(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0048697**

(51)[2020.01] **G01L 15/00**; G06F 3/00; G06F 17/00 (13) **B**

(21) 1-2021-04220 (22) 09/07/2021
(45) 25/07/2025 448 (43) 27/09/2021 402A
(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VIETTEL) (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội
(72) Đỗ Văn Hải (VN); Lê Nhật Minh (VN); Nguyễn Tùng Lâm (VN); Lê Quang Trung (VN); Nguyễn Tiến Thành (VN); Lê Đăng Linh (VN); Đặng Đình Sơn (VN); Nguyễn Thị Ngọc Ánh (VN); Phạm Minh Khang (VN); Nguyễn Ngọc Dũng (VN); Trần Mạnh Quân (VN); Nguyễn Mạnh Quý (VN).
(74) Công ty TNHH Tư vấn Quốc Dân (NACILAW)

(54) PHƯƠNG PHÁP PHÂN TÁCH, NHẬN DẠNG TIẾNG NÓI

(21) 1-2021-04220

(57) Sáng chế đề xuất phương pháp phân tách, nhận dạng tiếng nói. Sáng chế khắc phục được nhược điểm của các phương pháp hiện có bằng cách đề xuất phân tách và nhận dạng tiếng nói tự động giúp người quản lý có thể biết được điện thoại viên và khách hàng của mình đang nói những gì. Từ đó biết được mong muốn, băn khoăn của khách hàng cũng như các điện thoại viên của mình có tư vấn chính xác, đúng mực hay không một cách nhanh chóng và khách quan. Ngoài ra hệ thống được cập nhật liên tục dựa trên cơ chế huấn luyện bán giám sát, tức hệ thống có thể tự học từ dữ liệu thực tế trong quá trình vận hành, từ đó giúp độ chính xác của hệ thống càng ngày càng được cải thiện.

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp phân tách, nhận dạng tiếng nói. Cụ thể, sáng chế đề cập đến phương pháp phân tách, nhận dạng tiếng nói của điện thoại viên, khách hàng và huấn luyện bán giám sát trong tổng đài chăm sóc khách hàng, giúp tăng cường cho việc giám sát tự động các cuộc gọi chăm sóc khách hàng.

Tình trạng kỹ thuật của sáng chế

Ngày nay, số lượng các cuộc gọi chăm sóc khách hàng tăng lên nhanh chóng trong rất nhiều lĩnh vực như viễn thông, tài chính, điện lực, bán lẻ,... Do đó, làm sao để biết được mong muốn, băn khoăn của khách hàng cũng như các điện thoại viên của mình có tư vấn chính xác, đúng mực hay không là một nhu cầu cấp thiết đối với người quản lý. Việc này có thể thực hiện thủ công bằng cách sử dụng người giám sát nghe ngẫu nhiên một số cuộc gọi. Tuy nhiên phương pháp này tốn kém về nhân lực, chậm trễ về thời gian trong khi thông tin thu được lại phụ thuộc vào chủ quan của người giám sát. Do vậy cần thiết có một phương pháp để có thể tự động phân biệt và nhận dạng được tiếng nói của điện thoại viên và khách hàng trong các cuộc gọi chăm sóc khách hàng. Ngoài ra cũng cần có một phương pháp huấn luyện tự động để giúp hệ thống ngày càng nhận dạng chính xác hơn khi đưa vào sử dụng.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế này là đưa ra phương pháp phân tách, nhận dạng tiếng nói của điện thoại viên, khách hàng và huấn luyện bán giám sát trong tổng đài chăm sóc khách hàng nhằm tự động hóa quá trình giám sát các cuộc gọi chăm sóc khách hàng.

Cụ thể, sáng chế đề xuất phương pháp phân tách, nhận dạng tiếng nói được đề cập trong sáng chế, bao gồm:

Bước 1: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài để phân tách; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như ổ đĩa cứng, băng từ, ... hoặc thông qua các kết nối mạng dữ liệu, mỗi một tệp ứng với một cuộc gọi tổng đài; ta có thể lấy tệp tiếng nói trực tiếp trên thiết bị lưu trữ của người dùng hoặc sử dụng các giao thức truyền tệp qua mạng dữ liệu như FTP để lấy được các tín hiệu tiếng nói;

Bước 2: phân tách và gán nhãn văn bản cho các tệp tiếng nói; tại bước này, đưa các tệp tiếng nói ở bước 1 lên hệ thống gán nhãn để người gán nhãn nghe, phân tách và gán nhãn văn bản cho phần nói của điện thoại viên và khách hàng; đầu ra của bước này là các tập tiếng nói đã được phân loại và gán nhãn riêng biệt thành tập tiếng nói của điện thoại viên và tập tiếng nói của khách hàng;

