BÁO CÁO

TÌM HIỂU VẤN ĐỀ

MỤC LỤC

Chương I - Giới thiệu bài toán

I.1. Giới thiệu vấn đề

Trong môi trường hội nhập quốc tế hiện nay, nhu cầu sử dụng tiếng Anh trở nên bức thiết. Từ đó các bài thi quốc tế ra đời nhằm mục đánh giá khả năng sử dụng tiếng Anh của một người. Các kỳ thi ngày trước như thi bằng A, B, C tiếng Anh hoặc ngày nay thường được dung như TOEIC, TOEFT, … có một đặc điểm chung là sử dụng bài tập trắc nghiệm ghi điểm làm đánh giá khả năng của người học. Hiện nay, việc tự học có được lợi thế cao từ mặt thời gian và độ linh động. Vì thế, việc thường xuyên rèn luyện trước các kỳ thi tiếng Anh là mộ trong các phương pháp nâng cao trình độ và bổ sung kiến thức. Tuy nhiên, khó khăn của việc từ học và làm bài tập là cần có được sự hướng dẫn trực tiếp. Ngày nay, đã có một số ứng dụng được đưa ra để người tự học có thể có đời lời giải thích cho một câu hỏi thuộc các lĩnh vực khác nhau như: Toán, Vật Lý, Hóa Học, … Tuy nhiên, chưa có công trình hay ứng dụng nào nhằm vào mục tiêu giải câu hỏi tiếng Anh. Một câu hỏi tiếng Anh, để có thể giải được cần một người có kinh nghiệm và hiểu biết để chọn đáp án chính xác và giải thích cho người học hiểu, đây là hạn chế. Vì thế khóa luận đưa ra ý tưởng sử dụng máy tính để giải và đưa ra câu trả lời mang tính tham khảo cho người học tiếng Anh.

Đến thời điểm này có ít các nghiên cứu và giải pháp tập trung vào tự động hóa quá trình chọn câu trả lời trả lời cho câu hỏi trắc nghiệm tiếng Anh. Việc ít các công trình nghiên cứu do độ khó trong việc giải quyết bài toán involves logical reasoning in addition to both general and semantic knowledge. Hiện nay, đây vẫn đang là bài toán thánh thức nếu sử dụng phương pháp semantic modeling. Các mô hình đòi hỏi đánh giá ngữ nghĩa cả câu.

I.2. Phát biểu bài toán và giới hạn đề tài

Bài tập trắc nghiệm tiếng Anh có nhiều dạng khác nhau như:

- Bài tập điền khuyết
- Tìm lỗi sai trong câu
- Đọc hiểu văn bản chọn câu đúng nhất
- Chọn từ thích hợp cho đoạn văn
- Chọn từ có trọng âm khác với từ còn lại
- Chọn từ đồng nghĩa
- …

Khóa luận giới hạn trong việc trả lời câu hỏi trắc nghiệm tiếng Anh dạng điều khuyết với một vị trí điền khuyết cùng với các lựa chọ phương án. Hệ thống sẽ xử lý và chọn một đáp án được cho là đúng.

Ví dụ:

```
Certain clear patterns in the metamorphosis of a butterfly indicate that the
process is ___.
(A) systematic
(B) voluntary
```

```
(C) spontaneous
(D) experimental
(E) clinical
```

Với (A), (B), (C), (D), (E) là các lựa chọn, (___) là vị trí điền khuyết trong câu.

Chương II - Lịch sử phát triển & khó khăn thách thức

Chương III - Các hướng tiếp cận

Ít các công trình nghiên cứu tập trung hẳn vào việc giải quyết bài toán một cách trực tiếp là "trả lời câu hỏi tiếng Anh dạng điền khuyết". Bài toán quy về một số hướng tiếp được đưa ra trước đây cho các vấn đề liên quan như: grammar check, sentence completion, …

Tuy nhiên vẫn có một số nghiên cứu trực tiếp tập trung vào việc giải quyết bài toán.

