PHÂN LOẠI CÂU HỎI SỬ DỤNG NHIỀU LOẠI ĐẶC TRƯNG

Nguyễn Văn Tú[1], Lê Anh Cường[2], Nguyễn Hà Nam[3]
[1]Trường ĐH Tây Bắc, [2]Trường ĐH Tôn Đức Thắng, [3]Trường ĐH Công nghệ-ĐHQGHN

Tóm tắt

Phân loại câu hỏi là một thành phần quan trọng trong các hệ thống hỏi đáp tự động. Nhiệm vụ chính của phân loại câu hỏi là dự báo kiểu thực thể của câu trả lời của các câu hỏi viết bằng ngôn ngữ tự nhiên. Phân loại câu hỏi có thể được thực hiện bằng các tiếp cận khác nhau như: tiếp cận dựa trên luật, tiếp cận dựa trên học máy. Các đặc trưng khác nhau về từ vựng, cú pháp và ngữ nghĩa có thể được trích xuất tự động từ các câu hỏi để phục vụ việc phân loại. Trong nghiên cứu này chúng tôi kết hợp các đặc trưng về từ vựng, cú pháp, ngữ nghĩa trong phân loại câu hỏi. Chúng tôi đề xuất sử dụng mẫu câu hỏi (Question pattern) như là một đặc trưng mới để kết hợp với các đặc trưng khác trong phân loại câu hỏi. Chúng tôi cũng đề xuất sử dụng các tập đặc trưng khác nhau cho mỗi nhóm câu hỏi với các từ để hỏi khác nhau. Chúng tôi nhận thấy rằng khi sử dụng mẫu câu hỏi như là một đặc trưng và kết hợp với các đặc trưng từ vựng, cú pháp, ngữ nghĩa khác có thể cải thiện đáng kể độ chính xác của phân loại câu hỏi. Chúng tôi đã kiểm tra những đề xuất của mình bằng cách sử dụng bộ phân loại Support Vector Machine trên bộ dữ liệu TREC và đã đạt được độ chính xác phân loại câu hỏi cao hơn so với những nghiên cứu trước đó trên cùng nguyên tắc phân loại và tập dữ liệu.

Keywords: Question classification, question answering systems, question pattern, support vector machines.

Phương pháp

- Tiếp cận dựa trên học máy.
- Sử dụng bộ phân loại SVM với hàm nhân tuyến tính.
- Xây dựng các tập đặc trưng: các đặc trưng từ vựng, các đặc trưng cú pháp, các đặc trưng ngữ nghĩa.
- Trích rút, lựa chọn và kết hợp các đặc trưng tốt.

Kết quả

Kết quả phân loại câu hỏi sử dụng sự kết hợp của nhiều đặc trưng

Bảng 1: Độ chính xác của bộ phân loại khi sử dụng các đặc trưng U, B, Wh, WS, H, QE, QC, RW

Số lượng câu hỏi trong tập huấn luyện	1000	2000	3000	4000	5500
Coarse 2	90.20	91.20	92.00	92.60	94.20
Fine 2	79.00	85.40	86.60	88.00	90.40

Kết quả phân loại câu hỏi sử dụng sự kết hợp của nhiều đặc trưng

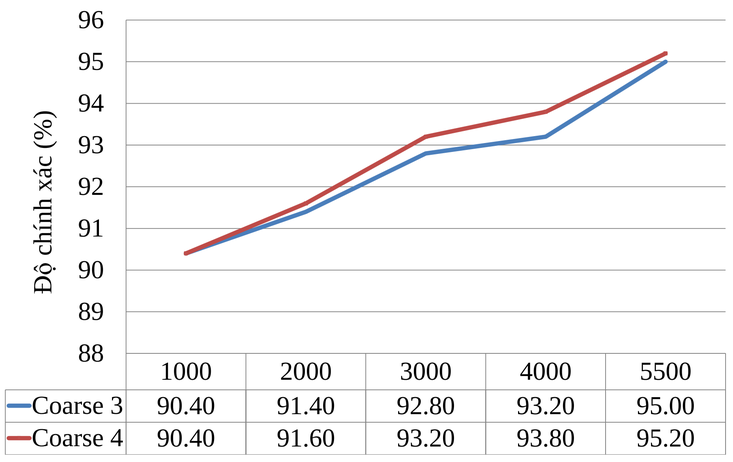
Bảng 2: Độ chính xác của bộ phân loại khi sử dụng các đặc trưng U, B, Wh, WS, H, QE, QC, RW, QP