Bước 3: tạo tập huấn luyện và kiểm thử; theo đó, khi dữ liệu tiếng nói được gán nhãn trong tệp của điện thoại viên và tệp của khách hàng ở bước 2 đều $\geq H_{label\_min}$ giờ dữ liệu, trong đó $H_{label\_min} \geq 10$ giờ nhằm đảm bảo tập dữ liệu đủ lớn; người quản trị quyết định lựa chọn một số tệp tiếng nói đã được gán nhãn ở bước 2 để tạo tập huấn luyện, các tệp còn lại được sử dụng để tạo tập kiểm thử với yêu cầu kích thước tập kiểm thử cần lớn hơn H_{test_min} giờ dữ liệu, trong đó $H_{test\_min} \geq 2$ giờ nhằm đảm bảo tập kiểm thử đủ lớn và tin cậy;

Bước 4: xây dựng hai mô hình ngôn ngữ; LM_a cho điện thoại viên và LM_b cho khách hàng dựa trên tập dữ liệu huấn luyện được tạo ở bước 3 nhằm lưu trữ những đặc điểm về ngôn ngữ nói như các cụm từ thường xuyên nói của điện thoại viên và khách hàng từ đó để phân biệt được câu nói của điện thoại viên hay khách hàng ở các bước sau; các mô hình ngôn ngữ có thể được sử dụng như là n-gram hay các mô hình dựa trên mạng nơ ron nhân tạo được tạo tự động bằng các phần mềm chuyên dụng;

Bước 5: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài cần phân tách, nhận dạng tự động; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như ổ đĩa cứng, băng từ, … hoặc thông qua các kết nối mạng dữ liệu, mỗi một tệp ứng với một cuộc gọi tổng đài; ta có thể lấy tệp tiếng nói trực tiếp trên thiết bị lưu trữ của người dùng hoặc sử dụng các giao thức truyền tệp qua mạng dữ liệu như FTP để lấy được các tín hiệu tiếng nói;

Bước 6: tự động cắt tệp tiếng nói thành các đoạn nhỏ; với mỗi tệp tiếng nói thu được ở bước 5, tiếng nói được tự động cắt thành các đoạn dựa theo các đặc tính về tín hiệu; ta có thể dựa vào các phương pháp phổ biến như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể dựa trên các hệ thống nhận dạng tiếng nói;

Bước 7: trích chọn các véc tơ đặc trưng người nói; tất cả các các đoạn tiếng nói thu được ở bước 6 được trích chọn véc tơ đặc trưng người nói bằng cách sử dụng một mạng trích chọn đặc trưng được huấn luyện trước như mạng nơ rơn học sâu (DNN), với mỗi đoạn tiếng nói sẽ thu được một véc tơ đặc trưng người nói tương ứng;

Bước 8: phân cụm các đoạn tiếng nói; với mỗi tệp tiếng nói, phân cụm các đoạn tiếng nói ở bước 6 thành hai cụm C_1 và C_2 dựa trên các véc tơ đặc trưng người nói được trích xuất ở bước 7;

Bước 9: chuyển đổi tiếng nói sang văn bản; tất cả các đoạn tiếng nói ở bước 6 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói, với mỗi đoạn tiếng nói thu được một văn bản tương ứng và một chỉ số độ tin cậy nhận dạng DTC có giá trị từ 0 đến 1;

Bước 10: lựa chọn đoạn tiếng nói thỏa mãn điều kiện làm căn cứ phân loại; với mỗi một tệp tiếng nói, lựa chọn đoạn tiếng nói trong bước 9 thỏa mãn điều kiện: có độ tin cậy DTC $\geq \alpha$, trong đó $0,5 \leq \alpha \leq 0,95$ nhằm loại bỏ những đoạn tiếng nói có độ tin cậy quá thấp thường là những đoạn tiếng nói có chất lượng quá kém hoặc môi trường quá nhiễu ảnh hưởng đến chất lượng hệ thống phân loại; nếu không lựa chọn được đoạn tiếng nói nào thỏa mãn, bỏ qua tệp này và chuyển sang tệp tiếng nói mới;

