NHẬN DẠNG THỰC THỂ ĐỊNH DANH TRONG VĂN BẢN TIẾNG VIỆT

NGUYỄN LÊ TRUNG THÀNH
Trường Đại học Sư phạm – Đại học Huế
ĐT: 0902 615 658, Email: nguyenthanh224@gmail.com

Tóm tắt: Nhận dạng thực thể định danh là bài toán xác định lớp của các thực thể trong văn bản (thực thể chỉ tên người, tên tổ chức, tên địa điểm,..). Nhận dạng thực thể định danh là bài toán cơ bản trong nhiều vấn đề của xử lý ngôn ngữ tự nhiên như truy vấn thông tin, trích xuất thông tin, dịch máy, hệ thống hỏi đáp, tóm tắt văn bản tự động. Bài báo giới thiệu hệ thống nhận dạng thực thể định danh trong văn bản tiếng Việt dựa trên tập luật. Luật được xây dựng để tìm kiếm các mẫu qua quá trình so khớp. Các thực thể trong mẫu sau đó sẽ được phân loại vào từng lớp cụ thể dựa vào thông tin ngữ cảnh mà mẫu cung cấp. Kết quả thực nghiệm của hệ thống là tương đối khả quan với độ đo F đạt 80,64%.

Từ khóa: nhận dạng thực thể định danh, hệ thống nhận dạng dựa trên tập luật, xử lý ngôn ngữ tự nhiên, văn bản tiếng Việt.

1. GIỚI THIỆU

Nhận dạng thực thể định danh là bài toán cơ bản và quan trọng trong xử lý ngôn ngữ tự nhiên. Nhận dạng thực thể định danh bao gồm xác định và phân loại các thực thể trong văn bản vào các lớp gồm lớp Người, Tổ chức, Địa điểm và lớp Khác (các thực thể không thuộc ba lớp trên). Kết quả của quá trình nhận dạng thực thể định danh được sử dụng trong nhiều lĩnh vực như truy vấn thông tin, trích xuất thông tin, dịch máy, hệ thống hỏi đáp, tóm tắt văn bản.

Bài báo này giới thiệu hệ thống nhận dạng thực thể định danh tiếng Việt dựa trên tập luật. Phần 2 đề cập đến các nghiên cứu liên quan. Phần 3 mô tả cách xây dựng hệ thống nhận dạng thực thể định danh dựa trên tập luật và trình bày về tập luật mà tác giả xây dựng được. Thực nghiệm trên hệ thống và hiệu quả được đánh giá ở phần 4. Phần 5 trình bày kết luận và các hướng phát triển tiếp trong tương lai.

2. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Nhiều nghiên cứu về nhận dạng thực thể định danh được thực hiện với các cách tiếp cận khác nhau. Có thể phân chia làm hai cách tiếp cận chính: tiếp cận dựa trên tập luật và cách tiếp cận dựa vào các kỹ thuật học máy.

Với cách tiếp cận học máy, học có giám sát hiện đang là kỹ thuật chiếm ưu thế. Một số các kỹ thuật học có giám sát bao gồm mô hình Markov ẩn, mô hình entropy cực đại [1], máy vectơ hỗ trợ, và trường điều kiện ngẫu nhiên [5]. Đối với tiếng Việt, Tu và các cộng sự [6] xây dựng hệ thống sử dụng trường điều kiện ngẫu nhiên trong khi Tran và các cộng sự [8] sử dụng máy vectơ hỗ trợ để nhận dạng thực thể định danh.

Tạp chí Khoa học và Giáo dục, Trường Đại học Sư phạm Huế
ISSN 1859-1612, Số 02(42)/2017: tr. 109-115
Ngày nhận bài: 22/02/2017; Hoàn thành phản biện: 05/4/2017; Ngày nhận đăng: 10/4/2017

Bên cạnh học có giám sát, một kỹ thuật học bán giám sát thường được sử dụng để nhận dạng thực thể là bootstrapping. Kỹ thuật bootstrapping chỉ cần tập dữ liệu huấn luyện tương đối nhỏ là có thể bắt đầu quá trình học. Một trong những nghiên cứu sử dụng kỹ thuật bootstrapping có ảnh hưởng là của Riloff và Jones [10].