III.1. Hướng tiếp cận theo grammar check

Hướng tiếp cận này giải quyết bài toán "trả lời câu hỏi tiếng Anh dạng điền khuyết" dựa vào các mô hình được dùng để kiểm tra lỗi chính tả của câu. Ngữ pháp của một ngôn ngữ tự nhiên được biểu diễn bằng các cú pháp và hình thái từ. Do đó, kiểm tra ngữ pháp có thể hiểu là việc kiểm tra tính chính xác của cú pháp và hình thái đối với ngôn ngữ đang xét. Có nhiều phương pháp khác nhau để kiểm tra tính chính xác về ngữ pháp trên một đoạn văn bản. Từ các dữ liệu nhập vào, chương trình sẽ lần lượt thế các phương án vào chỗ trống, từ đó tìm ra phương án được cho là thích hợp nhất trả về cho người dùng. [1]

III.2. Mô hình n-gram

Lợi thế trong việc sử dụng mô hình n-gram là khả năng tính toán được xác suất xuất hiện của một chuỗi *token*. Dễ trong việc training trên các corpus không được dán nhãn. Tuy nhiên mô hình n-gram bị giới hạn do sử dụng nguồn dữ liệu đã thông qua training dẫn đến đánh giá dựa trên những câu đã được training, không thể phân tích những câu phức tạp, mang tính ngữ nghĩa cao do khoảng cách lớn giữa các token trong câu. [2]

III.3. Mô hình cây phụ thuộc (*Dependency models)*

Dependency models giải quyết được giới hạn của mô hình n-gram bằng cách biểu diễn mỗi từ bằng 1 node trong cây phụ thuộc. Mô hình *cây phụ thuộc không dán nhãn* coi mỗi từ là mỗi từ là một từ độc lập một cách có điều kiện so với những từ phía trước, được xử lý độc lập với mỗi quan hệ ngữ nghĩa.

Để giải quyết việc tính toán giá trị của câu, 2 câu khác nhau về trật từ giữa động từ và đối số của nó, mô hình *labeled dependency language* coi mỗi từ độc lập một cách có điều kiện và được gán nhãn bên ngoài.

Ưu điểm là đưa ra được hiệu suất cao hơn so với mô hình n-gram, lợi thế của cách biểu diễn nằm bao gồm việc training và ước tính dễ dàng cũng như khả năng tận dụng phương pháp làm mịn chuẩn (standard smoothing methods). Tuy nhiên, kết quả của mô hình

phụ thuộc vào phương pháp *automatic dependency extraction* và sự thưa thớt trong dữ liệu được thu thập. [2]

III.4. Continuous Space Models

Mạng neural giảm thiểu vấn đề thưa thớt dữ liệu bằng cách học các biểu diễn phân tán của các từ, chứng mình mô hình nổi trội trong việc bảo tồn những qui luật tuyến tính giữa các token. Mặc dù nhược điểm bao gồm độ mờ, xu hướng *overfitting*, và tăng yêu cầu tính toán. *Neural language models* đã vượt trội hơn mô hình n-gram và *dependency models*.

Mô hình kiến trúc Log-linear đã được đề xuất để giải quyết chi phí tính toán cho mô hình mạng neural. Mô hình *continuous bag-of-words* cố gắng đoán từ hiện tại bằng cách sử dụng *n* từ trong tương lai *và n* từ trong quá khứ làm ngữ cảnh. Ngược lại, *continuous skip-gram model* sử dụng từ hiện tại làm đầu vào để dự đoán những từ xung quanh. Sử dụng kiến trúc tổng thể bao gồm *skip-gram model* và mạng *neural*, đạt được hiệu suất cao trong *MSR Sentence Completion Challenge*.

III.5. PMI Model

Cách tiếp cận mô hình PMI dựa trên pointwise mutual information. Mô hình được thiếu kế nhằm vào nguồn thông tin gần và xa để tính toán tổng thể sự gắn kết trong câu. PMI dự trên lý thuyết đo đạc thông tin. PMI thể hiện sử tương quan giữa 2 từ *i* và *j* bằng cách so sánh xác suất của chúng dựa trên quan sát các từ trong cùng bối cảnh so với xác suất của việc quan sát các từ một cách độc lập.