Bước 11: phân loại các đoạn tiếng nói của điện thoại viên và khách hàng; với các đoạn tiếng nói được lựa chọn ở bước 10 được chia thành hai cụm ở bước 8, tính $w = \frac{PPL_{a1} * PPL_{b2}}{PPL_{a2} * PPL_{b1}}$, trong đó PPL_{a1}, PPL_{a2}, PPL_{b1}, PPL_{b2} là chỉ số độ hỗn loạn (perplexity) được cho bởi các mô hình ngôn ngữ LM_a, LM_b ở bước 4 tính với tập dữ liệu văn bản của các đoạn tiếng nói được lựa chọn ở bước 10; PPL_{a1}, PPL_{b1} được tính ứng với các đoạn trong cụm C_1; PPL_{a2}, PPL_{b2} ứng với các đoạn trong cụm C_2; nếu $w \leq \theta$, toàn bộ các đoạn tiếng nói trong cụm C_1 được xác định là điện thoại viên, toàn bộ các đoạn tiếng nói trong cụm C_2 được xác định là khách hàng và ngược lại nếu $w > \theta$, toàn bộ các đoạn tiếng nói trong cụm C_2 được xác định là điện thoại viên, toàn bộ các đoạn tiếng nói trong cụm C_1 được xác định là khách hàng; ngưỡng θ có giá trị trong dải từ 0,5 đến 2,0; sau bước này ta đã phân tách và nhận dạng được tiếng nói của điện thoại viên và khách hàng, nếu có nhu cầu về huấn luyện bán giám sát để nâng cao chất lượng hệ thống thực hiện tiếp bước 12, nếu không thì dừng lại;

Bước 12: lựa chọn đoạn tiếng nói thỏa mãn điều kiện để đưa vào tập huấn luyện bán giám sát; lựa chọn đoạn tiếng nói trong bước 9 thỏa mãn điều kiện: có độ tin cậy DTC $\geq \beta$, trong đó $0,8 \leq \beta \leq 0,99$ nhằm lựa chọn những đoạn tiếng nói có độ tin cậy nhận dạng rất cao để đưa vào tập dữ liệu huấn luyện bán giám sát; mỗi đoạn tiếng nói đều đã được gán nhãn là điện thoại viên hay khách hàng từ bước 11;

Bước 13: lựa chọn thời điểm cập nhật mô hình ngôn ngữ; là khi dữ liệu huấn luyện trong tập bán giám sát lớn hơn một ngưỡng H_{semi_min} giờ dữ liệu và khi có quyết định của người quản trị, trong đó $H_{semi\_min} \geq 10$ giờ với mục đích dữ liệu huấn luyện bán giám sát đủ lớn và tin cậy;

Bước 14: xây dựng mô hình ngôn ngữ dựa trên dữ liệu bán giám sát; tại bước này sử dụng dữ liệu trong tập bán giám sát để xây dựng hai mô hình ngôn ngữ LM_{a_semi} với dữ liệu của điện thoại viên và LM_{b_semi} với dữ liệu của khách hàng; sau đó kết hợp với hai mô hình ngôn ngữ LM_a, LM_b ở bước 4 để tạo ra hai mô hình ngôn ngữ $LM_{a'}$, $LM_{b'}$ với hệ số kết hợp k, trong đó $0,8 \geq k \geq 0,1$;

Bước 15: cập nhật mô hình ngôn ngữ; tính chỉ số $w_0 = \frac{PPL_{a1} * PPL_{b2}}{PPL_{a2} * PPL_{b1}}$, trong đó PPL_{a1}, PPL_{a2}, PPL_{b1}, PPL_{b2} là chỉ số độ hỗn loạn được cho bởi các mô hình ngôn ngữ LM_a, LM_b ở bước 4 tính với các tập kiểm thử ở bước 3; PPL_{a1}, PPL_{b1} được tính ứng với tập kiểm thử gồm các đoạn tiếng nói của điện thoại viên; PPL_{a2}, PPL_{b2} được tính ứng với tập kiểm thử gồm các đoạn tiếng nói của khách hàng; sau đó tính chỉ số $w_1 = \frac{PPL_{a'1} * PPL_{b'2}}{PPL_{a'2} * PPL_{b'1}}$ tương tự như w_0 bằng cách thay thế hai mô hình ngôn ngữ ở bước 4 bằng $LM_{a'}$ và $LM_{b'}$ ở bước 14; nếu $w_0 > q * w_1$ thì cập nhật LM_a bằng $LM_{a'}$, LM_b bằng $LM_{b'}$; trong đó $q \geq 1,0$.

