Kết hợp BERTopic và Mô hình tóm tắt Đa văn bản cho Dữ liệu tin tức trên các trang mạng xã hội

Lưu Bảo Uyên[1], Lê Vy[2], Trần Lương Vân Nhi[3]
[1,2,3]Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh
[1]22521640@gm.uit.edu.vn, [2]22521703@gm.uit.edu.vn, [3]22521044@gm.uit.edu.vn

***Tóm tắt nội dung*—Trước sự bùng nổ của thông tin trên các nền tảng mạng xã hội, người dùng ngày càng gặp khó khăn trong việc theo dõi và tổng hợp tin tức. Bài toán phân cụm chủ đề và tóm tắt văn bản đóng vai trò quan trọng trong việc hỗ trợ người dùng nắm bắt nhanh chóng các sự kiện chính. Trong nghiên cứu này, chúng tôi đề xuất một hệ thống kết hợp giữa BERTopic để phân cụm các bài báo theo chủ đề và mô hình tạo sinh văn bản ViT5 và BARTpho để tạo ra các bản tóm tắt ngắn gọn cho từng bài báo trong cụm chủ đề. Hệ thống được thực nghiệm trên bộ dữ liệu Vietnamese Multi Document Summarization Dataset (ViMs), một tập dữ liệu phù hợp cho bài toán tóm tắt đa văn bản tiếng Việt. Kết quả thực nghiệm cho thấy BERTopic thực hiện tốt trong việc phân cụm chủ đề, tuy nhiên vẫn có thể tinh chỉnh embedder để có kết quả tốt hơn. Đối với kết quả tóm tắt văn bản, ViT5 có khả năng ghi nhận thông tin quan trọng và chính xác hơn BARTpho, nhưng BARTpho lại có xu hướng đưa ra từ khóa tóm tắt gần với văn bản tham chiếu hơn.**

***Từ khóa*—phân cụm chủ đề, tóm tắt đa văn bản, BERTopic, ViT5, BARTpho, xử lý ngôn ngữ tự nhiên, tiếng Việt**

I. Giới thiệu

Trong bối cảnh hiện đại hóa, hiện tượng bùng nổ thông tin đã và đang trở thành một thách thức lớn đối với người dùng trong việc tiếp nhận và xử lý dữ liệu. Bùng nổ thông tin là hiện tượng lượng thông tin được tạo ra đang ngày càng gia tăng mạnh mẽ và nhanh chóng. Sự phát triển vượt bậc của công nghệ thông tin, đặc biệt là mạng Internet và thiết bị di động thông minh, đã thúc đẩy quá trình tạo ra và lan truyền thông tin. Mỗi ngày, khối lượng thông tin khổng lồ với hàng tỷ dữ liệu mới được tạo ra và lan truyền với tốc độ nhanh chóng trên các trang mạng xã hội. Người dùng có thể tiếp cận thông tin dễ dàng và tức thời. Tuy nhiên, mặt trái của hiện tượng này là sự quá tải thông tin. Sự đa dạng của các nguồn tin, sự trùng lặp nội dung, hoặc thậm chí là thông tin sai lệch sẽ khiến người dùng gặp khó khăn trong việc tìm kiếm, phân tích và chọn lọc những nội dung quan trọng, đáng tin cậy. Đặc biệt trong lĩnh vực tin tức, khi một sự kiện diễn ra, sẽ có hàng loạt bài báo khác nhau có thể được xuất bản bởi nhiều cơ quan truyền thông, mỗi bài mang một góc nhìn hoặc mức độ chi tiết riêng biệt.

Nhóm chúng tôi đề xuất xây dựng một hệ thống tự động nhằm hỗ trợ người dùng trong việc tiếp nhận và tổng hợp thông tin hiệu quả hơn. Hệ thống có khả năng phân cụm các bài báo theo sự kiện mà chúng đề cập đến, đồng thời tạo ra bản tóm tắt ngắn gọn cho mỗi cụm bài báo tương ứng. Nhờ đó, người dùng có thể nhanh chóng nắm bắt được: trong một khoảng thời gian cụ thể, các bài báo đang nói về những sự kiện gì, và nội dung chính của các sự kiện đó là gì.

II. Khái quát bài toán

Ý tưởng nghiên cứu được xây dựng dựa trên sự kết hợp giữa hai hướng tiếp cận trong xử lý ngôn ngữ tự nhiên: phân cụm chủ đề (Topic Clustering) và tóm tắt văn bản (Text Summarization). Hai bài toán này được triển khai nhằm giải quyết một mục tiêu tổng thể là hỗ trợ người dùng nắm bắt nội dung chính của các bài báo trên mạng xã hội có hệ thống và theo từng chủ đề. Cụ thể, nghiên cứu này hướng đến hai mục tiêu chính:

- Phân cụm bài báo theo chủ đề: Sử dụng các kỹ thuật mô hình chủ đề (Topic Modeling) để nhóm các bài báo có nội dung tương đồng vào cùng một cụm, đại diện cho một chủ đề hoặc sự kiện.
- Tạo sinh bản tóm tắt cho mỗi bài báo: Sử dụng các mô hình tóm tắt đa văn bản để tạo ra các đoạn văn ngắn nhằm tóm lược nội dung chính của từng bài báo trong cụm chủ đề.

Để đánh giá hiệu quả của mô hình, chúng tôi sử dụng Vietnamese Multi-document Summarization Dataset (ViMs) – một tập dữ liệu gồm các nhóm bài báo tiếng Việt được gắn nhãn theo chủ đề và chứa bản tóm tắt tham chiếu. Trong phần I và phần II, chúng tôi trình bày tổng quan đề tài, bối cảnh và mục tiêu nghiên cứu. Các nghiên cứu liên quan đến bài toán sẽ được đề cập tại phần III. Phần IV cung cấp thông tin chi tiết về bộ dữ liệu. Phần V trình bày chi tiết các mô hình và phương pháp được sử dụng để giải quyết hai bài toán thành phần. Các kết quả thực nghiệm và phân tích được trình bày trong Phần VI.

Đầu vào của hệ thống là tập hợp các bài báo tiếng Việt ở định dạng văn bản.

