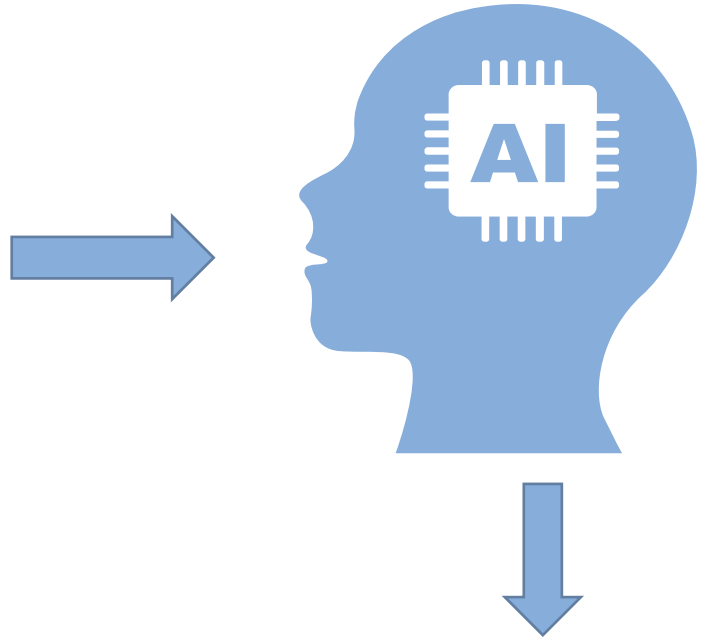
Extract Information From Contract

Using BERT (Deep Learning , NLP)

Extract Information From Contract

Contract

\ufeff_____ .. . II . สัญญาเลขที่ TMV/CM212015-1575/BBB/ys Project : FTTX Fund Code : ..... Fund Center : สัญญาว่าจ้างสร้าง และ/หรือ ปรับปรงข่ายสายสัญญาณโทรคมนาคม สัญญาฉบับนี้ทำขึ้นเมื่อวันที่ 0 1 .ค. 2558 ณบริษัท ทรู มูฟ จำกัดระหว่าง (1) บริษัท ทรู มูฟ จำกัด โดย นายศุภชัย เจียรวนนท์ และศาสตราจารย์พิเศษ อสึก อัศวานันท์ กรรมการผู้มีอำนาจ สำนักงานตั้งอยู่เลขที่ 18 อาคารทรูทาวเวอร์ ถนนรัชดาภิเษก แขวงห้วยขวาง เขตห้วยขวาง กรุงเทพมหานคร ซึ่งต่อไปนี้ จะเรียกว่า "ผู้ว่าจ้าง" ฝ่ายหนึ่ง กับ อพ (2) ห้างหุ้นส่วนจำกัด บรอดแบนด์ โดยนายปัญญาวุฒิ พลราช หุ้นส่วนผู้จัดการสำนักงานตั้งอยู่เลขที่ 154 หมู่ ที่ 12 ตำบลกุดลาด อำเภอเมืองอุบลราชธานี จังหวัดอุบลราชธานี ซึ่งต่อไปนี้จะเรียกว่า "ผู้รับจ้าง"อีกฝ่ายหนึ่ง โซกา นง CA2-007 หน้า 19 Standard Form 2015



Result

ห้างหุ้นส่วนจำกัด หรือ บริษัทใดเป็นผู้รับจ้าง ?