The first step toward applying PMI to the sentence completion task involved constructing a word-context frequency matrix from the training corpus. The context was specified to include all words appearing in a single sentence, which is consistent with the hypothesis that it is necessary to examine word co-occurrences at the sentence level to achieve appropriate granularity. During development/test set processing, all words were converted to lowercase and stop words were removed based on their part-of-speech tags. To determine whether a particular part-of-speech tag type did, in fact, signal the presence of uninformative words, tokens assigned a hypothetically irrelevant tag were removed if their omission positively affected performance on the development portion of the MSR data set. This non-traditional approach, selected to increase specificity and eliminate dependence on a non-universal stop word list, led to the removal of determiners, coordinating conjunctions,pronouns, and proper nouns.1 Next, feature sets were defined to capture the various sources of information available in a sentence. While feature set number and type is configurable, composition varies, as sets are dynamically generated for each sentence at run time. Enumerated below are the three feature sets utilized by the PMI model.

Chương IV - Các công trình liên quan

IV.1. Hướng tiếp cận cận theo grammar check

Các công trình sử dụng *grammar check* dựa trên ý tưởng chính là lần lượt thế các đáp án vào vị trí trống. Chọn ra đáp án có tần số xuất hiện cao nhất dựa trên ngữ liệu đã học được. Một sinh viên trước đây ở trường cũng đã có khóa luận tốt nghiệp về vấn đề này. Anh tiếp cận đề tài bằng cách quy bài toán về vấn đề grammar check và giải quyết bài toán bằng n-gram kết hợp gán nhãn chủ ngữ với chủ ngữ là ngôi 1, ngôi thứ 2 và ngôi thứ 3. Việc gom nhóm chủ ngữ này giúp tăng tần số xuất hiên của các trường hợp tương đồng, giảm bớt sự phân tán tần số xuất hiện không đáng có cho các chủ ngữ khác nhau nhưng cùng ngữ pháp chia động từ. Ở bước kiểm tra so sánh để tìm ra đáp án ta cũng thực hiện việc gom nhóm chủ ngữ tương tự, nhờ đó với n-grams rơi vào các trường hợp chung sẽ cho ra kết quả chính xác hơn. [1]

Ở một số nghiên cứu khác, bài toán *grammar check* cũng được giải quyết bằng mô hình n-gram và xác suất thống kê như “Mô hình kiểm tra lỗi chính tả dựa trên xác suất”. Ý tưởng chính dựa trên xác suất, thu thập cái bi- tri- quad- và pentagram của một ngôn ngữ thông qua quá trình training dữ liệu. Trong quá trình training, thu thập xác suất của các n-gram. Sử dụng *"Word Class Agreements"* nhằm giải quyết 2 vấn đề đặc thù trong trong kiểm tra lỗi chính tả tiếng Anh là: *Adverb-verb-agreement* và *Adjective-noun-agreement* bằng cách lưu trữ song song các tù thường đi chung với nhau. Ví dụ: trạng từ "yesterday" sẽ được lưu trữ chung với tag động từ "verb (past tense)". [3]

IV.2. Hướng tiếp cận theo mô hình n-gram

Một công trình nghiên cứu khác sử dụng trực tiếp ngữ liệu n-gram của Google để chọn đáp án đúng trong câu hỏi multiple question tiếng Anh. Nghiên cứu chọn câu trả lời thông qua việc tách các n-gram xung quanh khoảng trắng để tra khảo trong cơ sở dữ liệu n-gram và chọn kết quả nào có số lần xuất hiện cao nhất. Sử dụng đề thi TOEIC để làm dataset kiểm nghiệm hệ thống. Nhóm tác giả đề xuất sử dụng lần lượt quad-gram và tri-gram để giải quyết bài toán sau khi lần lượt sử dụng các n-gram khác nhau để tính toán độ chính xác của từng n-gram với bài toán cụ thể. [4]