Đầu ra gồm hai thành phần:

- Tập các cụm chủ đề được phát hiện từ bộ dữ liệu
- Các văn bản tóm tắt tương ứng cho từng bài báo trong mỗi cụm

III. Nghiên cứu liên quan

Trong những năm gần đây, phân cụm chủ đề (topic clustering) và tóm tắt văn bản (text summarization) là đề tài nhận được nhiều sự quan tâm trong cộng đồng xử lý ngôn ngữ tự nhiên. Trên thế giới, rất nhiều bài toán trong lĩnh vực

này được đề xuất và phát triển, tuy nhiên tiếng Việt lại là một trong những ngôn ngữ ít tài nguyên trong lĩnh vực này.

Đối với bài toán phân cụm chủ đề, các nghiên cứu thường áp dụng Latent Dirichlet Allocation hoặc các phương pháp phân cụm dựa trên biểu diễn vector truyền thống. Một số công trình như "An integrated clustering and BERT framework for improved topic modeling"[1] của Geogre đã đề xuất một mô hình lai (hybrid model) kết hợp Bidirectional Encoder Representations from Transformers (BERT) và LDA trong mô hình hóa chủ đề. Đặc biệt, công trình này còn tích hợp kỹ thuật phân cụm (clustering) dựa trên các phương pháp giảm chiều dữ liệu như PCA, t-SNE và UMAP, nhằm giải quyết độ phức tạp tính toán của thuật toán phân cụm khi số lượng đặc trưng tăng lên. Cuối cùng, một khung làm việc thống nhất dựa trên phân cụm, sử dụng BERT và LDA, được trình bày với mục tiêu khai thác các chủ đề có ý nghĩa từ kho dữ liệu văn bản lớn. Hoặc một công trình khác như "Vietnamese Facebook Posts Classification using Fine-Tuning BERT"[2] đã đề xuất một mô hình phân loại bài viết công khai trên Facebook bằng cách tinh chỉnh mô hình BERT với ba chiến lược cắt ngắn văn bản khác nhau, đồng thời nhóm tác giả đã xây dựng một bộ dữ liệu mới gồm 5.191 bài viết tiếng Việt, được gán nhãn theo chủ đề và chia thành ba tập huấn luyện, kiểm định và kiểm tra.

Bài toán tóm tắt văn bản đã được nghiên cứu trong nhiều cuộc thi như VLSP (Vietnamese Language and Speech Processing) [3]. Gần đây, cũng đã thực hiện một bài khảo sát "A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods"[4], tổng hợp các phương pháp tóm tắt văn bản tự động, đặc biệt chú trọng vào tính thực tiễn khi áp dụng vào hệ thống thực tế. Đây cũng là một trong những khảo sát đầu tiên tập trung sâu vào các phương pháp mới nhất sử dụng mô hình ngôn ngữ lớn (Large Language Model).

Tuy nhiên việc kết hợp phân cụm chủ đề và tóm tắt văn bản vẫn còn tương đối mới. Có một số nghiên cứu quốc tế đã từng đề cập đến pipeline gồm phân cụm văn bản trước sau đó sinh tóm tắt cho từng nhóm nội dung. Nhưng các hướng tiếp cận này chưa được phổ biến và chưa được kiểm chứng trên các tập dữ liệu đa tài liệu tiếng Việt.

IV. Bộ dữ liệu

Tên bộ dữ liệu: Vietnamese Multi Document Summarization Dataset [5]

Tác giả: Nghiem Quoc Minh.

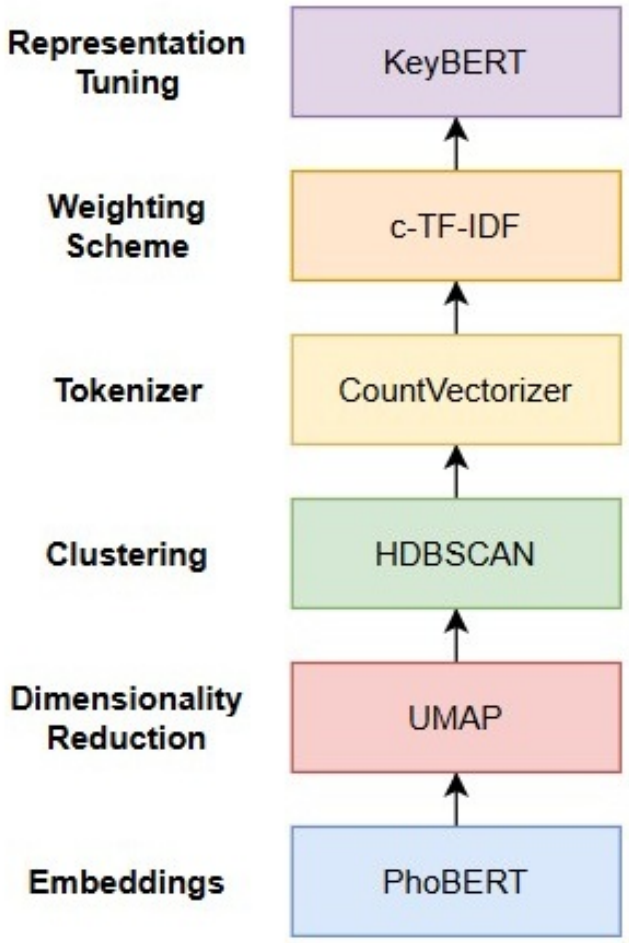
Tổ chức: Bộ môn Công nghệ Tri Thức, Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh.

Đây là bộ dữ liệu được nhóm tác giả thu thập nhằm mục đích "Xây dựng công cụ tổng hợp tin tức tiếng Việt và ứng dụng", có sự hỗ trợ của Sở Khoa học và Công nghệ Thành phố Hồ Chí Minh. Dữ liệu được thu thập một cách thủ công từ các trang báo khá nổi tiếng và phổ biến ở Việt Nam như: Vnexpress[6], Dân Trí[7], Tuổi Trẻ[8],... Với số lượng 300 nhóm văn bản, trong đó mỗi nhóm văn bản sẽ có ít nhất là 5 bài, nhiều nhất là 10 bài. Các bài báo trong bộ dữ liệu, bao gồm: tin thế giới, tin trong nước, giải trí, kinh tế và thể thao.

