examines the efficiency of different ontological concepts as features in CCCP using Support Vector Machine (SVMs). The experiment shows that an integration of ontological knowledge and bags-of-words (BoW) produces a higher precision for CCCP than BoW only with an overall increase of 2.1% and 2.2% in terms of precision and *F*-score.

Application of Ontology in CCCP

The use of ontologies has been applied to the task of automatic text classification (Cumbo, Iiritano and Rullo 2004; Melo and Siersdorfer 2007; Janik and Kochut 2008; Netzer et al. 2009). In this paper, we propose to apply ontological knowledge to the task of automatically classifying classical Chinese poetry (CCCP) according to authorship. The ontological information is extracted from a dedicated database of Chinese classical poetry (Lo 2008). LibSVM (Chang and Lin 2001) is adopted for the evaluation of differentiating features.

Preparation of Features

There are 10,669 word tokens in poems from Liu Yong and 12,416 word tokens from Su Shi. The two collections have been pre-processed by segmenting the poems into word tokens and linking each word token to a concept layer in the ontology. Each poem p_i is represented as a collection of the features f_m as $p_i = \{f_1, f_2, f_3, \ldots, f_m\}$, where m denotes the frequency of feature f in p_i

The candidate features f_m are selected as bags of word tokens (BoW) in the raw poems and/or their mapped ontological concepts $O_{k,j}$ in different layers ($1 \leq k \leq 5$) with respect to the j classes under each layer, while j varies from 6 to 2,440 from the first layer to the fifth layer. Ontological concepts $\{\forall_j O_{kj}\}$ are individually saved over all poems based on each ontological layer k. For example, the six classes of ontology concepts $O_1 = \{O_{1j}\}(j=1..6)$ in the first layer over all poems are stored separately from the ones in other four layers. To compare the effectiveness of the ontological concepts $O_{k,j}$ from different layers in CCCP, the different feature sets from all combinations of five layers are built up. For example, $BoWO_1O_2O_3O_4O_5$ constitute one feature set, and $O_2O_3O_4$ form another feature set. In all, 62 feature sets were extracted and tested.

Term weighting, term normalization and feature selection with Weka (Witten and Frank 2005) are applied to refine the default features as the input to LibSVM. In the process of term weighting, term frequency in inversed document frequency (*tf*idf*) weighting scheme is applied to represent each poem. The normalization of word frequencies to the average length of the poem is also conducted in this process. Due to the unbalancing of more unique words in poems but higher word frequency in different ontology layers, the term normalization process is used to normalize all numeric values in whole selection to the result values of [0,1]. Each normalized value x' is calculated by $x' = (x_{current} - x_{min})/(x_{max} - x_{current})$. Finally the selection of features is applied by applying the package of attributeSelection from Weka. We use the search method of BestFirst to search the space of attribute subsets from each different feature combinations of BoW and/or O_k ($1 \leq k \leq 5$), and the evaluation method of CfsSubsetEval to evaluate the performance of the attribute subsets.

Evaluation of Ontologies in CCCP

Figures 1 and 2 show the performance of top 30 feature combinations sorted according to the overall F-score.

Figure 1 expresses in overall evaluation scheme, most of the features from BoW integrated by the ontological concepts (for simplicity, WO_{145} stands for $BoWO_1O_4O_5$ in Figures 1 and 2) outperform the features of BoW, some of the features from ontology concepts only may also achieve better performance than BoW, such as the ontological features of $O_1O_3O_4$ outperform the features selected from BoW. From Figure 2, the F-score variation pattern holds consistent with respect to the evaluation of Liu Yong, Su Shi, and a combination of both (*Overall*). However, the improvement through the use of ontological knowledge for classifying the poems of Liu Yong is more substantial than classifying the ones written by Su Shi. An improvement of 3.4% of the F-score is observed for Liu Yong compared with 1.6% improvement in F-score for Su Shi.

