Journal of Southeast University (English Edition) Vol. 29, No. 1, pp. 22 - 26 Mar. 2013 ISSN 1003—7985

Feature combination via importance-inhibition analysis

Yang Sichun[1,2] Gao Chao[3] Yao Jiamin[2] Dai Xinyu[1] Chen Jiajun[1]

([1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)
([2]School of Computer Science, Anhui University of Technology, Maanshan 243032, China)
([3]School of Computer Science and Information Engineering, Chuzhou University, Chuzhou 239000, China)

Abstract: A new method for combining features via importance-inhibition analysis (IIA) is described to obtain more effective feature combination in learning question classification. Features are combined based on the inhibition among features as well as the importance of individual features. Experimental results on the Chinese questions set show that, the IIA method shows a gradual increase in average and maximum accuracies at all feature combinations, and achieves great improvement over the importance analysis(IA) method on the whole. Moreover, the IIA method achieves the same highest accuracy as the one by the exhaustive method, and further improves the performance of question classification.

Key words: question answering system; question classification; feature combination; importance-inhibition analysis

doi: 10.3969/j.issn.1003-7985.2013.01.005

Automatic question answering (QA) [1] is a hot research direction in the field of natural language processing (NLP) and information retrieval (IR), which allows users to ask questions in natural language, and returns concise and accurate answers. QA systems include three major modules, namely question analysis, paragraph retrieval and answer extraction. As a crucial component of question analysis, question classification classifies questions into several semantic categories which indicate the expected semantic type of answers to questions. The semantic category of a question helps to filter out irrelevant answer candidates, and determine the answer selection strategies.

In current research on question classification, the method based on machine learning is widely used, and features are the key to building an accurate question classifier[2-10]. Li et al.[2-3] presented a hierarchical classifier based on the sparse network of winnows (SNoW) architecture, and made use of rich features, such as words, parts of speech, named entity, chunk, head chunk, and class-specific words. Zhang et al.[4] proposed a tree kernel support vector machine classifier, and took advantage of the structural information of questions. Huang et al.[5-6] extracted head word features and presented two approaches to augment hypernyms of such head words using WordNet. However, when used to train question classifiers, these features were almost combined incrementally via importance analysis (IA) which is based on the importance of individual features. This method is effective when using only a few features, but for very rich features, it may prevent question classification from further improvement due to the problem of ignoring the inhibition among features.

In order to alleviate this problem, this paper proposes a new method for combining features via importance-inhibition analysis (IIA). By taking into account the inhibition among features as well as the importance of individual features, the IIA method more objectively depicts the process of combining features, and can further improve the performance of question classification. Experimental results on the Chinese questions set show that the IIA method performs more effectively than the IA method on the whole, and achieves the same highest accuracy as the one by the exhaustive method.

1 Feature Extraction

We use an open and free available language technology platform (LTP) (http://ir.hit.edu.cn/demo/ltp) which integrates ten key Chinese processing modules on morphology, word sense, syntax, semantics and other document analysis, and take the question "中国哪一条河流经过的省份最多?(Which river flows through most provinces in China?)" as an example. The result of word segmentation, POS tagging, named entity recognition and dependency parsing of the sample question is presented in Fig. 1.

We extract bag-of-words (BOW), part-of-speech (POS), word sense (WSD, WSDm), named entity (NE), dependency relation (R) and parent word (P) as basic features. Here, WSD is the 3-layer coding, i.e., coarse, medium and fine grained categories in the semantic dictionary "TongYiCiCiLin", while WSDm is the

Received 2012-05-03.

Biographies: Yang Sichun(1970—), male, graduate; Chen Jiajun (corresponding author), male, doctor, professor, chenjj@nju.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 61003112, 61170181), the Open Research Fund of State Key Laboratory for Novel Software Technology of China (No. KFKT2010B02), the Key Project of Natural Science Research for Anhui Colleges of China (No. KJ2011A048).