Chúng tôi sử dụng chính là thư mục original của bộ dữ liệu, thư mục này bao gồm 300 thư mục con chính là 300 cụm văn bản, trong mỗi cụm văn bản là các văn bản thuộc cùng chủ đề. Có thể có 5 đến 10 bài cho mỗi cụm văn bản với tổng số lượng văn bản là 1,945.

V. Lý thuyết mô hình

A. BERTopic



Hình 1. Pipeline BERTopic

BERTopic[9] là mô hình được phát triển bởi Maarten Grootendorst và hiện đang được xem là một trong những mô hình đạt SOTA trong lĩnh vực mô hình hóa chủ đề (topic modeling). Nó tận dụng các mô hình nhúng văn bản (embedding models) và kỹ thuật phân cụm để khám phá các chủ đề trong một tập hợp tài liệu. Khác với các mô hình truyền thống như Latent Dirichlet Allocation (LDA)[10] và Non-negative Matrix Factorization (NMF)[11] dựa trên cách tiếp cận thống kê hoặc đại số tuyến tính và biểu diễn văn bản dưới dạng "túi từ"(bag-of-words), BERTopic sử dụng phương pháp nhúng (embedding approach) kết hợp phân cụm mật độ. Điều này cho phép BERTopic mã hóa ý nghĩa ngữ nghĩa của

văn bản, làm cho các văn bản có ý nghĩa tương tự sẽ gần nhau trong không gian vector.

BERTopic là một hệ thống gồm 6 thành phần cốt lõi có thể được tùy chỉnh để phù hợp với nhiều trường hợp sử dụng khác nhau. Kiến trúc điển hình của nó được thể hiện qua các bước sau:

- Embeddings: Chuyển đổi văn bản đầu vào thành các biểu diễn dạng vector (embedding) thể hiện ý nghĩa ngữ nghĩa, bằng cách sử dụng các mô hình sentence-transformer.
- Giảm chiều (Dimensionality Reduction): Giảm số chiều của vector embedding xuống không gian thấp hơn mà vẫn giữ được các mối quan hệ quan trọng. Các thuật toán sử dụng cho bước này là PCA, UMAP,...
- Phân cụm (Clustering): Gom nhóm các văn bản tương tự lại với nhau dựa trên embedding đã giảm chiều để hình thành các chủ đề riêng biệt, bằng cách sử dụng các thuật toán như HDBSCAN, K-Means,...
- Vectorizer: Sau khi các cụm chủ đề được hình thành, vectorizer chuyển văn bản thành ma trận tần suất từ vựng để phân tích chủ đề của cụm. Các thuật toán thường dùng là count vectorizer, online vectorizer,...
- c-TF-IDF: Tính điểm quan trọng của các từ trong và giữa các cụm chủ đề qua ma trận trọng số ở bước Vectorizer nhằm xác định các từ khóa chính.
- Mô hình biểu diễn (Representation Tuning): Tận dụng sự tương đồng ngữ nghĩa giữa embedding của từ khóa ứng viên và embedding của văn bản để tìm ra các từ khóa đại diện nhất cho chủ đề. Các mô hình phổ biến ở bước này là KeyBERT, các kỹ thuật dựa trên mô hình ngôn ngữ lớn (LLM),...

Embeddings là khối xây dựng BERTopic đầu tiên và ảnh hưởng quan trọng đến kết quả của những bước sau đó. Cho dữ liệu tiếng Việt, chúng tôi chọn PhoBERT[12] để tạo nhúng tài liệu. BERTopic sẽ tính toán mức độ tương đồng ngữ nghĩa giữa các văn bản đầu vào bằng cách sử dụng khoảng cách cosine giữa các vector embedding.

Phục vụ cho bài toán phân cụm chủ đề, embedder của PhoBERT cần được tinh chỉnh với mục tiêu hiểu ngữ nghĩa là chủ đề trong văn bản. Nhóm thực hiện tinh chỉnh PhoBERT embedder theo chiến lược học đối kháng (Contrastive Learning), bằng cách với mỗi cặp văn bản đầu vào:

- Gán nhãn cho chúng là tích cực (positive), nếu cặp văn bản được gán nhãn cùng chủ đề trong bộ dữ liệu.
- Gán nhãn cho chúng là tiêu cực (negative), nếu trong bộ dữ liệu, hai văn bản được chọn từ hai chủ đề khác nhau.
- Hàm Loss là CosineEmbeddingLoss, đúng với tinh thần của các sentence-transformer như SimCSE,...

Sau khi huấn luyện, embedding từ token [CLS] đại diện cho văn bản đầu vào.

Đầu ra của bước Embeddings là vector ngữ nghĩa 768 chiều, và chúng cần được giảm chiều để tăng hiệu quả cho mô hình. PCA áp dụng tốt cho trường hợp bộ dữ liệu chứa những cụm đồng đều về số lượng. Trong khi đó, mục tiêu của mô hình là linh hoạt với số lượng sự kiện chưa biết trước trong bộ dữ liệu thực tế, nên BERTopic ở đây sử dụng Uniform Manifold Approximation and Projection (UMAP)[13] để giảm chiều sâu dữ liệu. UMAP giúp giảm chiều của vector embedding mà vẫn giữ cấu trúc khoảng cách quan trọng và không hạn chế tính toán về kích thước chiều nhúng. Các tham số cần lựa chọn cho thuật toán này là n_neighbours và min_dist:

- n_neighbors xác định UMAP sẽ xem xét bao nhiêu hàng xóm gần nhất quanh mỗi điểm. Giá trị n_neighbors nhỏ sẽ khiến UMAP tập trung xem xét các cụm nhỏ. Giá trị n_neighbors lớn sẽ mở rộng vùng lân cận, giúp UMAP nắm bắt được cấu trúc tổng thể của dữ liệu tốt hơn. Các tham số được xem xét cho biến này là [2, 4, 7, 10].
- min_dist quy định khoảng cách tối thiểu giữa các điểm trong không gian khi UMAP giảm chiều. Giá trị min_dist nhỏ (ví dụ 0.001) giúp tách cụm rõ hơn. Giá trị min_dist lớn (0.5, 1) khiến các điểm dữ liệu nằm dàn trải, cân đối hơn. Vì tính cách mỗi sự kiện thực tế xảy ra không nên bị phân qua cụm của sự kiện khác chỉ để đảm bảo sự cân đối, nên min_dist được lựa chọn nhỏ với các tham số [0.0, 0.1].