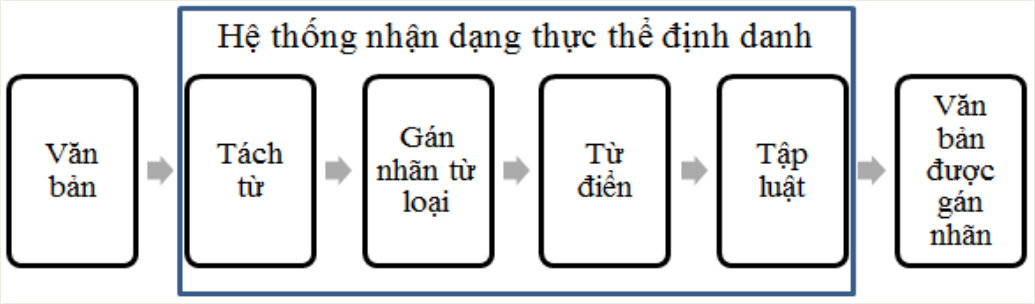
Với cách tiếp cận bằng tập luật, hệ thống sẽ nhận dạng các thực thể định danh thông qua các luật được thiết kế bởi con người. Các thực thể được nhận dạng bằng kỹ thuật so trùng mẫu dựa vào các đặc trưng như chữ viết thường, viết hoa, từ loại, từ đứng trước, từ đứng sau,... Với cách tiếp cận này, Appelt và các cộng sự [2] đã thiết kế hệ thống FASTUS nhận dạng thực thể định danh bằng biểu thức chính quy. Cao và các cộng sự [11] sử dụng các luật trong hệ thống VN KIM IE nhận biết và chú thích tự động cho các thực thể có tên trong trang web tiếng Việt.

Bên cạnh các hướng tiếp cận đã đề cập, một số hệ thống sử dụng hướng tiếp cận lai kết hợp tập luật và kỹ thuật học máy nhằm tận dụng ưu điểm của từng phương pháp. R. Sirhari và các cộng sự giới thiệu một hệ thống lai bằng cách kết hợp tập luật, mô hình Markov ẩn và entropy cực đại [9].

3. HỆ THỐNG NHẬN DẠNG THỰC THỂ ĐỊNH DANH

3.1 Kiến trúc hệ thống

Hệ thống nhận dạng thực thể định danh trên văn bản tiếng Việt được xây dựng như là một ứng dụng (application) trên GATE. GATE (General Architecture for Text Engineering) là kiến trúc tổng quát để phát triển các ứng dụng xử lý ngôn ngữ tự nhiên [3]. Kiến trúc của hệ thống được mô tả bởi hình 3.1. Hệ thống bao gồm bốn phần: bộ tách từ; bộ gán nhãn từ loại; bộ từ điển và tập các luật. Ban đầu, văn bản được tách từ và gán nhãn từ loại. Trong đó, bộ tách từ được sử dụng là vnTokenizer [4], bộ gán nhãn từ loại được sử dụng là vnTagger [7]. Các từ điển được xây dựng qua quá trình làm việc trên ngữ liệu tiếng Việt bằng cách liệt kê các thực thể chỉ người, địa điểm, tổ chức đã được nhận dạng cùng với các từ thường xuất hiện với các thực thể kể trên. Một từ thuộc văn bản nếu so khớp với từ thuộc từ điển thì GATE sẽ tự động gán nhãn chú thích Lookup trên từ đó. Thông tin về kiểu từ điển của nhãn Lookup kết hợp với các thông tin khác của các nhãn chú thích (annotation) khác trên từ như kiểu viết thường, viết hoa, từ loại, nội dung của từ,... sẽ làm cơ sở cho tập luật nhận dạng các thực thể định danh.