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây, cụ thể, phương pháp phân tách, nhận dạng tiếng nói của điện thoại viên, khách hàng và huấn luyện bán giám sát trong tổng đài chăm sóc khách hàng bao gồm các bước:

- Bước 1: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài để phân tách;
- Bước 2: phân tách và gán nhãn văn bản cho các tệp tiếng nói;
- Bước 3: tạo tập huấn luyện và kiểm thử;
- Bước 4: xây dựng hai mô hình ngôn ngữ;
- Bước 5: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài cần phân tách, nhận dạng tự động;
- Bước 6: tự động cắt tệp tiếng nói thành các đoạn nhỏ;
- Bước 7: trích chọn các véc tơ đặc trưng người nói;
- Bước 8: phân cụm các đoạn tiếng nói;
- Bước 9: chuyển đổi tiếng nói sang văn bản;

- Bước 10: lựa chọn đoạn tiếng nói thỏa mãn điều kiện làm căn cứ phân loại;

- Bước 11: phân loại các đoạn tiếng nói của điện thoại viên và khách hàng;

- Bước 12: lựa chọn đoạn tiếng nói thỏa mãn điều kiện để đưa vào tập huấn luyện bán giám sát;

- Bước 13: lựa chọn thời điểm cập nhật mô hình ngôn ngữ;

- Bước 14: xây dựng mô hình ngôn ngữ dựa trên dữ liệu bán giám sát;

- Bước 15: cập nhật mô hình ngôn ngữ.

Nội dung cụ thể của các bước này như sau:

Bước 1: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài để phân tách; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như ổ đĩa cứng, băng từ, … hoặc thông qua các kết nối mạng dữ liệu, mỗi một tệp ứng với một cuộc gọi tổng đài; ta có thể lấy tệp tiếng nói trực tiếp trên thiết bị lưu trữ của người dùng hoặc sử dụng các giao thức truyền tệp qua mạng dữ liệu như FTP để lấy được các tín hiệu tiếng nói.

Bước 2: phân tách và gán nhãn văn bản cho các tệp tiếng nói; tại bước này, đưa các tệp tiếng nói ở bước 1 lên hệ thống gán nhãn để người gán nhãn nghe, phân tách và gán nhãn văn bản cho phần nói của điện thoại viên và khách hàng; đầu ra của bước này là các tập tiếng nói đã được phân loại và gán nhãn riêng biệt thành tập tiếng nói của điện thoại viên và tập tiếng nói của khách hàng.

Bước 3: tạo tập huấn luyện và kiểm thử; theo đó, khi dữ liệu tiếng nói được gán nhãn trong tệp của điện thoại viên và tệp của khách hàng ở bước 2 đều $\geq H_{label\_min}$ giờ dữ liệu, trong đó $H_{label\_min} \geq 10$ giờ nhằm đảm bảo tập dữ liệu đủ lớn; người quản trị quyết định lựa chọn một số tệp tiếng nói đã được gán nhãn ở bước 2 để tạo tập huấn luyện, các tệp còn lại được sử dụng để tạo tập kiểm thử với yêu cầu kích thước tập kiểm thử cần lớn hơn H_{test_min} giờ dữ liệu, trong đó $H_{test\_min} \geq 2$ giờ nhằm đảm bảo tập kiểm thử đủ lớn và tin cậy.

Bước 4: xây dựng hai mô hình ngôn ngữ; LM_a cho điện thoại viên và LM_b cho khách hàng dựa trên tập dữ liệu huấn luyện được tạo ở bước 3 nhằm lưu trữ những đặc điểm về ngôn ngữ nói như các cụm từ thường xuyên nói của điện thoại viên và khách hàng từ đó để phân biệt được câu nói của điện thoại viên hay khách hàng ở các bước sau; các mô hình ngôn ngữ có thể được sử dụng như là n-gram hay các mô hình dựa trên mạng nơ ron nhân tạo được tạo tự động bằng các phần mềm chuyên dụng.

Bước 5: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài cần phân tách, nhận dạng tự động; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như ổ đĩa cứng, băng từ, … hoặc thông qua các kết nối mạng dữ liệu, mỗi một tệp ứng với một cuộc gọi tổng đài; ta có thể lấy tệp tiếng nói trực tiếp trên thiết bị lưu trữ của người dùng hoặc sử dụng các giao thức truyền tệp qua mạng dữ liệu như FTP để lấy được các tín hiệu tiếng nói.

Bước 6: tự động cắt tệp tiếng nói thành các đoạn nhỏ; với mỗi tệp tiếng nói thu được ở bước 5, tiếng nói được tự động cắt thành các đoạn nhỏ dựa theo các đặc tính về tín hiệu; ta có thể dựa vào các phương pháp phổ biến như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể dựa trên các hệ thống nhận dạng tiếng nói.