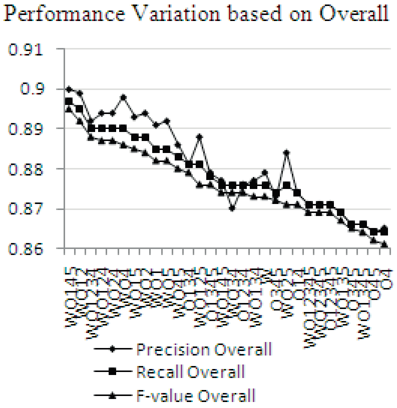

Figure 1. Overall Performance Variation for the Top 30 Features Combination Sorted in F-score

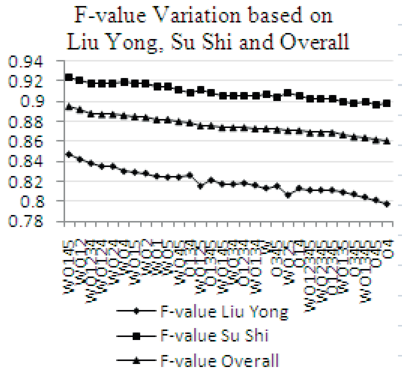

Figure 2. F-score Variation for the top 30 Features Combination Sorted in Overall F-score

From the above analysis, the best feature combination is shown to be $BoWO_1O_4O_5$, which is ranked as the top one in the sorted lists from F-score of Liu Yong, Su Shi, and overall. Table 1 summaries the performance improvement gained through $BoWO_1O_4O_5$ over BoW only, in terms of precision, recall, and F-score with respect to Liu Yong, Su Shi and a combination of both.

	Liu Yong	*Su Shi*	*Overall*
Precision (%)	2.6	1.9	2.1
Recall (%)	3.9	1.1	2.1
F-score (%)	3.4	1.6	2.2

Table 1. Improvement in precision, recall and F-score gained through $BoWO_1O_4O_5$ over BoW

Conclusion and Discussion

The study reported in this article shows that the use of ontological information produced better performance for the automatic attribution of Chinese classical poems according to poets. The ontological concepts conveyed more valuable background knowledge and therefore helped to produce a selection of better features of more differentiating power. By integrating ontological concepts with the BoW features, the experiment achieved an increase of 2.1% for precision and 2.1% for F-score over the performance measured according to BoW only.

Acknowledgements

The research reported in this article was supported in part by research grants from City University of Hong Kong (Project Nos 7008002, 7008062 and 9610126). The authors would also like to acknowledge the financial support received from the Dept of Chinese, Translation and Linguistics of the same university.

References

Chang, C.-C., and Lin, C.-J. 2001. *LibSVM: A Library for Support Vector Machines*. Software package available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cumbo, C., Iiritano, S., and Rullo, P. 2004. Combing Logic Programming and Domain Ontologies for Text Classification. *Convegno Italiano di Logica Computazionale*, Parma, Italy.

Janik, M., and Kochut, K.J. 2008. Wikipedia in Action: Ontological Knowledge in Text Categorization. *2nd International Conference on Semantic Computing*, CA, USA.

Lo, F. 2008. The Research of Building a Semantic Category System Based on the Language Characteristic of Chinese Poetry. *Proceedings of the 9th Cross-Strait Symposium on Library Information Science, Wuhan, China.*

Melo, G. de, and Siersdorfer, S. 2007. Multilingual Text Classification using Ontologies. *Proceedings of the 29th European Conference on Information Retrieval*, Italy.

Netzer, Y., Gabay, D., Adler, M., Goldberg, Y. and Elhadad, M. 2009. Ontology Evaluation through Text Classification. In Chang, K.C., Wang, W., Chen, L., Ellis, C.A., Hsu, C.-H., Tsoi, A.C., Wang, H., Lin, X., Yang, Y., and Xu, J. eds., *Advances in Web and Network Technologies, and Information Management*. Berlin: Springer. pp. 210-221.

Witten, I.H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.