Hình 1. *Hệ thống nhận dạng thực thể định danh trong văn bản tiếng Việt*

3.2. Tập luật nhận dạng

Về hình thức, luật là một cặp *mẫu* / *cách thực thi* (*pattern* / *action*). Trong đó, mẫu thể hiện khuôn dạng của nhóm từ thông qua thông tin về nhãn chú thích trên các từ đó; cách thực thi là hành động thực hiện khi mẫu được tìm thấy qua quá trình so khớp. Ví dụ, một mẫu giúp nhận dạng tên công ty được thể hiện như sau:

(*tiền tố công ty*) (*loại hình công ty*) (*ứng viên tên công ty*)

--->

ứng viên tên công ty được gán nhãn "Organization" (tổ chức)

Một cụm từ được phân loại thuộc lớp tổ chức nếu so khớp với mẫu ở vế trái của luật trên. Ví dụ, cụm từ "*công ty TNHH Phú Quốc*" sẽ được phân loại thuộc lớp tổ chức do có từ bắt đầu bằng tiền tố công ty ("*công ty*"), tiếp theo là từ chỉ loại hình công ty ("*TNHH*") và cuối cùng là ứng viên tên công ty gồm từ có các chữ cái đầu viết hoa ("*Phú Quốc*"). "*Phú Quốc*" trong trường hợp này được nhận dạng là tên tổ chức.

Các luật được cụ thể hóa trên GATE bằng các luật JAPE (Java Annotation Pattern Engine). Với luật JAPE, người dùng có thể tạo mẫu bằng biểu thức chính quy trên nhãn và tạo các nhãn mới trên các mẫu được so khớp. Cặp *mẫu* / *cách thực thi* được thể hiện bằng *vế trái --> vế phải* trên JAPE. Ví dụ, mẫu nhận dạng công ty có thể được thể hiện như sau:

```
Rule: Corporation1
 (
      ({Lookup.majorType == corporation-prefix})
      ({Lookup.majorType == corporation-type})
      (CANDIDATE):name
):corp
-->
:name.Organization = {type = "Corporation", rule = "Corporation1"},
:corp.OrganizationWrap = {type = "Corporation", rule = "Corporation1"}
```

Trong đó, *corporation-prefix* thể hiện cụm từ chỉ tiền tố công ty, *corporation-type* thể hiện loại hình công ty, *CANDIDATE* là thành phần thay thế (macro) thể hiện cụm từ bắt đầu bằng chữ viết hoa – là ứng viên của tên công ty. Một cụm từ so khớp với mẫu ở vế trái sẽ được gán nhãn là "OrganizationWrap" và cụm từ ứng viên trong thành phần thay thế được gán nhãn "Organization".

Các luật được thực hiện một cách tuần tự. Nhãn chú thích được sinh ra bởi các luật thực hiện trước có thể được sử dụng như dữ liệu đầu vào cho các luật thực hiện sau. Ví dụ, nếu cụm từ "*tỉnh Thừa Thiên Huế*" đã được nhận dạng là tên địa phương thì sẽ giúp nhận dạng cụm từ "*UBND tỉnh Thừa Thiên Huế*" là tổ chức khi so khớp mẫu:

(*tiền tố tổ chức*) (*thực thể chỉ địa điểm*)

-->

Gán toàn bộ cụm từ (*tiền tố tổ chức*) (*thực thể chỉ địa điểm*) *là thực thể chỉ tổ chức*

Thứ tự thực hiện các luật của hệ thống lần lượt là nhận dạng địa điểm, tổ chức, tên người. Sau các bước này, một số luật được xây dựng để nhận dạng lại các thực thể dựa trên những thông tin về địa điểm, tổ chức, tên người thu được từ các bước trước đó.