Citation: Yang Sichun, Gao Chao, Yao Jiamin, et al. Feature combination via importance-inhibition analysis[J]. Journal of Southeast University (English Edition), 2013, 29(1): 22 - 26. [doi: 10.3969/j.issn.1003-7985.2013.01.005]

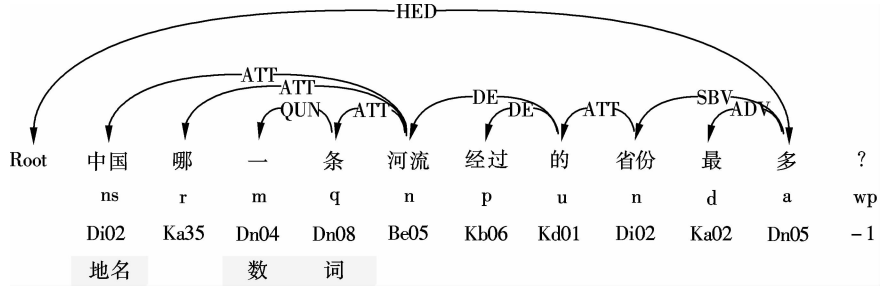

Fig. 1 Analysis result of the sample question with LTP platform

2-layer, i. e., coarse and medium grained word category. Tab. 1 gives the features and their values of the sample question.

Tab. 1 Features and their values of the sample question

Feature	Values
BOW	中国, 哪, 一, 条, 河流, 经过, 的, 省份, 最, 多, ?
POS	ns, r, m, q, n, p, u, d, a, wp
NE	(中国, Ns, S), (一, Nm, B), (条, Nm, E)
WSD	Di02, Ka35, Dn04, Dn08, Be05, Kb06, Kd01, Di02, Ka02, Dn05, -1
WSDm	Di, Ka, Dn, Be, Kb, Kd, Di, -1
R	ATT, QUN, DE, SBV, ADV, HED
P	条, 河流, 的, 省份, 多

2 Combining Features via Importance-inhibition Analysis

The basic features described above belong to different syntactic and semantic categories, and contribute to question classification from various levels of language knowledge. We combine these basic features to further improve the performance of question classification. Since the BOW feature is the basis of other features, it is always combined with other features. For example, the POS feature follows the BOW feature when these two types of features are combined.

With respect to the methods for combining features, the most intuitive one is the exhaustive method which lists all the feature combinations one by one. The exhaustive method is inefficient and not feasible in practical applications. In existing literature, combining features is conducted just on the basis of the importance of the features. However, this method may prevent it from further improvement on question classification due to the problem of ignoring the inhibition among features. For example, the dependency relation feature R and the POS feature belong to the same syntactic category, and they both contribute to question classification. However, since R covers POS to a large extent in syntactic expression, R will inhibit POS when they appear in the same feature combination. Similarly, the word sense features WSD and WSDm belong to the same semantic category, since the difference between WSD and WSDm is not obvious, they will inhibit each other when they are present at the same feature combination. From the above discussions, we find that an effective method for combining features should take into account the inhibition among features as well as the importance of individual features.

In this paper, we propose a new method for combining features via importance-inhibition analysis. Before introducing the IIA method in detail, we should specify some notations. In our importance-inhibition analysis setting, the feature set is a basic concept following the common feature combination.

A feature set F consists of each feature f_i extracted from a question, i. e. $F = \{f_i \mid i = 1, 2, \ldots\}$; F' is a subset of F, and consists of each feature $f^{(i)}$ which has side effects for feature combinations, i. e. $F' = \{f^{(i)} \mid i = 1, 2, \ldots\}$; $F_i^{(j)}$ denotes the j-th one in the i-th round of feature combination, and it is a subset of F; F_i^* denotes a feature combination with the highest accuracy in the i-th round, and it is also a subset of F.

Now we can give some formal definitions.

Definition 1 (importance) Given features f_i and f_j, f_i is more important than f_j if the accuracy of f_i is higher than that of f_j.

Definition 2 (inhibition) Given a feature f_i and a feature combination $F_i^{(j)}$, there exists inhibition between $F_i^{(j)}$ and f_i if the accuracy of the feature combination $F_i^{(j)} \cup \{f_i\}$ is lower than that of $F_i^{(j)}$ or f_i.

Definition 3 (k_ary combination) Given a feature set $F_i^{(j)}$, it is a k_ary feature combination in which k features are contained.

Definition 4 (best k_ary combination) Given a $(k-1)$_ary combination $F_i^{(j)}$ and a candidate feature f_i, $F_i^{(j)} \cup \{f_i\}$ is the best k_ary combination if it has the highest accuracy in the current round of feature combinations.