Số lượng câu hỏi trong tập huấn luyện	1000	2000	3000	4000	5500
Coarse 3	90.40	91.40	92.80	93.20	95.00
Fine 3	79.20	86.00	87.00	88.60	91.00

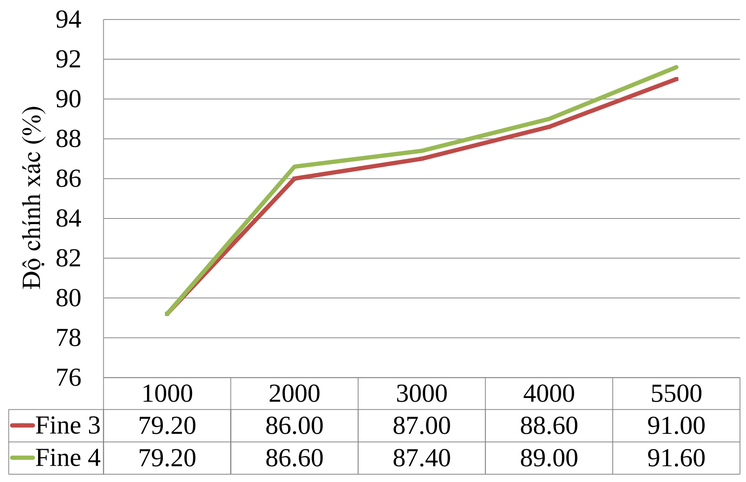
Kết quả phân loại câu hỏi sử dụng sự lựa chọn các tập đặc trưng khác nhau cho từng loại câu hỏi

Bảng 3: Độ chính xác khi sử dụng các tập đặc trưng khác nhau cho các câu hỏi với từ để hỏi khác nhau

Số lượng câu hỏi trong tập huấn luyện	1000	2000	3000	4000	5500
Coarse 4	90.40	91.60	93.20	93.80	95.20
Fine 4	79.20	86.60	87.40	89.00	91.60



Biểu đồ 1: So sánh độ chính xác của bộ phân loại lớp thô khi sử dụng nhiều đặc trưng (coarse 3) và khi sử dụng các nhóm đặc trưng theo question type (coarse 4)



Biểu đồ 2: So sánh độ chính xác của bộ phân loại lớp mịn khi sử dụng nhiều đặc trưng (fine 3) và khi sử dụng các nhóm đặc trưng theo Question type (fine 4)

Kết luận

Trong nghiên cứu này, chúng tôi đã trình bày một tiếp cận dựa trên học máy để phân loại câu hỏi. Để huấn luyện một thuật toán học, chúng tôi đã trích rút một tập các đặc trưng tốt từ từ vựng, cú pháp và ngữ nghĩa. Chúng tôi đề xuất sử dụng mẫu câu hỏi như là một đặc trưng mới để kết hợp với các đặc trưng từ vựng, cú pháp và ngữ nghĩa. Chúng tôi cũng đề xuất sử dụng các tập đặc trưng khác nhau cho việc phân loại các kiểu câu hỏi khác nhau. Các kết quả thực nghiệm đã chứng minh rằng các đề xuất của chúng tôi cho hiệu quả phân loại cao hơn so với các công trình nghiên cứu trước đó trên cùng nguyên tắc phân loại và tập dữ liệu. Chúng tôi cũng nhận thấy rằng trong phân loại câu hỏi thì các câu hỏi với từ để hỏi "What" thì khó phân loại hơn các câu hỏi khác. Vì vậy muốn tăng hiệu suất phân loại trong việc phân loại các câu hỏi thì cần phải cải thiện việc phân loại các câu hỏi với từ để hỏi "What".