Sau khi giảm chiều, các nhúng tài liệu được gom nhóm bằng thuật toán Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[14]. HDBSCAN là một phần mở rộng của DBSCAN[15], có khả năng tìm các cụm có mật độ khác nhau và coi các tài liệu không liên quan là nhiễu (outliers). Điều này giúp ngăn chặn các tài liệu không liên quan bị gán vào bất kỳ cụm nào và cải thiện biểu diễn chủ đề. HDBSCAN gom các bài báo tương đồng thành một cụm, mỗi cụm lý tưởng là một sự kiện tồn tại trong bộ dữ liệu. Các biến số cần được điều chỉnh cho thuật toán bao gồm:

- min_cluster_size quy định kích thước nhỏ nhất của một cụm, trong đó cụm nhỏ hơn số lượng đó sẽ bị coi là nhiễu. Do mỗi sự kiện thực tế thường được nhắc đến bởi số lượng nhỏ bài báo (bộ dữ liệu cho thấy mỗi sự kiện gồm 5-10 bài báo), nên các tham số cho biến này được nhóm chọn là [2, 4, 6, 8].
- cluster_selection_method là cách chọn cụm từ cây phân cấp. ['eom', 'leaf'] là các tham số cần chọn.

Sau khi hoàn thành việc phân cụm văn bản, BERTopic sử dụng công cụ CountVectorizer để trích xuất các từ khóa đại diện cho từng cụm chủ đề nhằm giúp mô tả nội dung tổng quát của cụm đó. CountVectorizer có chức năng chuyển đổi văn bản trong mỗi cụm thành ma trận đếm từ, trong đó mỗi hàng tương ứng với một tài liệu, và mỗi cột biểu thị số lần một từ cụ thể xuất hiện. Thông qua ma trận này, hệ thống có thể xác định các từ có tần suất xuất hiện cao và ổn định trong cụm. Chúng tôi áp dụng bỏ stopwords tiếng Việt và quy định kích thước của các cụm từ (n-gram) sẽ được trích xuất từ văn bản là (1, 2).

Để tăng độ chính xác và khả năng khái quát của từ khóa đại diện, BERTopic áp dụng một biến thể mở rộng của phương pháp TF-IDF[16] truyền thống là class-based TF-IDF (c-TF-IDF). Trong đó thay vì tính toán độ quan trọng của một từ trong từng tài liệu riêng lẻ, c-TF-IDF đánh giá mức độ đại diện của một từ đối với toàn bộ cụm tài liệu, tức là một chủ

đề. Cụ thể, tất cả các văn bản trong cùng một cụm được nối lại thành một văn bản lớn duy nhất, từ đó mô hình xem mỗi cụm như một "lớp" (class) độc lập. Khi đó, tần suất xuất hiện của từ trong một cụm (term frequency theo cụm, ký hiệu là tft,c) được tính toán tương tự như TF truyền thống. Phương pháp này cho thấy sự hiệu quả rõ hơn so với các phương pháp biểu diễn dựa trên tâm cụm vì nó có thể phân tích tần suất từ ở cấp độ cụm và giảm các từ xuất hiện nhiều ở mọi cụm. Từ đó hạn chế được việc phương pháp gây sai lệch nếu dữ liệu không phân bố quanh trung tâm.

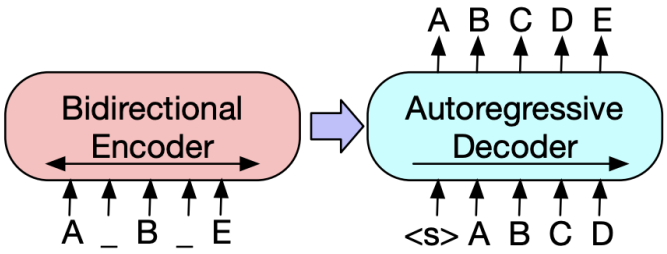
Cuối cùng, Representation Model là bước chọn ra từ khóa đại diện cho mỗi chủ đề. Nó nhận dữ liệu được xử lý tốt từ CountVectorizer và c-TF-IDF, sử dụng embedder PhoBERT đã được tinh chỉnh ở trên để chọn từ có embedding gần nhất với vector. Từ đó đưa ra danh sách từ khóa đại diện cho cụm chủ đề có tính ngữ nghĩa cao hơn. KeyBERTInspired, áp dụng kỹ thuật của KeyBERT được chọn vì tiết kiệm tài nguyên tính toán. Đây là một module được phát triển trong thư viện BERTopic.

B. BARTpho

BARTpho có hai phiên bản là BARTpho$_{syllable}$ và BARTpho$_{word}$, đây là những mô hình sequence-to-sequence đơn ngữ quy mô lớn đầu tiên được huấn luyện trước (pre-trained) cho tiếng Việt. Các mô hình này được phát triển để giải quyết việc thiếu các mô hình sequence-to-sequence đơn ngữ cho tiếng Việt và để khắc phục hạn chế không nhận biết được sự khác biệt đặc trưng về mặt ngôn ngữ giữa âm tiết và từ trong tiếng Việt của các mô hình đa ngôn ngữ hiện có. Mô hình này đặc biệt phù hợp cho các tác vụ NLP tạo sinh văn bản.

Trước BARTpho, chưa có mô hình sequence-to-sequence đơn ngữ nào được huấn luyện trước cụ thể cho tiếng Việt. Các mô hình ngôn ngữ lớn như BERT[17] và các biến thể của nó đã cải thiện hiệu suất cho nhiều tác vụ hiểu ngôn ngữ tự nhiên, nhưng chúng khó áp dụng trực tiếp cho các tác vụ tạo sinh văn bản. Các mô hình sequence-to-sequence huấn luyện trước đã được đề xuất để giải quyết vấn đề này, nhưng thành công của chúng chủ yếu giới hạn ở tiếng Anh. BARTpho được tạo ra để giải quyết vấn đề này cho tiếng Việt.