Now let us move to the details of the IIA method. From the above definitions, we can easily see that, given features f_i, f_j and a feature combination $F_i^{(j)}$, the accuracy of $F_i^{(j)} \cup \{f_i\}$ is not always higher than that of $F_i^{(j)} \cup \{f_j\}$ when f_i is more important than f_j. By taking into account the inhibition among features, we combine features via a heuristic algorithm. First, choose BOW as the best 1_ary feature combination, and combine each candidate feature from the rest with BOW to form 2_ary feature combinations. Then choose the one with the highest accuracy as

the best 2_ary feature combination, and filter out those features lower than the best 1_ary feature combination. Finally, repeat the above steps until the current candidate feature set is empty or all the feature combinations are no longer higher than the highest in the previous round.

Algorithm 1 gives the implement of the IIA method.

Algorithm 1 Importance-inhibition analysis algorithm

Input: F

Output: $F_i^{(j)}$

1) n features to form feature set F;

2) $F_1^* = F_1^{(1)} = \{\text{BOW}\}$;

3) $F = F - F_1^*$; $F' = \varnothing$;

4) For $i = 2$ to n

for $j = 1$ to $|F|$

$F_i^{(j)} = F_{i-1}^* \cup \{f^{(j)}\}$;

if the accuracy of $F_i^{(j)}$ is lower than that of F_{i-1}^* then $F' = F' \cup \{f^{(j)}\}$;

find the best combination F_i^* from $F_i^{(1)}$, $F_i^{(2)}$, …, $F_i^{(|F|)}$;

if the accuracy of F_i^* is lower than that of F_{i-1}^* then quit loop and output F_{i-1}^*;

$F = F - F'$;

5) Output $F_i^{(j)}$ is the final best combination.

The IIA method is on the basis of the $(k-1)$_ary feature combination to obtain the best k_ary one, so compared with the exhaustive method, it can significantly improve the efficiency of feature combination. In addition, since the IIA method takes into account the inhibition among features as well as the importance of individual features, compared with the IA method, it can more objectively depict the process of combining features and ensure a better performance of question classification.

3 Experimental Results and Analysis

3.1 Data set and evaluation

In our experiments, we use the Chinese questions set provided by IRSC lab of HIT (http://ir.hit.edu.cn), which contains 6 266 questions belonging to 6 categories and 77 classes.

The open and free available Liblinear-1.4 (http://www.csie.ntu.edu.tw/~cjlin/liblinear/) which is a linear classifier for data with millions of instances and features which is used to be the classifier. We use 10-fold cross validation on the total question set to evaluate the performance of the question classifications.

3.2 Combining features via IIA

According to the IIA method, we take BOW as the initial feature, and combine POS, NE, WSD, WSDm, R and P features gradually to form feature combinations, such as 2_ary, 3_ary, 4_ary and so on. The accuracies of individual features are presented in Fig. 2(a). Figs. 2(b) to (d) list all the accuracies of 2_ary, 3_ary and 4_ary feature combinations respectively, where Base1, Base2 and Base3 stand for the corresponding best 1_ary, 2_ary, 3_ary feature combinations.

