# BÁO CÁO

# TÌM HIỂU VẤN ĐỀ

MỤC LỤC

## Chương I - Vấn đề cần giải quyết

Vấn đề cần giải quyết là việc chọn câu trả lời giữa các lựa chọn trong một câu tiếng Anh được đục lỗ. 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 khoảng trắng được đục lỗ trong câu.

## Chương II - Các phương pháp giải quyết vấn đề đã được đưa ra

Bài toán trả lời câu hỏi như trên tương đương với một số vấn đề được đưa ra trước đây như:

- Sentence completion
- Grammar check
- Automatic text completion
- Spell checking
- …

Một số keyword được sử dụng để tìm hiểu vấn đề:

```
grammar check
natural language processing
heuristics
Fill-in-the-blank exercise
English grammar
corpus
English learner
computer support for english grammar
Analogical Comparisons
Second Language Learning
cloze exercises
Open cloze
Language exercises
Computer-assisted language learning (CALL)
English as a foreign language (EFL)
Multiple-choice gap-fill
multiple-choice synonym
answering multiple choice english question
automatic text completion
spell checking
EasyEnglish
```

Trích: KHOÁ LUẬN TỐT NGHIỆP XÂY DỰNG HỆ THỐNG TỰ ĐỘNG TRẢ LỜI CÂU HỎI TRẮC NGHIỆM TIẾNG ANH DẠNG ĐIỀN KHUYẾT

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, bao gồm:

- **Pattern matching**
- **Rule-based approach**
- **Statistical approach**

## II.1. Statical Approach

### II.1.a. N-gram Models

**Study: Exploiting Linguistic Features for Sentence Completion (2/5 - 439)**

Lợi thế trong việc sử dụng mô hình n-gram thuần thì là khả năng tính toán được xác suất xuất hiện của một chuỗi *token*. Mặc dù dễ trong việc training trên các corpus không được dán nhãn. Nhưng mô hình n-gram bị giới hạn trong việc sử dụng nguồn dữ liệu đã thông qua training. Thực tế, mô hình này đánh giá quá cao dựa trên những câu đã được training cục bộ, gần như không thể phân tích những câu phức tạp, mang tính ngữ nghĩa cao do khoảng cách giữa các token trong câu.

**Study: Solving English Questions through Applying Collective Intelligence**

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 như hình bên dưới. Tác giả sử dụng lần lượt tri- quad- và penta-gram để ứng dụng vào trả lời câu hỏi. Kết luận rằng độ trả về của penta-gram chỉ nằm ở mức 56.8% cho thấy rất nhiều câu được trích xuất từ các câu hỏi trong để thì không khớp với nhiều penta gram trong hệ thống. Tác giả đề xuất sử dụng lần lượt quad-gram và tri-gram để trảlời một câu hỏi khi mà độ trả về của tri-gram là 100%. Tức gần như 100% câu khi được so khớp với tri-gram đều có kết quả trả về.

| Question & Answer | | ~ in order to inform ( ) about the purpose of ~<br>(A) themselves (B) them (C) that (D) it | |
|---|---|---|---|
| 5gram candidates | themselves | - in order to inform themselves | 170 |
| | | - order to inform themselves about | 68 |
| | | - to inform themselves about the | 1858 |
| | them | - in order to inform them | 2825 |
| | | - order to inform them about | 720 |
| | | - to inform them about the | 8980 |
| | | - inform them about the purpose | 47 |
| | | - them about the purpose of | 302 |
| | that | - in order to inform that | 118 |
| | it | - in order to inform it | 158 |
| | | - to inform it about the | 467 |
| *NG* = 5, *TNG* = 20 | | | |
| Sum | | themselves: 2096<br>that: 118 | them: 12874<br>it: 652 |

*Figure Trích từ nghiên cứu Solving English Questions through Applying Collective Intelligence*

| | 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 Trích từ nghiên cứu Solving English Questions through Applying Collective Intelligence*

**Study:** 09520134.pdf **(16)**

Luận văn đề tài này của tác giả trước ở trường - Ngram kết hợp gán nhãn chủ ngữ

Với chủ ngữ là ngôi 1, ngôi thứ 2 và ngôi thứ 3 số nhiều, ở thì hiện tại đơn chia động từ ở dạng nguyên mẫu.

| | | |
|---|---|---|
| PRP (personal pronoun) | I | FSP |
| PRP | He, she, it | TSP |
| PRP | You, we, they | PP |
| NNP (proper noun, singular) | John | TSP |
| NNPS (proper noun, plural) | Vikings | PPS |