BARTpho sử dụng kiến trúc "large"với 12 lớp encoder và 12 lớp encoder, dựa trên cấu trúc của mô hình BART[18]. Giống như mBART[19], BARTpho cũng bổ sung một lớp chuẩn hóa (layer-normalization) vào đầu cả encoder và decoder. So với mô hình BART gốc, BARTpho dùng hàm kích hoạt GeLU thay vì ReLU và khởi tạo tham số từ phân phối chuẩn. BARTpho được huấn luyện trên một phiên bản đã bỏ token (detokenized) của tập dữ liệu huấn luyện trước PhoBERT, khoảng 4 tỷ token âm tiết. Bartpho sử dụng SentencePiece[20] được huấn luyện trước từ mô hình XLM-RoBERTa[21] (được dùng trong mBART) để tiếp tục phân đoạn. Từ vựng được chọn bao gồm 40K loại phổ biến nhất.

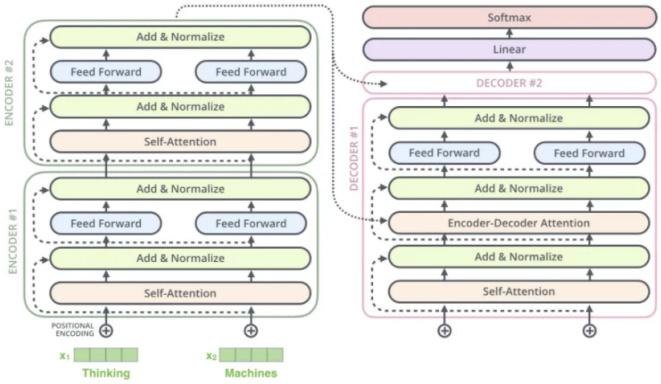


Hình 2. BART

C. ViT5

ViT5[22] là một mô hình Transformer được huấn luyện trước cho ngôn ngữ tiếng Việt. Mô hình này được thiết kế cho các tác vụ sinh văn bản tiếng Việt và được huấn luyện theo kiểu tự giám sát T5 (T5-style self-supervised pretraining). Sự phát triển của ViT5 được lấy cảm hứng từ thành công của các mô hình Transformer lớn như T5 cho tiếng Anh, và các nghiên cứu trước đây đã chỉ ra rằng kiến trúc T5 có thể vượt trội hơn BART trong một số khía cạnh.

Kiến trúc của ViT5 tuân theo kiến trúc encoder-decoder được đề xuất bởi Vaswani và khung T5[23] của Raffel với mục tiêu huấn luyện tự giám sát. Bằng việc sử dụng "span-corruption"với tỷ lệ lỗi là 15%, các đoạn (spans) ngẫu nhiên trong chuỗi đầu vào bị che đi, và mô hình cố gắng dự đoán các token gốc.



Hình 3. Mô hình T5

Sử dụng mục tiêu "span-corruption"với tỷ lệ lỗi là 15%. Đây là một phương pháp tự giám sát trong đó các đoạn (spans) ngẫu nhiên trong chuỗi đầu vào bị che đi, và mô hình cố gắng dự đoán các token gốc.

ViT5 được huấn luyện trên một tập dữ liệu tiếng Việt đơn ngữ lớn, chất lượng cao và đa dạng. Cụ thể, nó sử dụng tập con tiếng Việt đơn ngữ của bộ dữ liệu CC100 (Monolingual Datasets from Web Crawl Data)[24] với tổng kích thước của tập dữ liệu là 138GB văn bản thô. Sau khi xử lý và lọc, 69GB đoạn văn ngắn được sử dụng cho mô hình 256-length và 71GB đoạn văn dài cho mô hình 1024-length. Việc mô hình được huấn luyện với hai độ dài đầu vào và đầu ra khác nhau: 256

và 1024. Phân tích cho thấy độ dài ngữ cảnh trong quá trình huấn luyện tự giám sát có vai trò quan trọng đối với hiệu suất ở các downstream task. Đặc biệt, độ dài dài hơn (khớp với độ dài tài liệu của downstream task) dẫn đến kết quả tốt hơn. ViT5 sử dụng SentencePiece huấn luyện trên một tập từ vựng gồm 36K từ con (sub-words). Quá trình tiền xử lý dữ liệu để tạo từ vựng bao gồm chuẩn hóa dấu câu và chữ hoa, cũng như tách số.

VI. Thiết kế thực nghiệm

Các siêu tham số được lựa chọn dựa trên các thiết lập phổ biến trong huấn luyện mô hình transformer cho tiếng Việt, đồng thời đảm bảo hiệu quả tính toán trong điều kiện tài nguyên hạn chế.

A. BERTopic

Trong quá trình thực nghiệm, mô hình BERTopic được huấn luyện nhằm trích xuất văn bản vào từng cụm chủ đề. Trong đó dữ liệu đầu vào được mã hóa với độ dài tối đa (max length) là 256 token. Quá trình huấn luyện được tiến hành với kích thước (batch size) là 8, sử dụng thuật toán tối ưu hóa AdamW với tốc độ học (learning rate) đặt là 2e-5. Mô hình được huấn luyện trong 10 epochs.

	batch size	max length	LR		
BERTopic	8	256	2e-5	AdamW	10 epochs

B. BARTpho và ViT5

Trong quá trình thực nghiệm, mô hình BARTpho và ViT5 được huấn luyện nhằm tạo sinh văn bản cho từng văn bản trong cụm chủ . Trong đó dữ liệu đầu vào được mã hóa với độ dài tối đa (max length) là 1024 token. Quá trình huấn luyện được tiến hành với kích thước (batch size) là 4, sử dụng thuật toán tối ưu hóa AdamW với tốc độ học (learning rate) đặt là 2e-5. Mô hình được huấn luyện trong 10 epochs.

	batch size	max length	LR		
BARTpho$_{syllable}$	8	256	2e-5	AdamW	10 epochs
ViT5					

VII. Đánh giá thực nghiệm

A. Lý thuyết đánh giá

1) Độ đo đánh giá BERTopic: Để đánh giá hiệu quả của việc phân cụm chủ đề tin tức, chúng tôi sử dụng những chỉ tiêu đánh giá sau:

Density-Based Clustering Validation (DBCV)[25]Chỉ số đánh giá mức độ chặt chẽ và tách biệt của các cụm chủ đề được tạo ra. Giá trị của chỉ số thuộc [0; 1], càng gần 1 thì chất lượng phân cụm càng tốt.