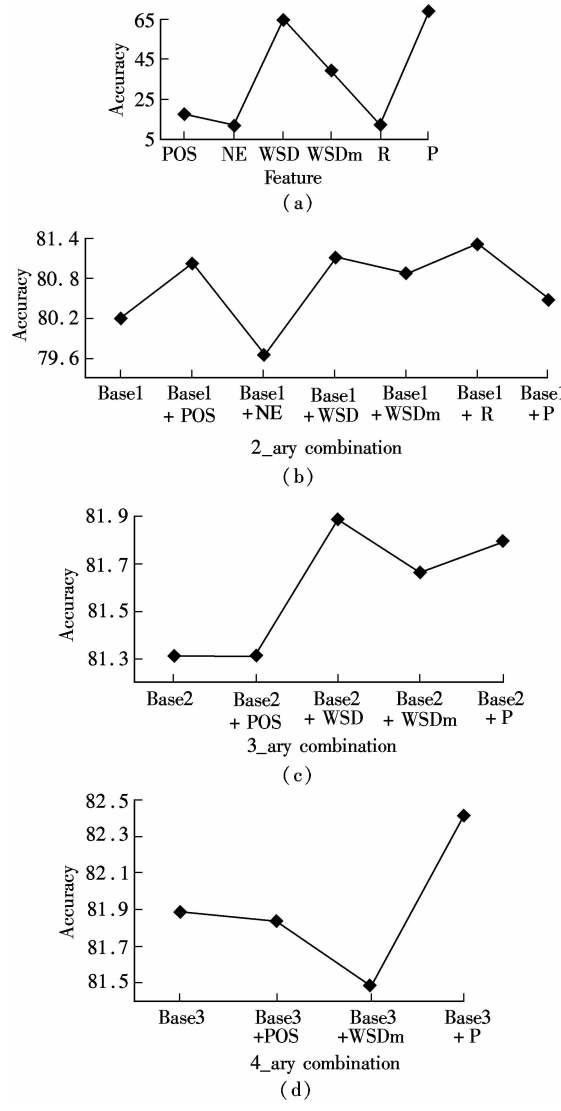

Fig. 2 Accuracies of n_ary feature combinations. (a) 1_ary; (b) 2_ary; (c) 3_ary; (d) 4_ary

In Fig. 2(b) and Fig. 2(c), the P feature has the highest classification accuracy among all the candidates, but the accuracies of Base1 + P and Base2 + P are not the highest in all the 2_ary and 3_ary feature combinations, respectively. In particular, the accuracy of Base1 + P is the last but one in all the 2_ary feature combinations.

In Fig. 2(b), the accuracy of Base1 + NE is lower than that of Base1, so NE is no longer considered in subsequent rounds. Similarly, in Fig. 2(d), the accuracies of Base3 + POS and Base3 + WSDm are both lower than that of Base3, so POS and WSDm are not considered in subsequent rounds. This is greatly convenient for filtering noise features.

In Fig. 2(c) and Fig. 2(d), the accuracies of Base1 + NE, Base3 + POS, Base3 + WSDm are lower than those

of Base1 and Base3, respectively. The reason is that R covers POS to a large extent in syntactic expression, and the difference between WSD and WSDm is very small. As a result, there exists the inhibition among features when they are in the same feature combination.

3.3 Performance comparison with IA

In order to verify the efficiency and effectiveness of IIA, we conduct performance comparison with IA. Tab. 2 shows the accuracies of the feature combinations via IIA and IA, respectively, where the "2_ary" column means 2_ary combinations, the "Base" row denotes the best ($n-1$)_ary combinations, "+POS" row means the feature combined with its baseline, the accuracy in bold means the maximum of n_ary combinations, and the one in bold with underline shows the maximum of all the combinations.

Tab. 2 Accuracies of feature combinations via IIA and IA %

Feature	IIA			IA					
	2_ary	3_ary	4_ary	2_ary	3_ary	4_ary	5_ary	6_ary	7_ary
Base	80.194 7	81.311 8	81.886 4	80.194 7	80.497 9	81.678 9	81.583 1	81.583 1	82.269 4
+POS	81.008 6	81.311 8	81.838 5				**81.583 1**		
+NE	79.668								**82.189 6**
+WSD	81.104 4	**81.886 4**			**81.678 9**				
+WSDm	80.880 9	81.662 9	81.487 4			**81.583 1**			
+R	**81.311 8**							**82.269 4**	
+P	80.497 9	81.790 6	**82.413**	**80.497 9**					

Fig. 3 conducts the comparison of average and maximum accuracies between IIA and IA, where the X axis denotes n_ary feature combinations, the Y axis denotes classification accuracies.

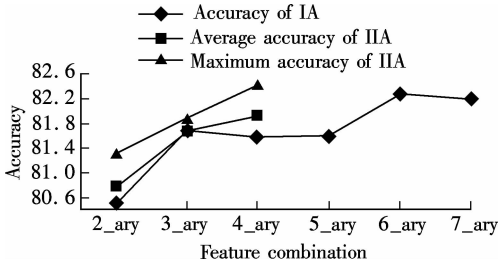

Fig. 3 Performance comparison between IIA and IA

From Fig. 3, we can see that IIA shows a gradual increase in average and maximum accuracies in all the feature combinations, while IA shows a slight decline in accuracy at the 4_ary and 7_ary ones. The reason is that IIA is based on the best previous feature combination to obtain the current one. In addition, IIA performs as well as IA in average accuracy at 3_ary feature combinations, and achieves a great improvement over IA in average and maximum accuracies at 2_ary and 4_ary feature combinations. In particular, IIA achieves 0.813 9% and 0.829 9% higher than IA in average and maximum accuracies at 4_ary feature combinations, so we can draw a conclusion that IIA performs significantly better than IA on the whole.

