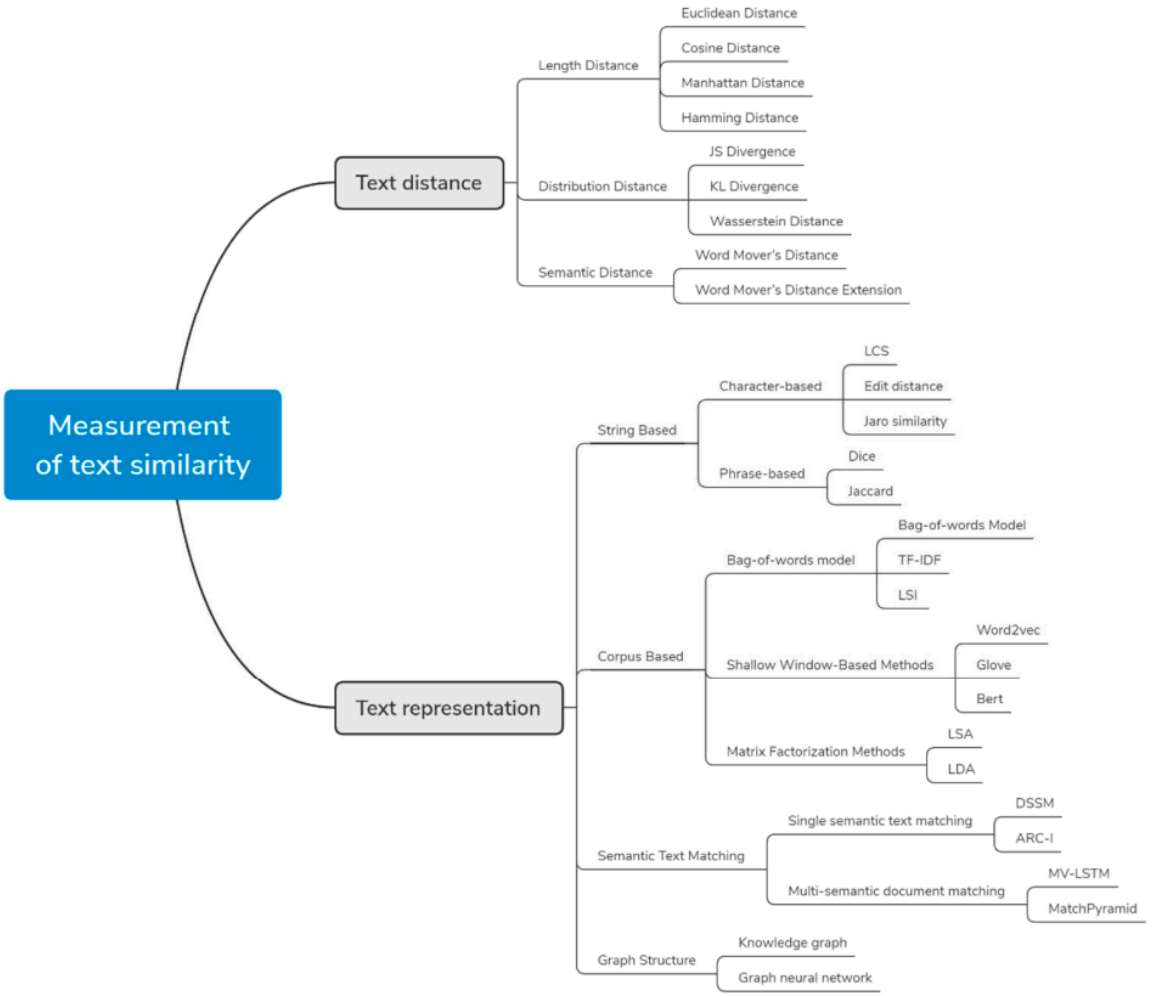
Báo cáo tổng quan về Semantic Textual Relatedness và Measurement of Text Similarity



[(PDF) Measurement of Text Similarity: A Survey (researchgate.net)]

Text Distance

Length Distance: sử dụng các đặc trưng số học để tính toán. Điều kiện tiên quyết là chúng ta vẫn phải biến đổi văn bản từ dạng chữ sang dạng số học để tính toán, mang lại hiệu quả kém bởi những lý do sau:

- Thứ nhất, nó phù hợp với các bài toán đối xứng, chẳng hạn Sim(A,B) = Sim(B,A), nhưng đối với câu hỏi: Liệu A có phải là câu trả lời của Q thì độ tương tự tương ứng là không đối xứng.
- Thứ hai, có rủi ro khi sử dụng độ dài và khoảng cách để đánh giá mức độ tương tự mà không biết các đặc tính thống kê của dữ liệu.

Distribution distance: Hai phương pháp được sử dụng phổ biến nhất hiện nay là: Jensen–Shannon divergence và Kullback–Leibler divergence

Semantic Distance: Khi không có từ chung trong văn bản, độ tương tự thu được bằng cách sử dụng thước đo khoảng cách dựa trên độ dài hoặc phân bổ có thể tương đối nhỏ, vì vậy chúng ta có thể xem xét tính khoảng cách tại mức độ ngữ nghĩa với **Word Mover's Distance**

Text Representation

String-based: có thể tính toán đơn giản dựa trên chuỗi ký tự:

Character-based: LCS (longest common substring), editing distance, Jaro similarity
Phrase-Based: dice coefficient, Jaccard

Corpus-Based:

Bag-of-Words Model: BOW (bag of words), TF-IDF (term frequency–inverse document frequency), LSA (latent semantic analysis)
Shallow Window-Based Methods: Word2vec and glove, BERT
Matrix Factorization Methods: LSA và biến thể PLSA, LDA (latent dirichlet allocation)
Skip-Thought Vectors