Outlier Ratio[26] đánh giá tỷ lệ bài báo đầu vào không được phân loại vào bất kỳ chủ đề nào. Thông thường, mỗi sự kiện xảy ra trong thực tế thường được nhắc đến trong ít nhất một vài bài báo. Do đó, mà tỷ lệ này nên được tối ưu cho tối thiểu.

Adjusted Rand Index (ARI)[27] đo mức độ các cặp bài báo cùng nhãn được phân vào cùng một cụm chủ đề. Giá trị của chỉ số thuộc [0; 1], càng gần 1 thì các văn bản càng được gom lại thành một cụm chủ đề mà nó nhắc đến.

Normalized Mutual Information (NMI)[28] đo độ tương đồng của cụm được tạo ra với các cụm thực sự. Giá trị của chỉ số thuộc [0; 1], càng gần 1 thì chất lượng phân cụm càng gần với ngữ nghĩa "sự kiện"được định nghĩa thông qua tinh chỉnh embedder PhoBERT.

2) Độ đo đánh giá BARTpho và ViT5: Để đánh giá hiệu quả của mô hình tóm tắt, chúng tôi sử dụng các chỉ số phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên, bao gồm các độ đo sau:

Generated Length là độ dài trung bình của bản tóm tắt do mô hình sinh ra (số lượng token). Chỉ số này giúp đánh giá xem mô hình sinh bản tóm tắt có quá ngắn/dài so với thực tế không. Nếu độ dài quá ngắn có thể dẫn đến việc mất thông tin, ngược lại nếu văn bản tóm tắt quá dài có thể gây dư thừa thông tin đến với người đọc.

ROUGE Scores[29] là bộ độ đo phổ biến để tóm tắt văn bản tạo sinh và văn bản tham chiếu, bao gồm các độ đo nhỏ sau ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum

- ROUGE1 dùng để đo mức độ trùng khớp giữa các từ đơn để xem khả năng giữ các từ khóa quan trọng.
- ROUGE2 để đấnh giá mức độ trùng cặp từ liên tiếp để khả năng giữ cấu trúc ngữ nghĩa cơ bản trong câu.
- ROUGE3 sẽ đo mức độ trùng 3 từ liên tiếp. Điều này để đánh giá việc mô hình có duy trì được các cụm từ dài mang nhiều thông tin hay không. Tuy nhiên điều này thường khó xảy ra vì xác suất trùng 3 từ liên tiếp thường thấp.
- ROUGE-L dùng để đánh giá độ dài "chuỗi con chung"dài nhất. Vậy nên sẽ có xảy ra những trường hợp ROUGE1 cao vì giữ đúng từ nhưng ROUGE-L thấp hơn vì thông tin sai thứ tự. Chỉ số này quan trọng để đánh giá mức độ bảo toàn thứ tự thông tin quan trọng.
- ROUGE-SUM đo cặp từ liên tiếp, ROUGE-SUM chỉ yêu cầu cặp từ đúng thứ tự. Điều này linh hoạt hơn, cho phép mô hình tóm tắt mô hình có thể dùng từ chêm vào mà vẫn giữ ý chính và logic.

B. Kết quả đánh giá

1) BERTopic: Kết quả đánh giá BERTopic được trình bày ở bảng I

NMI = 0.9597 cho thấy mô hình tạo ra cụm có độ khớp cao với nhãn thật trong bộ dữ liệu. Tuy nhiên, ARI = 0.7968, ở mức khá tốt, thể hiện rằng vẫn còn một số bài báo bị lẫn vào cụm sự kiện khác. Như trong kết quả demo sau, một số bài báo thuộc sự kiện "Thân nhân phi hành đoàn MH17 kiện Malaysia Airlines"đang bị lẫn vào cụm sự kiện "Tai nạn máy bay của Ai Cập", có thể do "máy bay"và "tai nạn"là chủ đề chính của hai cụm sự kiện này.

DBCV	0.641
Outliner Ratio	0.0329
ARI	0.7968
NMI	0.9597

Bảng I
KẾT QUẢ THỰC NGHIỆM MÔ HÌNH BERTopic

{
"topic": 1,
"count": 15,
"docs": [
"Các nước chia buồn vụ tai nạn máy bay của Ai Cập. ...",
"Máy bay Ai Cập rơi: Những câu hỏi cho chính phủ Pháp và Ai Cập. ...",
"Giải mã bí ẩn máy bay rơi của EgyptAir. ...",
"Phát hiện mảnh vỡ nghi của máy bay MS804 gặp nạn. ...",
"Máy bay EgyptAir có thể bị tấn công bằng tên lửa đất đối không. ...",
"Danh tính một số nạn nhân mất tích cùng máy bay MS804 của Ai Cập. ...",
"Nữ tiếp viên máy bay mất tích từng đăng ảnh máy bay chìm. ...",
"Tiếp viên trên chuyến bay MS804 từng chụp ảnh cùng máy bay đang chìm. ...",
"Máy bay Ai Cập đột ngột chuyển hướng trước khi biến mất. ...",
"Bí ẩn bao trùm vụ máy bay EgyptAir mất tích ở Địa Trung Hải. ...",
"Gia đình thành viên phi hành đoàn MH17 kiện Malaysia Airlines. ...",
"Gia đình phi hành đoàn máy bay MH17 đâm đơn kiện. ...",
"Gia đình phi hành đoàn MH17 kiện hãng hàng không. ...",
"Gia đình phi hành đoàn máy bay MH17 kiện Malaysia Airlines. ...",
"Thân nhân phi hành đoàn MH17 kiện Malaysia Airlines. ..."]
}

Chất lượng của các cụm sự kiện này ở mức tốt (DBCV = 0.641). Tỷ lệ văn bản không được chia vào cụm chủ đề nào cũng đã được giảm xuống gần như bằng 0 (Outliner Ratio = 0.0329).