Bước 7: trích chọn các véc tơ đặc trưng người nói; tất cả các các đoạn tiếng nói thu được ở bước 6 được trích chọn véc tơ đặc trưng người nói bằng cách sử dụng một mạng trích chọn đặc trưng được huấn luyện trước như mạng nơ rơn học sâu (DNN), với mỗi đoạn tiếng nói sẽ thu được một véc tơ đặc trưng người nói tương ứng.

Bước 8: phân cụm các đoạn tiếng nói; với mỗi tệp tiếng nói, phân cụm các đoạn tiếng nói ở bước 6 thành 2 cụm C_1 và C_2 dựa trên các véc tơ đặc trưng người nói được trích xuất ở bước 7.

Bước 9: chuyển đổi tiếng nói sang văn bản; tất cả các đoạn tiếng nói ở bước 6 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói, với mỗi đoạn tiếng nói thu được một văn bản tương ứng và một chỉ số độ tin cậy nhận dạng DTC có giá trị từ 0 đến 1.

Bước 10: lựa chọn đoạn tiếng nói thỏa mãn điều kiện làm căn cứ phân loại; với mỗi một tệp tiếng nói, lựa chọn đoạn tiếng nói trong bước 9 thỏa mãn điều kiện: có độ tin cậy DTC $\geq \alpha$, trong đó $0,5 \leq \alpha \leq 0,95$ nhằm loại bỏ những đoạn tiếng nói có độ tin cậy quá thấp thường là những đoạn tiếng nói có chất lượng quá kém hoặc môi trường quá nhiễu ảnh hưởng đến chất lượng hệ thống phân loại; nếu không lựa chọn được đoạn tiếng nói nào thỏa mãn, bỏ qua tệp này và chuyển sang tệp tiếng nói mới.

Bước 11: phân loại các đoạn tiếng nói của điện thoại viên và khách hàng; với các đoạn tiếng nói được lựa chọn ở bước 10 được chia thành hai cụm ở bước 8, tính $w = \frac{PPL_{a1} * PPL_{b2}}{PPL_{a2} * PPL_{b1}}$, trong đó PPL_{a1}, PPL_{a2}, PPL_{b1}, PPL_{b2} là chỉ số độ hỗn loạn (perplexity) được cho bởi các mô hình ngôn ngữ LM_a, LM_b ở bước 4 tính với tập dữ liệu văn bản của

các đoạn tiếng nói được lựa chọn ở bước 10; PPL_{a1}, PPL_{b1} được tính ứng với các đoạn trong cụm C_1; PPL_{a2}, PPL_{b2} ứng với các đoạn trong cụm C_2; nếu $w \leq \theta$, toàn bộ các đoạn tiếng nói trong cụm C_1 được xác định là điện thoại viên, toàn bộ các đoạn tiếng nói trong cụm C_2 được xác định là khách hàng và ngược lại nếu $w > \theta$, toàn bộ các đoạn tiếng nói trong cụm C_2 được xác định là điện thoại viên, toàn bộ các đoạn tiếng nói trong cụm C_1 được xác định là khách hàng; ngưỡng θ có giá trị trong dải từ 0,5 đến 2,0; sau bước này ta đã phân tách và nhận dạng được tiếng nói của điện thoại viên và khách hàng, nếu có nhu cầu về huấn luyện bán giám sát để nâng cao chất lượng hệ thống thực hiện tiếp bước 12, nếu không thì dừng lại.

Bước 12: lựa chọn đoạn tiếng nói thỏa mãn điều kiện để đưa vào tập huấn luyện bán giám sát; lựa chọn đoạn tiếng nói trong bước 9 thỏa mãn điều kiện: có độ tin cậy $DTC \geq \beta$, trong đó $0,8 \leq \beta \leq 0,99$ nhằm lựa chọn những đoạn tiếng nói có độ tin cậy nhận dạng rất cao để đưa vào tập dữ liệu huấn luyện bán giám sát; mỗi đoạn tiếng nói đều đã được gán nhãn là điện thoại viên hay khách hàng từ bước 11.

Bước 13: lựa chọn thời điểm cập nhật mô hình ngôn ngữ; là khi dữ liệu huấn luyện trong tập bán giám sát lớn hơn một ngưỡng H_{semi_min} giờ dữ liệu và khi có quyết định của người quản trị, trong đó $H_{semi\_min} \geq 10$ giờ với mục đích dữ liệu huấn luyện bán giám sát đủ lớn và tin cậy.