IV.3. Kiểm nghiệm độ chính xác giữa các hướng tiếp cận

Ở nghiên cứu [2] có đưa ra kết quả độ chính xác trong việc hoàn thành câu (sentence completion) giữa các mô hình thuật toán với nhau. Bài kiểm tra dựa trên data set của Microsoft Research Sentence Completion Challenge - bộ tổng hợp 1040 câu chứa khoảng trống và có đáp án được rút trích từ tác phẩm Sherlock Homles. Kết quả cho thấy mô hình PMI cho kết quả tốt hơn rất nhiều so với các mô hình tiền nhiệm trước đó.

Language Model	MSR
Random chance	20.00
N-gram [Zweig (2012b)]	39.00
Skip-gram [Mikolov (2013)]	48.00
LSA [Zweig (2012b)]	49.00
Labeled Dependency [Gubbins (2013)]	50.00
Dependency RNN [Mirowski (2015)]	53.50
RNNs [Mikolov (2013)]	55.40
Log-bilinear [Mnih (2013)]	55.50
Skip-gram + RNNs [Mikolov (2013)]	58.90
PMI	**61.44**

Figure 1: Best performance of various models on the MSR Sentence Completion Challenge. Values reflect overall accuracy [2]

Ở nghiên cứu [4], công trình nghiên cứu sử dụng corpus n-gram của Google để chọn đáp án đúng trong câu hỏi multiple question tiếng Anh. Nghiên cứu chọn option thông qua việc tách các n-gram xung quanh khoảng trắng (___) để tra khảo trong cơ sở dữ liệu n-gram và chọn kết quả nào có số lần xuất hiện cao nhất. Trong đó, tác giả sử dụng đề thi TOEIC để làm dataset kiểm nghiệm hệ thống. Kết quả cho thấy khi sử dụng lần lược quad-gram và tri-gram thì xác suất chính xác và hiệu suất tăng lên.

	Measurement	Vocabulary	Grammar	Total
5gram	Recall(%)	56.8	46.667	**53**
	Precision(%)	78.873	100	**85.849**
	F1-measure	66.041	63.636	**65.538**
4gram	Recall(%)	90.16	79.92	**85**
	Precision(%)	85.455	86.667	**85.882**
	F1-measure	87.746	881.504	**85.438**
Trigram	Recall(%)	100	98.611	**99.5**
	Precision(%)	75.781	85.915	**79.397**
	F1-measure	86.222	91.826	**88.318**
Trigram & 4gram	Recall(%)	100	97.436	**99**
	Precision(%)	83.607	86.842	**84.848**
	F1-measure	91.071	91.831	**91.379**

Figure 2 Evaluation results [4]

IV.4. Một số cách tiếp cận khác

Một số cách tiếp cận khác nhắc đến các công trình giải quyết các vấn đề khác nhưng gần giống với bài toán "trả lời câu hỏi trắc nghiệm tiếng Anh dạng điền khuyết".

Chương V - Tài liệu tham khảo

[1] Lê Quang Khải, "Xây dựng hệ thống tự động trả lời câu hỏi trắc nghiệm tiếng Anh dạng điền khuyết," Trường Đại học Công nghệ Thông Tin - Đại học Quốc gia thành phố Hồ Chí Minh, Hồ Chí Minh, 2013.

[2] A. M. Woods, "Exploiting Linguistic Features for Sentence Completion," Carnegie Mellon University, Pittsburgh, PA 15213, USA, 2016.

[3] V. Henrich and T. Reuter, "LISGrammarChecker: Language Independent Statistical Grammar Checking," Hochschule Darmstadt - University of Applied Sciences, 2009.

[4] D. Choi, M. Hwang, B. Ko and a. P. Kim, "Solving English Questions through Applying Collective Intelligence," Dept. Of Computer Engineering Chosun University; Korea Institute of Science and Technology Information, Gwangju, South Korea; Daejeon, South Korea, 2011.