Ví dụ về một luật dựa trên thông tin về địa điểm, tổ chức đã được nhận dạng trước đó để nhận dạng tên người

(tiền tố chức vụ) (thực thể chỉ tổ chức) (thực thể chỉ địa điểm) (ứng viên)

-->

Gán cụm từ (ứng viên) là thực tể chỉ tên người

Với luật trên thì cụm từ "*CEO Microsoft Việt Nam Vũ Minh Trí*" nếu có "*CEO*" được nhận dạng là chức vụ, "*Microsoft*" được nhận dạng là tổ chức, "*Việt Nam*" được nhận dạng là địa điểm thì cụm từ ứng viên "*Vũ Minh Trí*" sẽ được nhận dạng là tên người.

4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Tập dữ liệu

Dữ liệu được thu thập từ 200 bài báo của các trang báo điện tử *thanhnien.vn, tuoitre.vn* và *vnexpress.net*. Các tài liệu được chuẩn hóa để có cùng một loại mã (encoding) là UTF-8. Các ký hiệu không cần thiết phát sinh từ quá trình sao chép nội dung trang web như *,^ và các chú thích ảnh viết bằng tiếng Việt không dấu như "*tong-thong-Barack-Obama-doc-dien-van*" sẽ được loại bỏ trước khi văn bản được đưa vào tập dữ liệu chính thức.

Sau đó, dữ liệu được trộn lẫn và chia làm 2 phần một cách ngẫu nhiên: phần 1 (ký hiệu D1) gồm 150 bài báo, phần 2 (ký hiệu D2) gồm 50 bài báo. Tập dữ liệu D1 được sử dụng để tạo các từ điển và tập luật. Tập dữ liệu D2 được dùng để kiểm tra. Các thực thể định danh trong tập dữ liệu D2 được gán nhãn bằng phương pháp thủ công.

4.2. Độ đo

Hiệu quả hoạt động của hệ thống nhận dạng thực thể định danh được đánh giá qua các độ đo sau:

Độ chính xác P (Precision): $P = \frac{N1}{N2} \times 100\%$

Độ đầy đủ R (Recall): $R = \frac{N1}{N3} \times 100\%$

Độ đo F (F-score): $F = 2 \times \frac{P \times R}{P + R} \times 100\%$

Trong đó, N1 là số thực thể được nhận dạng chính xác bởi hệ thống, N2 là số thực thể được nhận dạng bởi hệ thống (có thể chính xác hoặc không), N3 là số thực thể thực tế.

4.3. Kết quả trên tập dữ liệu kiểm tra

Hệ thống nhận dạng thực thể định danh trên tập dữ liệu D2 gồm gồm 989 câu, 19846 từ. Kết quả nhận dạng sử dụng tập luật do tác giả xây dựng được thể hiện ở bảng 1.

Kết quả thu được trên tập dữ liệu kiểm tra khá khả quan với độ đo F trên tổng thể đạt 80,64%. Trong đó, độ đo F của các thực thể chỉ người là 81,20%; thực thể chỉ tổ chức là 68,51% và thực thể chỉ địa điểm là 84,85%.

Trong 3 loại thực thể, thực thể chỉ địa điểm được nhận dạng hiệu quả tốt hơn cả với độ chính xác P 82,44%, độ bao phủ R 87,41% và độ đo F 84,85%.