Điều này cho thấy BERTopic thực hiện tốt phần phân cụm chủ đề. Trong đó, nếu các sự kiện có tính chất và hướng các bài báo thảo luận của mỗi sự kiện rất giống nhau, thì vẫn còn khả năng mô hình không nhận ra sự khác biệt ngữ nghĩa và phân loại văn bản nhiễu vào cụm chủ đề. Kết quả này nêu ra một vấn đề nghiên cứu cách tinh chỉnh Embedder để hiểu ngữ nghĩa tốt hơn nữa.

2) BARTpho và ViT5: Kết quả đánh giá BERTopic được trình bày ở ảnh II

	ViT5	BARTpho$_{syllable}$
Generate Length	38.6506	41.0942
ROUGE1	56.0904	55.9397
ROUGE2	26.2323	26.9945
ROUGEL	36.4953	37.4335
ROUGELSUM	37.5475	38.8

Bảng II
KẾT QUẢ THỰC NGHIỆM MÔ HÌNH BARTPHO VÀ VIT5

Bảng kết quả cho thấy sự khác biệt về kết quả giữa hai mô hình. Chúng tôi sẽ phân tích chi tiết để có thể đánh giá để hiểu rõ hơn mức độ phù hợp của mô hình.

Ví dụ 1:

- ViT5: "Thị trưởng thành phố Tokyo, ông Yoichi Masuzoe, hôm nay (2/6) đã cúi đầu xin lỗi trước người dân vì bị cáo buộc tiêu xài xa xỉ."
- BARTpho: "Thị trưởng Tokyo ông Yoichi Masuzoe ngày 23/5 đã cúi đầu xin lỗi người dân vì những rắc rối tài chính gây ra trong thời gian gần đây."

Văn bản tham chiếu: "Thị trưởng Tokyo (Nhật Bản) Yoichi Masuzoe hứa sẽ không bay các chuyến hạng nhất và cắt giảm chi phí các chuyến đi nước ngoài."

Ví dụ 2:

- ViT5: "Cậu bé Yamato Tanooka, Nhật Bản được tìm thấy ngày 2-6 sau khi bị bố mẹ bỏ lại trong một khu rừng ở Hokkaido, Nhật Bản đã được tìm thấy và được đưa đến bệnh viện kiểm tra sức khỏe."
- BARTpho: "Sau khi bị bố mẹ bỏ lại trong một khu rừng trên đảo Hokkaido hôm 28-5, ông Takayuki Tanooka, ông bà của cậu bé Yamato đã được tìm thấy gần nơi cậu bé bị bắt giam."

Văn bản tham chiếu: "Cậu bé Yamato Tanooka, 7 tuổi, bị bố mẹ phạt bỏ trong rừng rồi mất tích gần cả tuần vừa được tìm thấy."

Ví dụ 3:

- ViT5: "Coca-Cola tuyên bố, hãng sẽ ngừng sản xuất ở Venezuela do thiếu đường."
- BARTpho: "Coca-Cola cho biết, hãng sẽ tiếp tục sản xuất các đồ uống không đường như Coca-Cola dành cho người ăn kiêng."

Văn bản tham chiếu: "Hãng giải khát CocaCola sẽ tạm ngừng sản xuất nước ngọt ở Venezuela do tình trạng thiếu đường tại quốc gia Nam Mỹ này."

⇒ Đánh giá: ViT5 ghi nhận thông tin quan trọng và chính xác hơn BARTpho. ViT5 có xu hướng tạo tóm tắt ngắn hơn BARTpho. Đây thể là lý do ROUGE1 ViT5 cao hơn nhưng các chỉ số khác không bằng BARTpho.

Ví dụ 4:

- ViT5: "Đường hầm tàu Gothard xuyên dãy núi Alpe dài nhất thế giới."
- BARTpho: "Ngày 1/6, giới chức Thuỵ Sĩ sẽ khánh thành đường hầm tàu hỏa Gothard xuyên dãy núi Alpe dài nhất thế giới (57,104 km), nối hai TP Erstfeld và Bodio vào ngày 1/6."

Văn bản tham chiếu: "Ngày 1/6, giới chức Thuỵ Sĩ sẽ khánh thành đường hầm tàu hỏa Gothard xuyên dãy núi Alpe dài nhất thế giới."

Ví dụ 5:

- ViT5: "Tình hình kinh tế Venezuela đang trở nên bất ổn định."
- BARTpho: "Tình hình tại Venezuela đang diễn biến theo chiều hướng tồi tệ hơn, thậm chí viễn cảnh đất nước này sụp đổ đã trở nên xấu hơn."

Văn bản tham chiếu: "Tình hình ở Venezuela đang diễn biến theo chiều hướng xấu trở nên xấu hơn."

Ví dụ 6:

- ViT5: "Hãng hàng không giá rẻ VietJet Air vừa ký hợp đồng đặt mua 100 máy bay Boeing 737 MAX 200, trị giá 11,3 tỉ USD."
- BARTpho: "Hãng hàng không VietJet Air vừa ký hợp đồng đặt mua 100 máy bay Boeing 737 MAX 200, trị giá 11,3 tỉ USD."

Văn bản tham chiếu: "Hãng hàng không VietJet Air ký hợp đồng mua 100 máy bay Boeing 737 MAX 200 trong hôm nay 23.5, giữa lúc Tổng thống Mỹ Barack Obama thăm Việt Nam."

⇒ Đánh giá: BARTpho có xu hướng học cụm từ gần với label hơn, nên kết quả ROUGE2, ROUGE3 cao hơn nhẹ.

TÀI LIỆU

[1] L. George **and** P. Sumathy, *An Integrated Clustering and BERT Framework for Improved Topic Modeling*, **august** 2022. DOI: 10.21203/rs.3.rs-1986180/v1.

[2] D. T. Tuan, D. Van Thin, V.-H. Pham **and** N. L.-T. Nguyen, "Vietnamese Facebook Posts Classification using Fine-Tuning BERT," **in***2020 7th NAFOSTED Conference on Information and Computer Science (NICS)* 2020, **pages** 314–319. DOI: 10 . 1109 / NICS51282.2020.9335865.