Bước 14: xây dựng mô hình ngôn ngữ dựa trên dữ liệu bán giám sát; tại bước này sử dụng dữ liệu trong tập bán giám sát để xây dựng hai mô hình ngôn ngữ LM_{a_semi} với dữ liệu của điện thoại viên và LM_{b_semi} với dữ liệu của khách hàng; sau đó kết hợp với hai mô hình ngôn ngữ LM_a, LM_b ở bước 4 để tạo ra hai mô hình ngôn ngữ $LM_{a'}$, $LM_{b'}$ với hệ số kết hợp k, trong đó $0,8 \geq k \geq 0,1$.

Bước 15: cập nhật mô hình ngôn ngữ; tính chỉ số $w_0 = \frac{PPL_{a1} * PPL_{b2}}{PPL_{a2} * PPL_{b1}}$, trong đó PPL_{a1}, PPL_{a2}, PPL_{b1}, PPL_{b2} là chỉ số độ hỗn loạn được cho bởi các mô hình ngôn ngữ LM_a, LM_b ở bước 4 tính với các tập kiểm thử ở bước 3; PPL_{a1}, PPL_{b1} được tính ứng với tập kiểm thử gồm các đoạn tiếng nói của điện thoại viên; PPL_{a2}, PPL_{b2} được tính ứng với tập kiểm thử gồm các đoạn tiếng nói của khách hàng; sau đó tính chỉ số $w_1 = \frac{PPL_{a'1} * PPL_{b'2}}{PPL_{a'2} * PPL_{b'1}}$ tương tự như w_0 bằng cách thay thế hai mô hình ngôn ngữ ở bước 4 bằng $LM_{a'}$ và $LM_{b'}$ ở bước 14; nếu $w_0 > q * w_1$ thì cập nhật LM_a bằng $LM_{a'}$, LM_b bằng $LM_{b'}$; trong đó $q \geq 1,0$.

Ví dụ thực hiện sáng chế

Giải pháp đã được đưa vào hoạt động để xây dựng phương pháp phân tách, nhận dạng tiếng nói điện thoại viên, khách hàng và huấn luyện bán giám sát trong tổng đài chăm sóc khách hàng của Viettel.

Tại trung tâm chăm sóc khách hàng Viettel, chúng tôi sử dụng phương pháp phân tách, nhận dạng tiếng nói điện thoại viên và khách hàng để phân tách và chuyển đổi các cuộc gọi chăm sóc khách hàng sang văn bản. Từ đó có thể giám sát, thống kê được nội dung của các cuộc gọi một cách tự động và nhanh chóng. Ngoài ra, ta còn có thể biết được tâm tư, bức xúc của khách hàng cũng như việc trả lời khách hàng của điện thoại viên có đúng mực hay không. Hệ thống được cập nhật liên tục dựa trên cơ chế huấn luyện bán giám sát, từ đó giúp độ chính xác của hệ thống càng ngày càng được cải thiện.

Hiệu quả đạt được của sáng chế

Ưu điểm đặc biệt liên quan đến sáng chế này đó là xây dựng được một phương pháp phân tách, nhận dạng tiếng nói điện thoại viên, khách hàng và huấn luyện bán giám sát trong tổng đài chăm sóc khách hàng. Với phương pháp đề xuất này, người quản lý có thể biết được điện thoại viên và khách hàng của mình đang nói những gì. Từ đó biết được mong muốn, băn khoăn của khách hàng cũng như các điện thoại viên của mình có tư vấn chính xác, đúng mực hay không một cách nhanh chóng và khách quan. Ngoài ra hệ thống được cập nhật liên tục dựa trên cơ chế huấn luyện bán giám sát, tức hệ thống có thể tự học từ dữ liệu thực tế trong quá trình vận hành, từ đó giúp độ chính xác của hệ thống càng ngày càng được cải thiện.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

Yêu cầu bảo hộ

1. Phương pháp phân tách, nhận dạng tiếng nói, bao gồm:

bước 1: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài để phân tách; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như ổ đĩa cứng, băng từ, … hoặc thông qua các kết nối mạng dữ liệu, mỗi một tệp ứng với một cuộc gọi tổng đài; ta có thể lấy tệp tiếng nói trực tiếp trên thiết bị lưu trữ của người dùng hoặc sử dụng các giao thức truyền tệp qua mạng dữ liệu như FTP để lấy được các tín hiệu tiếng nói;

bước 2: phân tách và gán nhãn văn bản cho các tệp tiếng nói; tại bước này, đưa các tệp tiếng nói ở bước 1 lên hệ thống gán nhãn để người gán nhãn nghe, phân tách và gán nhãn văn bản cho phần nói của điện thoại viên và khách hàng; đầu ra của bước này là các tập tiếng nói đã được phân loại và gán nhãn riêng biệt thành tập tiếng nói của điện thoại viên và tập tiếng nói của khách hàng;