Bảng 1. *Kết quả nhận dạng thực thể định danh trên tập dữ liệu kiểm tra*

Loại	Số thực thể thực tế	Số thực thể được nhận dạng	Số thực thể nhận dạng đúng	Độ chính xác P (%)	Độ bao phủ R (%)	Độ đo F (%)
Người	324	309	257	83,17	79,32	81,20
Tổ chức	238	194	148	76,28	62,18	68,51
Địa điểm	564	598	493	82,44	87,41	84,85
Tất cả	1126	1101	898	81,56	79,75	80,64

Các thực thể địa điểm với đặc điểm thông thường gồm 2 đến 3 tiếng và được viết hoa chữ cái đầu tiên tạo nên sự thuận lợi cho việc nhận dạng. Trong khi đó, thực thể tổ chức có hiệu quả nhận dạng thấp nhất với độ đo F là 68,51% do sự phức tạp trong cấu tạo tên tổ chức như "*Hội Khoa học Phát triển Nguồn nhân lực và nhân tài Việt Nam*", "*Hiệp hội các trường CĐ, trung cấp kinh tế, kỹ thuật*". Bên cạnh đó, tên các tổ chức thường được viết trực tiếp mà không đi kèm với các tiền tố chỉ tổ chức cũng gây khó khăn cho việc nhận dạng. Ví dụ câu sau:

Ông Yuri vừa nâng cổ phần của mình ở Bank Rossiya lên 60%.

"*Bank Rossiya*" không được nhận dạng tên tổ chức do dấu hiệu nhận biết là "*cổ phần*" nằm ngoài ngữ cảnh nhận dạng. Hiệu quả tương đối thấp trong việc nhận dạng tên tổ chức ảnh hưởng đến việc nhận dạng tên người, đặc biệt tên người có liên quan đến tổ chức như thể hiện ở câu sau:

Chủ tịch Hiệp hội các trường CĐ, trung cấp kinh tế, kỹ thuật Hoàng Lâm vừa có chuyến thăm và làm việc với Đại học Huế.

Do "*Hiệp hội các trường CĐ, trung cấp kinh tế, kỹ thuật*" không được nhận dạng là tổ chức nên chủ tịch "*Hoàng Lâm*" cũng không được nhận dạng là tên người trong câu trên. Bên cạnh đó, việc nhận dạng tên người cũng gặp một số khó khăn do sự nhập nhằng giữa tên người và tên địa điểm như ví dụ sau:

Sinh viên Huế tham dự cuộc thi "Đường chạy nghị lực VNU will run" 2016.

Trong trường hợp này, "*Huế*" nếu được hiểu là sinh viên tên Huế hay sinh viên của (Đại học) Huế đều hợp lý. Chính vì sự nhập nhằng của tên người nên mặc dù có cấu trúc đơn giản nhưng trong một số trường hợp tên người rất khó để nhận ra. Hiệu quả nhận dạng tên người thể hiện qua độ đo F đạt 81,20%.

Hiệu quả nhận dạng chung của hệ thống hứa hẹn sẽ được cải tiến nếu dữ liệu được mở rộng đồng nghĩa với bộ từ điển và tập luật phong phú hơn. Bên cạnh đó, quá trình nhận dạng cần sử dụng nhiều hơn yếu tố ngữ cảnh. Phân giải đồng tham chiếu là một trong các giải pháp tận dụng yếu tố ngữ cảnh để nhận dạng. Thêm vào đó, có thể kết hợp với các phương pháp học máy để tìm ra các ứng viên tiềm năng cho các thực thể định danh.