Bảng 3.1: Phân loại các nhóm chủ ngữ

| | |
|---|---|
| **English sentence** | He lives in Ho Chi Minh city |
| **Semantic assigned sentence** | He/PRP lives/VBZ in/IN Ho/NNP Chi/NNP Minh/NNP city/NN |
| **Bigram** | TSP lives; lives in; in TSP; TSP city |
| **Trigram** | TSP lives in; lives in TSP; in TSP city |

Bảng 3.2: Ví dụ phân loại nhóm chủ ngữ cho câu

Tuy nhiên, với "to be" lại có sự tương đồng giữa ngôi 1 số nhiều, ngôi 2 và ngôi 3 số nhiều. Tương tự, đối với danh từ riêng hoặc các chủ ngữ thuộc ngôi thứ ba số ít sẽ có cách chia khác nhau. Từ những nhận xét trên, chúng tôi chia chủ ngữ thành ba nhóm và đặt tên như sau:

- Ngôi thứ nhất số ít – First singular pronoun - FSP
- Chủ ngữ số nhiều (We, they, you, danh từ riêng số nhiều)
- Plural pronoun - PP - Chủ ngữ số ít (he, she, it, danh từ riêng số ít) – Third singular pronoun – TSP

Dựa vào các nhãn từ loại của Penn Treebank ta sẽ tiến hành phân nhóm và thay thế tên nhóm vào vị trí của từ đang xét theo như bảng 3.1

Để giảm bớt độ phức tạp nên ta sẽ bỏ qua, không phân nhóm chủ ngữ cho các trường hợp chủ ngữ là các từ có nhãn NN (danh từ số ít) và NNS (danh từ số nhiều) vì dễ bị trùng với trường hợp danh từ làm tân ngữ, bổ ngữ trong câu.

Đối với chương trình giải câu hỏi trắc nghiệm tiếng Anh dạng điền khuyết, bên cạnh các câu hỏi kiểm tra ngữ pháp, chia động từ còn có các dạng về chọn cụm từ phù hợp với ngữ nghĩa, thành ngữ. . . Do đó ta chỉ xử lý các nhãn ngữ nghĩa được liệt kê ở bảng trên, các thành phần còn lại vẫn giữ nguyên và tiến hành thu thập n-grams. Trong ví dụ ở bảng 3.2, “Ho Chi Minh” là danh từ riêng và được gán thành ba nhãn NNP đứng gần nhau nên ta sẽ gom lại chỉ còn 1, do đó tổng cộng ta có 4 bigrams và 3 trigrams.

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.

**Study:** [LISGrammarChecker Language Independent Statistical Grammar Checking Master Thesis to achieve the academic degree](#) **(33)**

Mô hình kiểm tra lỗi chính tả dự trên xác suất: Language Independent Statistical Grammar Checker - LISGrammarChecker

Ý tưởng chính dựa trên xác suất, tu 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.

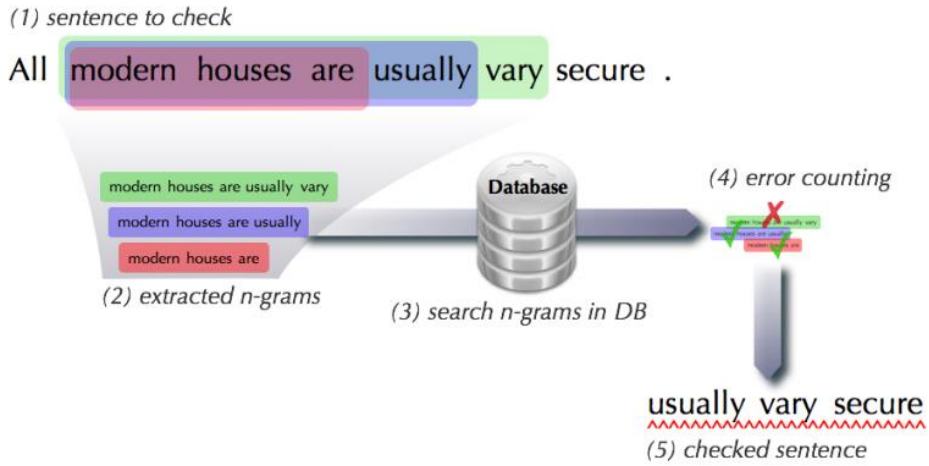


Figure 4.1.: Token n-gram check example

Điểm nhấn trong luận văn này là sử dụng khái niệm "Word Class Agreements"

"Word Class Agreements" nhằm giải quyết 2 vấn đề đặc thù trong trong sửa lỗi tiếng Anh. Ý tưởng chính của việc này không nằm ở việc kiểm tra chính xác ngữ pháp của cả câu, mà ý tưởng gần giống với cách tiếp cận của n-gram, cần nhiều dữ liệu để có thể training hệ thống.

