1. **Definition**

Machine translation is the process of using artificial intelligence to automatically translate text from one language to another without human involvement.

Dịch máy là quá trình sử dụng trí tuệ nhân tạo để tự động dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác mà không cần sự tham gia của con người.

**Advantage**

Machine translation works very fast, translating millions of words almost instantaneously. It can translate large amounts of data, such as real-time chat or large-scale legal cases.

Dịch vụ dịch máy hoạt động rất nhanh, dịch hàng triệu từ gần như ngay lập tức. Dịch vụ này có thể dịch một lượng lớn dữ liệu, chẳng hạn như cuộc trò chuyện theo thời gian thực hoặc các vụ án tố tụng có quy mô lớn.

Machine translation increases productivity and the ability to deliver translations faster, reducing the time to market. There is less human involvement in the process as machine translation provides basic but valuable translations, reducing both the cost and time of delivery. For example, in high-volume projects, you can integrate machine translation with your content management systems to automatically tag and organize the content before translating it to different languages.

Dịch máy làm tăng hiệu suất và giúp cung cấp bản dịch nhanh hơn, giảm thiểu thời gian đưa ra thị trường. Do sự tham gia của con người vào quá trình dịch ít hơn nên dịch vụ dịch máy cung cấp các bản dịch cơ bản nhưng cũng có giá trị, giúp giảm thiểu cả chi phí và thời gian trả dự án. Ví dụ: trong các dự án lớn, bạn có thể tích hợp dịch vụ dịch máy với hệ thống quản lý nội dung của mình để tự động gắn thẻ và sắp xếp nội dung trước khi dịch nội dung đó sang các ngôn ngữ khác.

Disadvantages

MT can’t account for certain phrases because of lack of context

It’s difficult for machine translation to accurately translate nuances and slang

Complicated or industry specific terms (i.e. medical terminology) may not be easily translated

Specific errors are hard to predict and difficult to correct

Content in the target language can feel choppy or pieced together

BERT, short for Bidirectional Encoder Representations from Transformers, is a Machine Learning (ML) model for natural language processing. It was developed in 2018 by researchers at Google AI Language and serves as a swiss army knife solution to 11+ of the most common language tasks, such as sentiment analysis and named entity recognition.

BERT can be used on a wide variety of language tasks:

Can determine how positive or negative a movie’s reviews are. (Sentiment Analysis)

Helps chatbots answer your questions. (Question answering)

Predicts your text when writing an email (Gmail). (Text prediction)

Can write an article about any topic with just a few sentence inputs. (Text generation)

Can quickly summarize long legal contracts. (Summarization)

Can differentiate words that have multiple meanings (like ‘bank’) based on the surrounding text. (Polysemy resolution)

[BERT](https://arxiv.org/pdf/1810.04805.pdf) là viết tắt của cụm từ Bidirectional Encoder Representation from Transformer có nghĩa là mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer. BERT được thiết kế để huấn luyện trước các biểu diễn từ (pre-train word embedding). Điểm đặc biệt ở BERT đó là nó có thể điều hòa cân bằng bối cảnh theo cả 2 chiều trái và phải.

Cơ chế attention của Transformer sẽ truyền toàn bộ các từ trong câu văn đồng thời vào mô hình một lúc mà không cần quan tâm đến chiều của câu. Do đó Transformer được xem như là huấn luyện hai chiều (bidirectional) mặc dù trên thực tế chính xác hơn chúng ta có thể nói rằng đó là huấn luyện không chiều (non-directional). Đặc điểm này cho phép mô hình học được bối cảnh của từ dựa trên toàn bộ các từ xung quanh nó bao gồm cả từ bên trái và từ bên phải.

**How does BERT Work?**

**Large amounts of training data**

BERT was specifically trained on Wikipedia (~2.5B words) and Google’s BooksCorpus (~800M words). These large informational datasets contributed to BERT’s deep knowledge not only of the English language but also of our world!

2.2 What is a Masked Language Model?

MLM enables/enforces bidirectional learning from text by masking (hiding) a word in a sentence and forcing BERT to bidirectionally use the words on either side of the covered word to predict the masked word

Biến word thành vector