[3] "Association for Vietnamese Language and Speech Processing." (), **url**: https://vlsp.org.vn/.

[4] Y. Zhang, H. Jin, D. Meng, J. Wang **and** J. Tan, *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*, 2025. arXiv: 2403 . 02901 [cs.AI]. **url**: https://arxiv.org/abs/2403.02901.

[5] N.-T. Tran, M.-Q. Nghiem, N. T. Nguyen, N. L.-T. Nguyen, N. Van Chi **and** D. Dinh, "ViMs: a high-quality Vietnamese dataset for abstractive multi-document summarization," *Language Resources and Evaluation*, **jourvol** 54, **number** 4, **pages** 893–920, 2020.

[6] "Vnexpress." (), **url**: https://vnexpress.net/.

[7] "Báo dân trí." (), **url**: https://dantri.com.vn/.

[8] "Báo tuổi trẻ." (), **url**: https://tuoitre.vn/.

[9] M. Grootendorst, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*, 2022. arXiv: 2203 . 05794 [cs.CL]. **url**: https://arxiv.org/abs/2203.05794.

[10] D. Blei, A. Ng **and** M. Jordan, "Latent Dirichlet Allocation," **volume** 3, **january** 2001, **pages** 601–608.

[11] W. Xu, X. Liu **and** Y. Gong, "Document clustering based on non-negative matrix factorization," **in***Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* **jourser** SIGIR '03, Toronto, Canada: Association for Computing Machinery, 2003, **pages** 267–273, ISBN: 1581136463. DOI: 10.1145/860435.860485. **url**: https://doi.org/10.1145/860435.860485.

[12] D. Q. Nguyen **and** A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," **in***Findings of the Association for Computational Linguistics: EMNLP 2020* 2020, **pages** 1037–1042.

[13] L. McInnes, J. Healy **and** J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2020. arXiv: 1802.03426 [stat.ML]. **url**: https://arxiv.org/abs/1802.03426.

[14] L. McInnes, J. Healy **and** S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, **jourvol** 2, **march** 2017. DOI: 10.21105/joss.00205.

[15] D. Wang, X. Lu **and** A. Rinaldo, *DBSCAN: Optimal Rates For Density Based Clustering*, 2019. arXiv: 1706.03113 [math.ST]. **url**: https://arxiv.org/abs/1706.03113.

[16] P. Bafna, D. Pramod **and** A. Vaidya, "Document clustering: TF-IDF approach," **in***2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* 2016, **pages** 61–66. DOI: 10 . 1109/ICEEOT.2016.7754750.

[17] J. Devlin, M.-W. Chang, K. Lee **and** K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019. arXiv: 1810.04805 [cs.CL]. **url**: https://arxiv.org/abs/1810.04805.

[18] M. Lewis, Y. Liu, N. Goyal **andothers**, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2019. arXiv: 1910.13461 [cs.CL]. **url**: https://arxiv.org/abs/1910.13461.

[19] Y. Liu, J. Gu, N. Goyal **andothers**, *Multilingual Denoising Pre-training for Neural Machine Translation*, 2020. arXiv: 2001.08210 [cs.CL]. **url**: https://arxiv.org/abs/2001.08210.

[20] T. Kudo **and** J. Richardson, *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*, 2018. arXiv: 1808.06226 [cs.CL]. **url**: https://arxiv.org/abs/1808.06226.

[21] A. Conneau, K. Khandelwal, N. Goyal **andothers**, *Unsupervised Cross-lingual Representation Learning at Scale*, 2020. arXiv: 1911.02116 [cs.CL]. **url**: https://arxiv.org/abs/1911.02116.

[22] L. Phan, H. Tran, H. Nguyen **and** T. H. Trinh, "ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation," **in***Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop* D. Ippolito, L. H. Li, M. L. Pacheco, D. Chen **and** N. Xue, **editors**, Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, **july** 2022, **pages** 136–142. DOI: 10.18653/v1/2022.naacl-srw.18. **url**: https://aclanthology.org/2022.naacl-srw.18/.

[23] C. Raffel, N. Shazeer, A. Roberts **andothers**, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, 2023. arXiv: 1910.10683 [cs.LG]. **url**: https://arxiv.org/abs/1910.10683.

[24] A. Conneau, K. Khandelwal, N. Goyal **andothers**, *Unsupervised Cross-lingual Representation Learning at Scale*, 2020. arXiv: 1911.02116 [cs.CL]. **url**: https://arxiv.org/abs/1911.02116.

[25] D. Moulavi, P. Andretta Jaskowiak, R. Campello, A. Zimek **and** J. Sander, "Density-Based Clustering Validation," **april** 2014. DOI: 10.1137/1.9781611973440.96.

[26] D. R. Bull **and** F. Zhang, "Chapter 10 - Measuring and managing picture quality," **in***Intelligent Image and Video Compression (Second Edition)* D. R. Bull **and** F. Zhang, **editors**, Second Edition, Oxford: Academic Press, 2021, **pages** 335–384, ISBN: 978-0-12-820353-8. DOI: https://doi.org/10.1016/B978-0-12-820353-8.00019-0. **url**: https://www.sciencedirect.com/science/article/pii/B9780128203538000190.

[27] Y. Yang, "Chapter 3 - Temporal Data Clustering," **in***Temporal Data Mining Via Unsupervised Ensemble Learning* Y. Yang, **editor**, Elsevier, 2017, **pages** 19–34, ISBN: 978-0-12-811654-8. DOI: https://doi.org/10.1016/B978-0-12-811654-8.00003-8. **url**: https://www.sciencedirect.com/science/article/pii/B9780128116548000038.

[28] A. F. McDaid, D. Greene **and** N. Hurley, *Normalized Mutual Information to evaluate overlapping community finding algorithms*, 2013. arXiv: 1110.2515 [physics.soc-ph]. **url**: https://arxiv.org/abs/1110.2515.

[29] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," **in***Text Summarization Branches Out* Barcelona, Spain: Association for Computational Linguistics, **july** 2004, **pages** 74–81. **url**: https://aclanthology.org/W04-1013/.