- **Adverb-verb-agreement**: lưu trữ adverb (trạng từ) chung với tag của verb (động từ) trong câu đó. Ví dụ trạng từ "yesterday" sẽ được lưu trữ chung với tag động từ "verb (past tense)"

- **Adjective-noun-agreement**: lưu trữ toàn bộ các tính từ dùng để diễn tả cho danh từ. Ví dụ: lưu trữ tính từ *young* đi kèm với danh từ *girl*.


Figure 5.4.: Extract adverb and verb

Ưu điểm của hệ thống:
- tính độc lập về ngôn ngữ khi toàn bộ quá trình kiểm tra lỗi văn bản dự trên xác suất, đếm số lỗi và chọn vị trí lỗi của câu dựa trên xác suất.

- việc phát triển hệ thống dựa trên xác suất thống kê có các thách thức riêng tuy nhiên độ phức tạp thấp hơn so với cách tiếp cận rule-base khi mà việc thêm nhiều các luật để kiểm tra ngữ pháp dẫn đến độ phức tạp cao hơn của hệ thống rule-base.

- Độ chính xác được cho là cao hơn so với rule-base approach.

Nhược điểm:

- Yêu cầu bộ nhớ lưu trữ rất lớn khi so với hệ thống rule-base (Hệ thống kiểm tra lỗi chính tả trên Microsoft Word 97 sử dụng rule-base approach chỉ sử dụng khoảng 3 MB lưu trữ).

- Yêu cầu phải có hệ thống lưu trữ, quản lý, rút trích, truy vấn dữ liệu hiệu quả để có thể đáp ứng tốt về mặt tốc độ thực thi do phải truy cập dữ liệu nhiều lần.

- Yêu cầu dữ liệu training phải có độ chính xác cao.

- Tagger phải hoạt động chính xác. Để giải quyết việc này, tác giả tiếp cận bằng cách sử dụng nhiều tagger để cùng đánh tag một câu sau đó chọn phương án nào mà nhiều tagger sử dụng nhất.

### II.1.b. Dependency Models

*Dependency models* giải quyết được vấn đề giới hạn của mô hình n-gram bằng cách biểu diễn mỗi từ bằng 1 node trong cây Dependency. Mô hình *unlabeled depedency tree* 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.

Study: Exploiting Linguistic Features for Sentence Completion (2/5 - 439)

### II.1.c. Continuous Space Models

Mô hình mạng neural giảm thiểu vấn đề thưa thớt dữ liệu bằng cách học *distributed representations* của các từ, điều đó dừng để *excel at preserving linear regularities*

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, Mikolov et al. (2013) đạt được hiệu suất state-of-the-art cao trong *MSR Sentence Completion Challenge*.

Study: Exploiting Linguistic Features for Sentence Completion (2/5 - 439)

### II.1.d. PMI Model

**Study: Exploiting Linguistic Features for Sentence Completion (2/5 - 439)**

Cách tiếp cận mô hình PMI dựa trên pointwise mutual information. Mô hình được thiếu kế nhằm vào nguồn thông tin gần và xa đẻ 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 (Toutanova et al., 2003). 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.

- **Reduced Context**. This feature set consists of words that remain following the preprocessing steps described above.

- **Dependencies**. Sentence words that share a semantic dependency with the candidate word(s) are included in this set (Chen and Manning, 2014). Absent from the set of dependencies are words removed during the pre-processing phase. Figure 2 depicts an example dependency parse tree along with features provided to the PMI model.
- **Keywords**. Providing the model with a collection of salient tokens effectively increases the tokens' associated weights. An analogous approach to the one described for stop word identification was applied to discover that common nouns consistently hold greater significance than other words assigned hypothetically informative part-of-speech tags.

II.1.e. Kiểm nghiệm độ chính xác

Ở tiểu luận Exploiting Linguistic Features for Sentence Completion 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 chứ khoảng trống và có đáp án được rút trích từ tác phẩm Sherlock Homles. Cho thấy bảng như sau:

| 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** |

Kết quả cho thấy mô hình PMI cho kết quả tốt hơn rất nhiều so với các mô hình tiền nhiệm trước đó.

Ở nghiên cứu: Solving English Questions through Applying Collective Intelligence
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.

II.2. Mộ số cách tiếp cận khác

II.2.a. Heuristic

**Study:** Automated Grammar Checking of Tenses for ESL Writing

Nhóm tác giả bài tiểu luận *Automated Grammar Checking of Tenses for ESL Writing* đã giải quyết bài toán sau khi phân tích rằng các bài tiểu luận của học sinh Malaysia thường gặp lỗi ở sử dụng thì trong tiếng Anh. Nhóm tác giả đưa ra một hệ thống gồm 2 tầng: tầng xử lý ngôn ngữ tự nhiên và tầng xử lý logic.