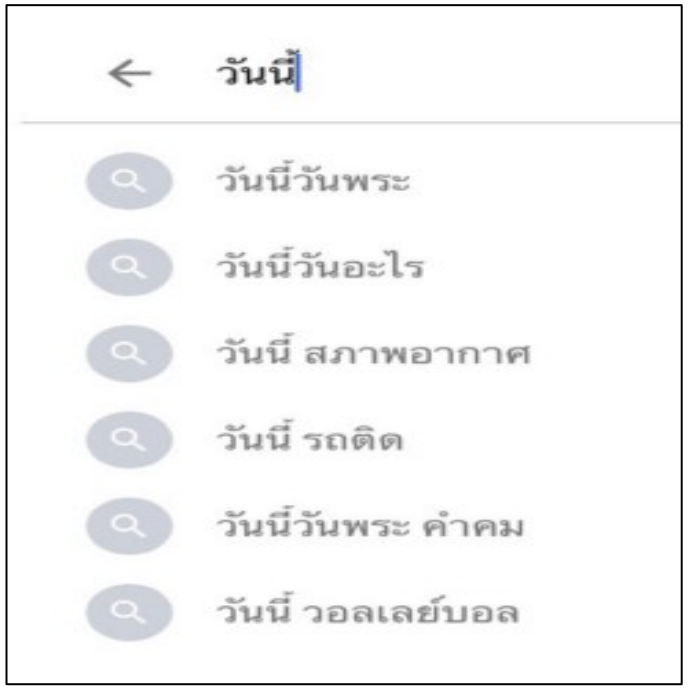
ห้างหุ้นส่วนจำกัด บรอดแบนด์

What is Natural Language Processing?

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

Chatbot

Search Engine

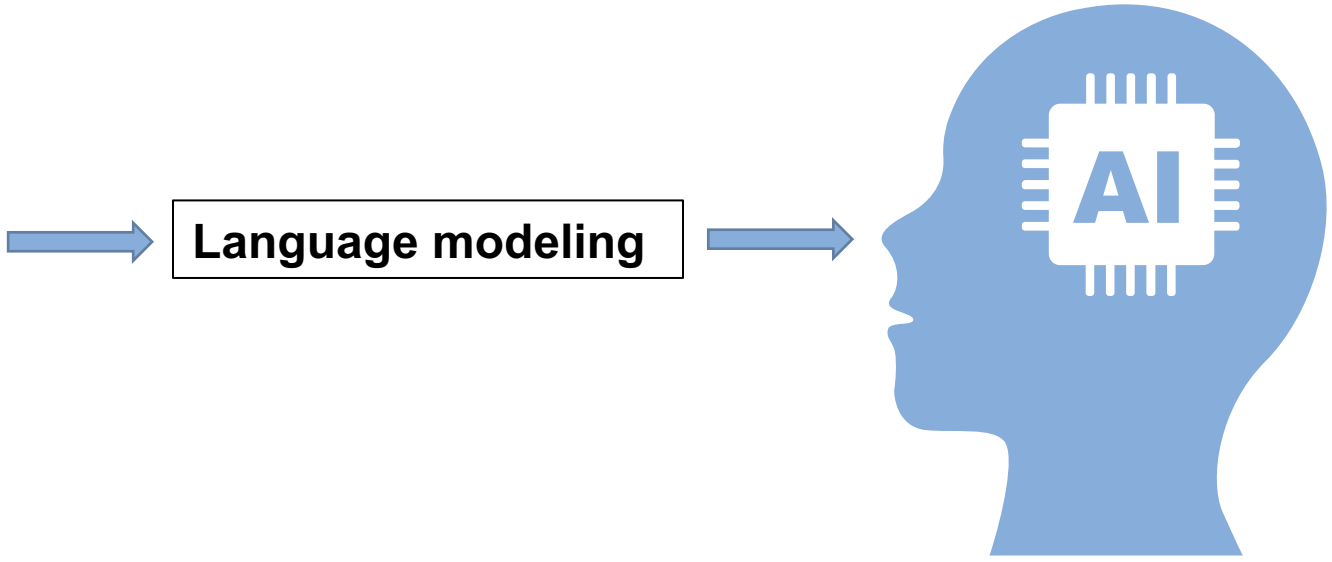


Translator

How do computers understand language?