Tài liệu tham khảo

[1] Phil Blunsom, Krystle Kocik, and James R. Curran, Question classification with log-linear Models, In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, 2006, pages 615–616.

[2] David A. Hull, Xerox TREC-8 question answering track report, In *In Voorhees and Harman*, 1999.

[3] Hardy, Yu-N Cheah, Question Classification Using Extreme Learning Machine on Semantic Features. *J. ICT Res. Appl.*, Vol. 7, No. 1, 2013, pages 36-58.

[4] Ulf Hermjakob, Eduard Hovy, and Chin yew Lin, Automated question answering in webclopedia - a demonstration, In *In Proceedings of ACL*-02, 2002.

[5] Zhiheng Huang, Marcus Thint, and Asli Celikyilmaz, Investigation of question classifier in question answering. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (EMNLP '09), 2009, pages 543–550.

[6] Zhiheng Huang, Marcus Thint, and Zengchang Qin, Question classification using head words and their hypernyms, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP '08)*, 2008, pages 927–936.

[7] A. Ittycheriah, M. Franz, W. J. Zhu, A. Ratnaparkhi, and R. J. Mammone, IBM's statistical question answering system. *In Proceedings of the 9th Text Retrieval Conference*, NIST, 2001.

[8] John Judge, Aoife Cahill, and Josef van Genabith, Questionbank: creating a corpus of parse-annotated questions, In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, 2006, pages 497–504.

[9]Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti, Enhanced answer type inference from questions using sequential models, In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, 2005, pages 315–322.

[10] Wendy G. Lehnert, A conceptual theory of question answering, In *Proceedings of the 5th international joint conference on Artificial intelligence*, Volume 1, 1977, pages 158–164.

[11] Xin Li and Dan Roth, Learning question classifiers, *In Proceedings of the 19th international conference on Computational linguistics*, COLING '02, 2002, pages 1–7.

[12] Xin Li and Dan Roth, Learning question classifiers: The role of semantic information, In *In Proc, International Conference on Computational Linguistics* (COLING), 2004, pages 556–562.

[13] Andreas Merkel and Dietrich Klakow, Improved methods of language model based question classification, In *In Proceedings of Interspeech Conference*, 2007.

[14] Donald Metzler and W. Bruce Croft, Analysis of statistical question classification for fact-based questions, *Inf. Retr*, 2005, pages 481–504.

[15] Dan Moldovan, Marius Pa¸sca, Sanda Harabagiu, and Mihai Surdeanu, Performance issues and error analysis in an open-domain question answering system, *ACM Trans*, Inf. Syst., 2003, pages 133–154.

[16] Yan Pan, Yong Tang, Luxin Lin, and Yemin Luo, Question classification with semantic tree kernel, *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 2008, pages 837–838.

[17] Slav Petrov and Dan Klein, Improved inference for unlexicalized parsing, In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 2007, pages 404–411.

[18] John Prager, Dragomir Radev, Eric Brown, and Anni Coden, The use of predictive annotation for question answering in trec8, In *In NIST Special Publication 500-246: The Eighth Text REtrieval Conference* (TREC 8), 1999, pages 399–411.

[19]Jo˜ao Silva, Lu´ısa Coheur, Ana Mendes, and Andreas Wichert, From symbolic to subsymbolic information in question classification, *Artificial Intelligence Review*, 2011, pages 137–154.

[20] Ellen M. Voorhees. Overview of the trec 2001 question answering track. In *In Proceedings of the Tenth Text REtrieval Conference (TREC)*, 2001, pages 42–51.

[21] Li Xin, HUANG Xuan-Jing, and WU Li-de, Question classification using multiple classifiers, In *Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network*, 2005.

[22] Dell Zhang and Wee Sun Lee, Question classification using support vector machines, In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '03, 2003, pages 26–32.