bước 3: tạo tập huấn luyện và kiểm thử; theo đó, khi dữ liệu tiếng nói được gán nhãn trong tệp của điện thoại viên và tệp của khách hàng ở bước 2 đều $\geq H_{label\_min}$ giờ dữ liệu, trong đó $H_{label\_min} \geq 10$ giờ nhằm đảm bảo tập dữ liệu đủ lớn; người quản trị quyết định lựa chọn một số tệp tiếng nói đã được gán nhãn ở bước 2 để tạo tập huấn luyện, các tệp còn lại được sử dụng để tạo tập kiểm thử với yêu cầu kích thước tập kiểm thử cần lớn hơn H_{test_min} giờ dữ liệu, trong đó $H_{test\_min} \geq 2$ giờ nhằm đảm bảo tập kiểm thử đủ lớn và tin cậy;

bước 4: xây dựng hai mô hình ngôn ngữ; LM_a cho điện thoại viên và LM_b cho khách hàng dựa trên tập dữ liệu huấn luyện được tạo ở bước 3 nhằm lưu trữ những đặc điểm về ngôn ngữ nói như các cụm từ thường xuyên nói của điện thoại viên và khách hàng từ đó để phân biệt được câu nói của điện thoại viên hay khách hàng ở các bước sau; các mô hình ngôn ngữ có thể được sử dụng như là n-gram hay các mô hình dựa trên mạng nơ ron nhân tạo được tạo tự động bằng các phần mềm chuyên dụng;

bước 5: thu thập dữ liệu tiếng nói các cuộc gọi tổng đài cần phân tách, nhận dạng tự động; bước này được thực hiện bằng các phương thức khác nhau như lấy tệp tiếng nói trực tiếp từ thiết bị lưu trữ như ổ đĩa cứng, băng từ, … hoặc thông qua các kết nối mạng dữ liệu, mỗi một tệp ứng với một cuộc gọi tổng đài; ta có thể lấy tệp tiếng nói trực tiếp trên thiết bị lưu trữ của người dùng hoặc sử dụng các giao thức truyền tệp qua mạng dữ liệu như FTP để lấy được các tín hiệu tiếng nói;

bước 6: tự động cắt tệp tiếng nói thành các đoạn nhỏ; với mỗi tệp tiếng nói thu được ở bước 5, tiếng nói được tự động cắt thành các đoạn nhỏ dựa theo các đặc tính về tín hiệu; ta có thể dựa vào các phương pháp phổ biến như: dựa trên năng lượng trung bình của tín hiệu hoặc ta có thể dựa trên các hệ thống nhận dạng tiếng nói;

bước 7: trích chọn các véc tơ đặc trưng người nói; tất cả các các đoạn tiếng nói thu được ở bước 6 được trích chọn véc tơ đặc trưng người nói bằng cách sử dụng một mạng trích chọn đặc trưng được huấn luyện trước như mạng nơ rơn học sâu (DNN), với mỗi đoạn tiếng nói sẽ thu được một véc tơ đặc trưng người nói tương ứng;

bước 8: phân cụm các đoạn tiếng nói; với mỗi tệp tiếng nói, phân cụm các đoạn tiếng nói ở bước 6 thành hai cụm C_1 và C_2 dựa trên các véc tơ đặc trưng người nói được trích xuất ở bước 7;

bước 9: chuyển đổi tiếng nói sang văn bản; tất cả các đoạn tiếng nói ở bước 6 được chuyển sang văn bản bằng cách sử dụng hệ thống nhận dạng tiếng nói, với mỗi đoạn tiếng nói thu được một văn bản tương ứng và một chỉ số độ tin cậy nhận dạng DTC có giá trị từ 0 đến 1;

bước 10: lựa chọn đoạn tiếng nói thỏa mãn điều kiện làm căn cứ phân loại; với mỗi một tệp tiếng nói, lựa chọn đoạn tiếng nói trong bước 9 thỏa mãn điều kiện: có độ tin cậy DTC $\geq \alpha$, trong đó $0,5 \leq \alpha \leq 0,95$ nhằm loại bỏ những đoạn tiếng nói có độ tin cậy quá thấp thường là những đoạn tiếng nói có chất lượng quá kém hoặc môi trường quá nhiễu ảnh hưởng đến chất lượng hệ thống phân loại; nếu không lựa chọn được đoạn tiếng nói nào thỏa mãn, bỏ qua tệp này và chuyển sang tệp tiếng nói mới;