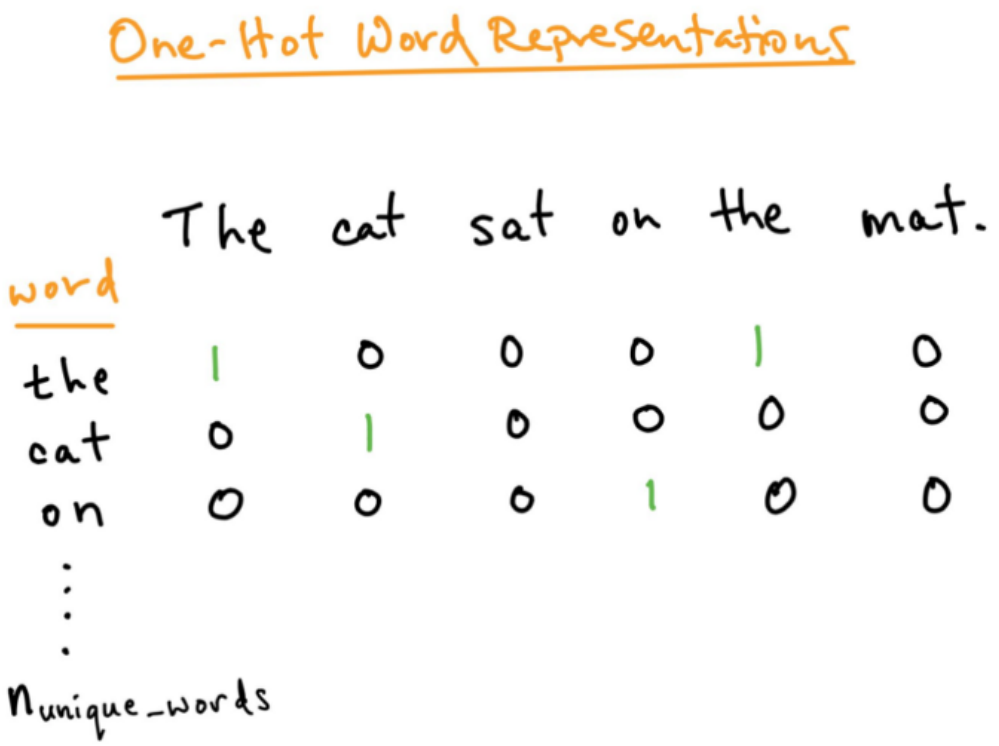
A statistical language model is a probability distribution over sequences of words. The language model provides context to distinguish between words and phrases that sound similar. For example, in American English, the phrases "recognize speech" and "wreck a nice beach" sound similar, but mean different things.

\ufeff_____ .. . II . สัญญาเลขที่ TMV/CM212015-1575/BBB/ys Project : FTTX Fund Code : ..... Fund Center : สัญญาว่าจ้างสร้าง และ/หรือ ปรับปรุงข่ายสายสัญญาณโทรคมนาคม สัญญาฉบับนี้ทำขึ้นเมื่อวันที่ 0 1 .ค. 2558 ณบริษัท ทรู มูฟ จำกัดระหว่าง (1) บริษัท ทรู มูฟ จำกัด โดย นายศุภชัย เจียรวนนท์ และศาสตราจารย์พิเศษ อสึก อัศวานันท์ กรรมการผู้มีอำนาจ สำนักงานตั้งอยู่เลขที่ 18 อาคารทรู ทาวเวอร์ ถนนรัชดาภิเษก แขวงห้วยขวาง เขตห้วยขวาง กรุงเทพมหานคร ซึ่งต่อไปนี้ จะเรียกว่า "ผู้ว่าจ้าง" ฝ่ายหนึ่ง กับ อพ (2) ห้างหุ้นส่วนจำกัด บรอดแบนด์ โดยนายปัญญาวุฒิ พลราช หุ้นส่วนผู้จัดการสำนักงานตั้งอยู่เลขที่ 154 หมู่ ที่ 12 ตำบลกุดลาด อำเภอเมืองอุบลราชธานี จังหวัดอุบลราชธานี ซึ่งต่อไปนี้จะเรียกว่า "ผู้รับจ้าง"อีกฝ่ายหนึ่ง โชกา นง CA2-007 หน้า 19 Standard Form 2015