Based on Graph Structure: Knowledge Graph và Graph Neural Network

Structure Based Sentence Similarity

Các phương pháp phổ biến được sử dụng là Grammar Based, Part of Speech POS, Using Word Order https://arxiv.org/pdf/1910.03940

Đối với SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages, có bài đăng liên quan sau đây mô tả cách tiếp cận và thực hiện của các nhóm: 2404.01490 (arxiv.org) , 2407.12426 (arxiv.org) , 2405.00659 (arxiv.org) , 2404.04513 (arxiv.org) , 2404.01860 (arxiv.org) ,

[2404.02570] MaiNLP at SemEval-2024 Task 1: Analyzing Source Language Selection in Cross-Lingual Textual Relatedness (arxiv.org)

[2403.14990] MasonTigers at SemEval-2024 Task 1: An Ensemble Approach for Semantic Textual Relatedness (arxiv.org)

2024.semeval-1.13.pdf (aclanthology.org) (nhóm người việt)

[2403.18933] SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages (arxiv.org) - Đây là bài viết tổng kết và đánh giá cho cả cuộc thi.

Xu hướng chung của các phương pháp được đưa ra cho phần huấn luyện supervised là (1) embedding các cặp câu vào văn bản và (2) huấn luyện mô hình hồi quy. Một số nhóm đã sử dụng phương pháp nhúng và hồi quy truyền thống các phương pháp tiếp cận (ví dụ: word2vec với Support vector regression (SVR)). Phần lớn sử dụng các phương pháp học sâu (ví dụ: BERT, RoBERTa) hoặc các mô hình pre-trained transformer lớn khác

2404.09047 (arxiv.org) sử dụng SVR

AAdaM Họ đã chọn tăng cường dữ liệu bằng cách dịch tập dữ liệu SemRel tiếng Anh và STSB (tương tự về ngữ nghĩa) để tạo và tăng cường dữ liệu bằng các ngôn ngữ khác. Nhóm đã khám phá cả fine-tune và adapter-based tuning dựa trên bộ điều hợp của mô hình AfroXLMR cross-encoder

NRK Họ đã tập hợp nhiều mô hình BERT khác nhau và sử dụng kỹ thuật bỏ phiếu có trọng số để cải thiện hiệu suất của mô hình của họ.

STR giúp đánh giá các LLM ở các phương diện sau:
Semantic similarity evaluation of LLMs (linkedin.com)

1. Hiểu bối cảnh và ý nghĩa
STR giúp đánh giá mức độ LLM hiểu ngữ cảnh và ý nghĩa của văn bản đầu vào. Điều này rất quan trọng đối với các nhiệm vụ như:

Diễn giải: Xác định xem mô hình có thể tạo ra các câu khác nhau có cùng ý nghĩa hay không.
Tóm tắt: Kiểm tra xem bản tóm tắt có nắm bắt chính xác bản chất của văn bản gốc hay không.
Trả lời câu hỏi: Đảm bảo rằng các câu trả lời do mô hình cung cấp có liên quan về mặt ngữ nghĩa với các câu hỏi được đặt ra.

2. Đo lường sự tương đồng
STR có thể được sử dụng để đo lường mức độ tương tự giữa văn bản được tạo và văn bản tham chiếu. Các kỹ thuật bao gồm:

Tương tự Cosine: So sánh cách biểu diễn vectơ của các câu.
Điểm BLEU, ROUGE, METEOR: Số liệu truyền thống đo lường sự trùng lặp của các từ và cụm từ.
BERTScore: Một số liệu gần đây hơn sử dụng các phần nhúng từ BERT để đánh giá sự tương đồng ở cấp độ ngữ nghĩa.

3. Đánh giá tính mạch lạc, nhất quán
LLM phải tạo ra văn bản mạch lạc và nhất quán. STR giúp đánh giá:

Hệ thống đối thoại: Đảm bảo các câu trả lời phù hợp với ngữ cảnh và liên quan đến ngữ nghĩa với cuộc trò chuyện trước đó.
Tạo câu chuyện: Đánh giá xem câu chuyện có còn mạch lạc xuyên suốt hay không

Các thang đo phổ biến được sử dụng:

Instruction:	Give some examples of what people usually say when someone arrives safely
Target Response:	Glad you made it safe and sound.
Model Response:	Thank goodness you arrived without any issues.
ROUGE-L:	0.143
BLEU:	6.57
Human Rating:	**A (best)**

Ý tưởng đằng sau SemScore rất đơn giản nhưng hiệu quả: tính toán độ tương tự cosine giữa hai câu. SemScore: Evaluating LLMs with Semantic Similarity (huggingface.co), 2401.17072 (arxiv.org) dùng all-mpnet-base-v2 và cosine score để đánh giá.

Metric	τ	r
SEMSCORE	**0.879**	**0.970**
*G-Eval-4**	0.855	0.863
*G-Eval-3.5**	0.855	0.831
BERTScore	0.848	0.944
G-Eval-3.5-instruct	0.840	0.911
ROUGE-L	0.788	0.933
BARTScore	0.788	0.621
BARTScore$_{para}$	0.697	0.884
BLEU	0.667	0.865
BLEURT	0.485	0.485
DiscoScore	0.364	0.583

SemScore và G-Eval-4* có cách đánh giá tương tự giống con người nhất. Nhưng thang đo tiêu chuẩn vẫn xuất phát từ con người.

Tìm teen cho khóa luận, đã có bộ dữ liệu tiesgn việt, chuẩn hóa lại dữ liệu

PhoBert, v2. llama, simeSE,
Xây dựng bộ dữ liệu chuẩn có thống kê, độ đo đồng thuận

thiết lập trước cách độ đo cơ bản trên git, share tài liệu