Tầng xử lý ngôn ngữ tự nhiên bao gồm việc parsing một câu tiếng Anh, gán nhãn POS. Sau đó sử dụng các thuật toán heuristic để tìm lỗi, sửa lỗi.

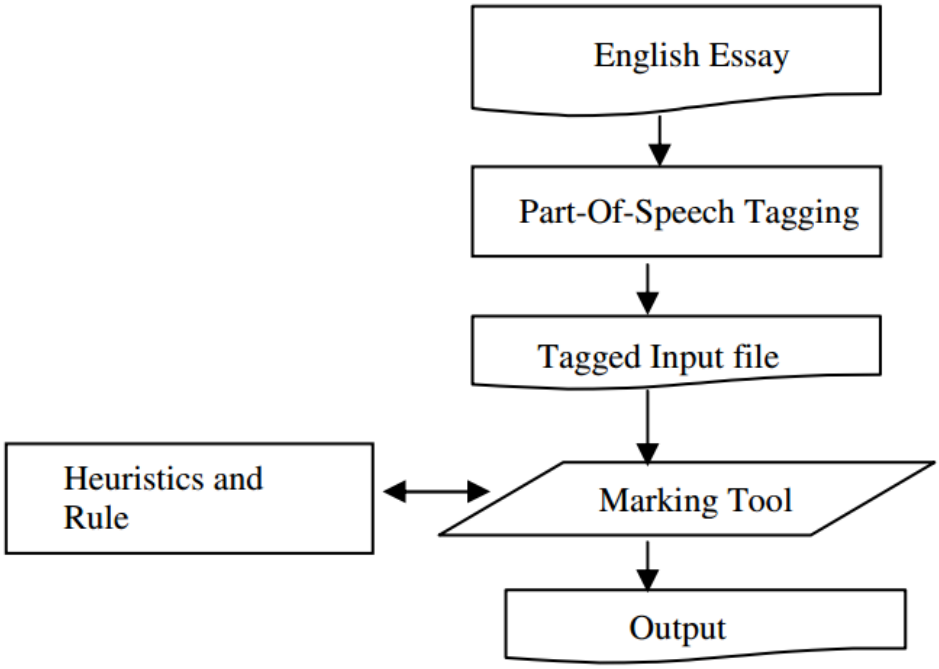


**Fig. 1.** Process in automated marking tool for ESL writing

II.3. Rule-based approach

**Study:** Reducing Grammar Errors for Translated English Sentences.pdf

**Study:** Developing a Chunk-based Grammar Checker for Translated English Sentences.pdf

Chunk-based approach

Mô hình chunk based sửa dụng xác suất lẫn rule-based approach để giải quyết vấn đề. Sửa dụng câu tiếng Anh làm input. Đầu tiên, câu sẽ được phân thành các *tokens* gán nhãn POS, sau được được gom nhóm lại thành các chunks. Parsing câu thành thành câu thành câu có cấu trúc  chunk based. Sau khi tạo thahf các chunks. Sau đó lỗi ngữ pháp sẽ được phát hiện dựa trên các sentence pattern. Nếu không tìm được sentence pattern thích hợp thì hệ thống sẽ tìm ra lỗi và sửa lỗi.

## Chương III - Đề xuất

Đề xuất sử dụng mô hình PMI được đề cập ở tiểu luận: Exploiting Linguistic Features for Sentence Completion. Vì tiểu luận đã đưa ra các lập luận để sử dụng mô hình PMI cũng như đã tính toán hiệu suất của mô hình cao hơn so với các mô hình trường đó. Việc đưa ra được hiệu suất và hiện cũng đã tìm được nhiều tiểu luận, luận văn gần đây sử dụng mô hình PMI để giải quyết vấn đề kể trên như giải đề thi SAT, TOEIC, … cho thấy mô hình cho hiệu suất cao, dễ hiện thực.

Đề xuất quy trình:

- Liên hệ chủ nhiệm đề tài trước để có thể sử dụng lại ngân hàng câu hỏi. Nhằm tiết kiệm chi phí và cũng để sử dụng, ước tính lại hiệu suất có cao hơn so với hệ thống trước đó hay không.
- Tìm hiểu mô hình PMI, các giải thuật, thuật toán liên quan
- Tìm hiểu các đề tài đã sử dụng mô hình PMI để giải quyết vấn đề
- Tìm hiểu vấn đề sử dụng corpus trong mô hình PMI, xem có thể sử dụng lại corpus có sẵn hay không hay phải tự dựng lại corpus
- Hiện thực hệ thống ở mức đơn giản, có thể truy vấn và lấy câu trả lời dựa trên corpus nhỏ
- Hiện thực hệ thống sử dụng corpus lớn hơn.
- Dựng demo.