bước 11: phân loại các đoạn tiếng nói của điện thoại viên và khách hàng; với các đoạn tiếng nói được lựa chọn ở bước 10 được chia thành hai cụm ở bước 8, tính $w = \frac{PPL_{a1} * PPL_{b2}}{PPL_{a2} * PPL_{b1}}$, trong đó PPL_{a1}, PPL_{a2}, PPL_{b1}, PPL_{b2} là chỉ số độ hỗn loạn (perplexity) được cho bởi các mô hình ngôn ngữ LM_a, LM_b ở bước 4 tính với tập dữ liệu văn bản của các đoạn tiếng nói được lựa chọn ở bước 10; PPL_{a1}, PPL_{b1} được tính ứng với các đoạn trong cụm C_1; PPL_{a2}, PPL_{b2} ứng với các đoạn trong cụm C_2; nếu $w \leq \theta$, toàn bộ các đoạn tiếng nói trong cụm C_1 được xác định là điện thoại viên, toàn bộ các đoạn tiếng nói trong cụm C_2 được xác định là khách hàng và ngược lại nếu $w > \theta$, toàn bộ các đoạn tiếng nói trong cụm C_2 được xác định là điện thoại viên, toàn bộ các đoạn tiếng nói trong cụm C_1 được xác định là khách hàng; ngưỡng θ có giá trị trong dải từ 0,5 đến 2,0; sau bước này ta đã phân

tách và nhận dạng được tiếng nói của điện thoại viên và khách hàng, nếu có nhu cầu về huấn luyện bán giám sát để nâng cao chất lượng hệ thống thực hiện tiếp bước 12, nếu không thì dừng lại;

bước 12: lựa chọn đoạn tiếng nói thỏa mãn điều kiện để đưa vào tập huấn luyện bán giám sát; lựa chọn đoạn tiếng nói trong bước 9 thỏa mãn điều kiện: có độ tin cậy DTC $\geq \beta$, trong đó $0,8 \leq \beta \leq 0,99$ nhằm lựa chọn những đoạn tiếng nói có độ tin cậy nhận dạng rất cao để đưa vào tập dữ liệu huấn luyện bán giám sát; mỗi đoạn tiếng nói đều đã được gán nhãn là điện thoại viên hay khách hàng từ bước 11;

bước 13: lựa chọn thời điểm cập nhật mô hình ngôn ngữ; là khi dữ liệu huấn luyện trong tập bán giám sát lớn hơn một ngưỡng H_{semi_min} giờ dữ liệu và khi có quyết định của người quản trị, trong đó $H_{semi\_min} \geq 10$ giờ với mục đích dữ liệu huấn luyện bán giám sát đủ lớn và tin cậy;

bước 14: xây dựng mô hình ngôn ngữ dựa trên dữ liệu bán giám sát; tại bước này sử dụng dữ liệu trong tập bán giám sát để xây dựng hai mô hình ngôn ngữ LM_{a_semi} với dữ liệu của điện thoại viên và LM_{b_semi} với dữ liệu của khách hàng; sau đó kết hợp với hai mô hình ngôn ngữ LM_a, LM_b ở bước 4 để tạo ra hai mô hình ngôn ngữ $LM_{a'}$, $LM_{b'}$ với hệ số kết hợp k, trong đó $0,8 \geq k \geq 0,1$;

bước 15: cập nhật mô hình ngôn ngữ; tính chỉ số $w_0 = \frac{PPL_{a1} * PPL_{b2}}{PPL_{a2} * PPL_{b1}}$, trong đó PPL_{a1}, PPL_{a2}, PPL_{b1}, PPL_{b2} là chỉ số độ hỗn loạn được cho bởi các mô hình ngôn ngữ LM_a, LM_b ở bước 4 tính với các tập kiểm thử ở bước 3; PPL_{a1}, PPL_{b1} được tính ứng với tập kiểm thử gồm các đoạn tiếng nói của điện thoại viên; PPL_{a2}, PPL_{b2} được tính ứng với tập kiểm thử gồm các đoạn tiếng nói của khách hàng; sau đó tính chỉ số $w_1 = \frac{PPL_{a'1} * PPL_{b'2}}{PPL_{a'2} * PPL_{b'1}}$ tương tự như w_0 bằng cách thay thế hai mô hình ngôn ngữ ở bước 4 bằng $LM_{a'}$ và $LM_{b'}$ ở bước 14; nếu $w_0 > q * w_1$ thì cập nhật LM_a bằng $LM_{a'}$, LM_b bằng $LM_{b'}$; trong đó $q \geq 1,0$.