5. KẾT LUẬN

Bài báo trình bày vấn đề nhận dạng thực thể định danh. Hệ thống nhận dạng thực thể định danh trong văn bản tiếng Việt được thiết kế trên nền tảng của khung làm việc GATE với tập luật nhận dạng được xây dựng bởi nhóm tác giả. Hệ thống được thử nghiệm trên tập dữ liệu 50 bài báo trực tuyến. Kết quả thu được khá khả quan với độ đo F trên tổng thể đạt 80,64%. Trong đó, độ đo F của các thực thể chỉ người là 81,20%; thực thể chỉ tổ chức là 68,51% và thực thể chỉ địa điểm là 84,85%. Kết quả thu được phần nào khẳng định sự hiệu quả của hệ thống nhận dạng các thực thể định danh dựa trên tập luật. Tuy nhiên, kết quả thu được vẫn còn khiêm tốn, hệ thống còn có thể tiếp tục phát triển theo các hướng: mở rộng kho ngữ liệu huấn luyện, từ đó phát hiện được nhiều mẫu hơn để làm phong phú thêm tập luật nhận dạng; mở rộng từ điển nhờ tận dụng sự phong phú của kho ngữ liệu. Có thể phát triển các từ điển một cách tự động nhờ vào các từ đồng nghĩa, từ điển WordNet; thực hiện phân giải đồng tham chiếu trên các cụm danh từ để hạn chế sự nhập nhằng giữa các thực thể; kết hợp với các phương pháp học máy để tìm ra các ứng viên của các thực thể. Các ứng viên sẽ được kiểm tra lại bằng tập luật trước khi được gán nhãn. Trong tương lai hệ thống sẽ tiếp tục được nghiên cứu và phát triển để đạt độ chính xác tốt hơn.

TÀI LIỆU THAM KHẢO

[1] D. Borthwick, Andrew; Sterling, J.; Agichtein, E.; Grishman, R. (1998). NYU: *Description of the MENE Named Entity System as used in MUC-7*. In Proc. Seventh Message Understanding Conference.

[2] D. Appelt, and et. al., (1993). *FASTUS: A finite state processor for information extraction from real-world text*. Proceedings of IJCAI.

[3] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan (2002). G*ATE, A Framework and Graphical Development Environment for Robust NLP Tools and Application*. Proceedings of ACL'02. Philadelphia.

[4] Hong-Phuong Le, Minh-Huyen Thi Nguyen, Azim Roussanaly, and Tuong-Vinh Ho (2008). *A Hybrid Approach to Word Segmentation of Vietnamese Texts*. Language and Automata Theory and Applications, page 240.

[5] McCallum, Andrew; Li, W. (2003). *Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons*. In Proc. Conference on Computational Natural Language Learning.

[6] Nguyen Cam Tu, Tran Thi Oanh, Phan Xuan Hieu, and Ha Quang Thuy (2005). *Named entity recognition in Vietnamese free-text and web documents using conditional random fields*. In Conference on Some Selection Problems of Information Technology and Telecommunication

[7] Phuong Le-Hong, Azim Roussanaly, T. M. Huyen Nguyen, Mathias Rossignol (2010). *An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts*. Traitement Automatique des Langues Naturelles.

[8] Q. Tri Tran, T.X. Thao Pham, Q. Hung Ngo, Dien Dinh, and Nigel Collier. (2007). *Named entity recognition in Vietnamese documents*. Progress in Informatics, 5:14–17

[9] R. Sirhari, C. Niu, W. Li. (2000). *A Hybrid Approach for Named Entity and Sub-Type Tagging*. In Proceedings of the sixth conference on Applied natural language processing, ACM

[10] Riloff, E. and Jones, R. (1999). *Learning Dictionaries for Information Extraction by MultiLevel Bootstrapping*. In Proceedings of the AAAI Conference on Artificial Intelligence, Orlando, Florida, pages 474–479. JOHN WILEY & SONS LTD.

[11] T. Cao (2007). *Automatic Extraction of Vietnamese Named Entities on the Web*. New Generation Computing, Springer.

Title: NAMED ENTITY RECOGNITION IN VIETNAMESE DOCUMENTS

Abstract: Named Entity Recognition (NER) is the process of classifying different entity types (e.g person, organization, location, etc.) in documents. NER is considered to be crucial in many natural language processing tasks such as information retrieval, information extraction, machine translation, question answering system, automatic text summarization. This paper presents a NER rule-based system which is applied to Vietnamese documents. Rules are created and used to find patterns through matching process. Entities in matched pattern are classified into specific categories based on its contextual information. The experimental result with an overall F-score of 80,64% shows that this system achieves significant accuracy.

Keywords: Named Entity Recognition (NER), rule-based system, natural language processing, Vietnamese documents