In order to further verify the efficiency and effectiveness of IIA, we conduct performance comparison with the exhaustive method. Experimental results show that the exhaustive method carries on 6 rounds for acquiring 63 feature combinations, while IIA does 3 rounds with 13 feature combinations gained. This demonstrates that IIA is much more efficient and feasible than the exhaustive method in practical applications. Furthermore, IIA gets the accuracy of 82.413% which is the highest one gained by the exhaustive method.

4 Conclusion

In this paper, we propose a new method called IIA to combine features via importance-inhibition analysis. The method takes into account the inhibition among various features as well as the importance of individual features. Experimental results on the Chinese question set show that the IIA method performs more effectively than the IA method on the whole, and achieves the same highest accuracy as the one gained by the exhaustive method.

The IIA method is a heuristic one in nature, and may be faced with the problem of a local optimum. In our further work, we will make great efforts to achieve more efficient and effective optimization for combining features.

Acknowlegement We would like to thank the IRSC laboratory of Harbin Institute of Technology for their free and available LTP platform.

References

[1] Zhang Z C, Zhang Y, Liu T, et al. Advances in open-domain question answering [J]. *Acta Electronica Sinica*, 2009, **37**(5): 1058 − 1069. (in Chinese)

[2] Li X, Roth D. Learning question classifiers[C]//*Proc of the* 19*th International Conference on Computational Linguistics*. Taipei, China, 2002: 1 − 7.

[3] Li X, Roth D. Learning question classifiers: the role of semantic information [J]. *Journal of Natural Language Engineering*, 2006, **12**(3): 229 − 250.

[4] Zhang D, Lee W. Question classification using support vector machines [C]//*Proc of the* 26*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada, 2003: 26 − 32.

[5] Huang Z H, Thint M, Qin Z C. Question classification using head words and their hypernyms[C]//*Proc of the* 2008 *Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii, USA, 2008: 927－936.

[6] Huang Z H, Thint M, Celikyilmaz A. Investigation of question classifier in question answering[C]//*Proc of the* 2009 *Conference on Empirical Methods in Natural Language Processing*. Singapore, 2009: 543－550.

[7] Li F T, Zhang X, Yuan J H, et al. Classifying what-type questions by head noun tagging[C]//*Proc of the* 22*nd International Conference on Computational Linguistics*. Manchester, UK, 2008: 481－488.

[8] Li X, Huang X J, Wu L D. Combined multiple classifiers based on TBL algorithm and their application in question classification [J]. *Journal of Computer Research and Development*, 2008, **45**(3): 535－541. (in Chinese)

[9] Sun J G, Cai D F, Lu D X, et al. HowNet based Chinese question automatic classification [J]. *Journal of Chinese Information Processing*, 2007, **21**(1): 90－95. (in Chinese)

[10] Zhang Z C, Zhang Y, Liu T, et al. Chinese question classification based on identification of cue words and extension of training set [J]. *Chinese High Technology Letters*, 2009, **19**(2): 111－118. (in Chinese)

基于重要性和抑制性分析的问句特征组合

杨思春[1,2] 高 超[3] 姚佳岷[2] 戴新宇[1] 陈家骏[1]

([1] 南京大学计算机软件新技术国家重点实验室, 南京 210093)

([2] 安徽工业大学计算机学院, 马鞍山 243032)

([3] 滁州学院计算机与信息工程学院, 滁州 239000)

摘要:针对基于机器学习的问题分类中问句特征的组合,提出了一种基于重要性和抑制性分析(importance-inhibition analysis, IIA)的特征组合方法. 该方法在组合问句特征时不仅考虑了单个特征本身的重要性,还考虑了待组合特征之间的抑制性. 在中文问题集上的实验结果表明,IIA 方法在所有的特征组合上都获得了平均精度和最高精度的提升,总体上比单纯基于重要性分析(importance analysis, IA)的特征组合方法要更加高效;同时,IIA 方法还获得了与穷举式特征组合方法同样的最高精度,进一步提升了当前中文问题分类的性能.

关键词:问答系统;问题分类;特征组合;重要性和抑制性分析

中图分类号:TP391