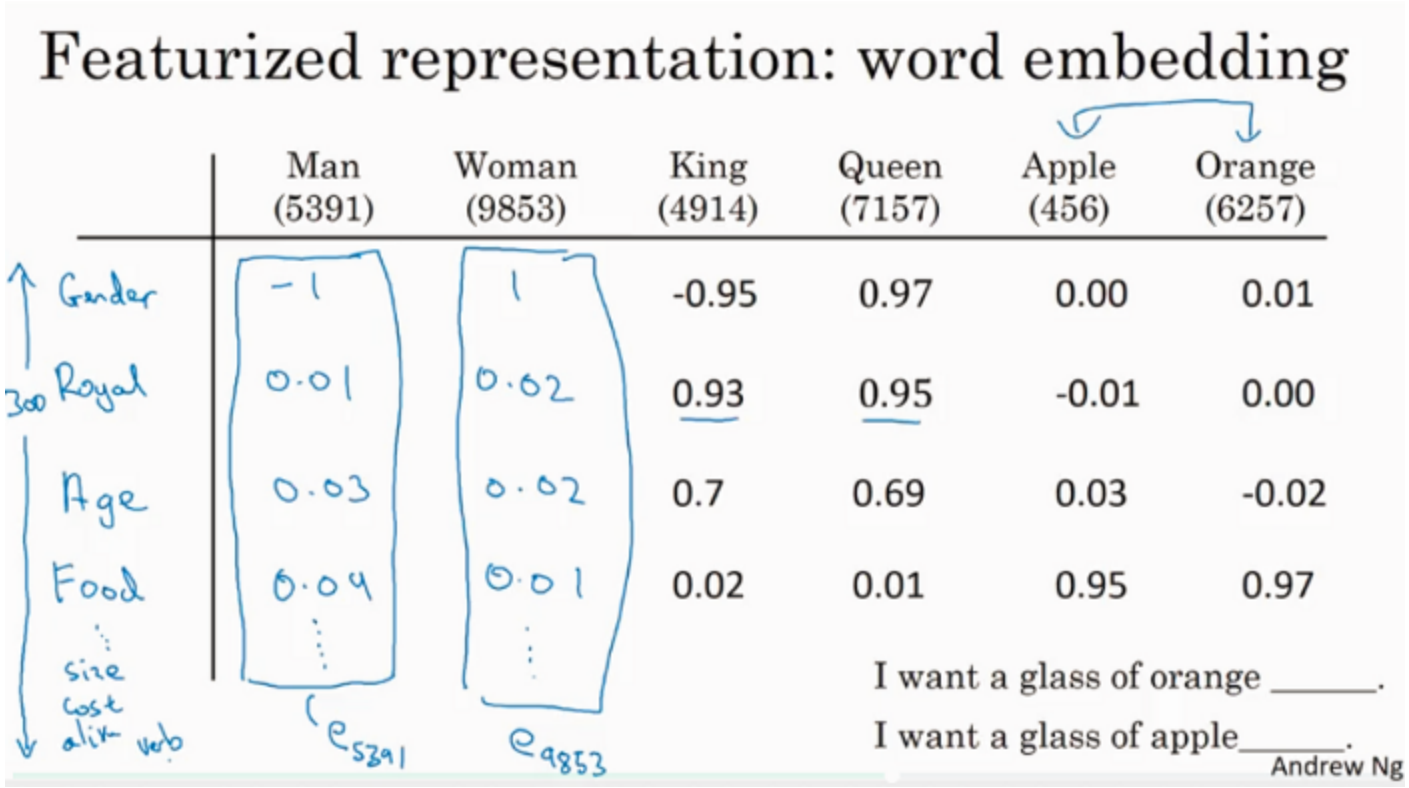
One-hot

One hot encodings are another way of representing words in numeric form. The length of the word vector is equal to the length of the vocabulary, and each observation is represented by a matrix with rows equal to the length of vocabulary and columns equal to the length of observation, with a value of 1 where the word of vocabulary is present in the observation and a value of zero where it is not.

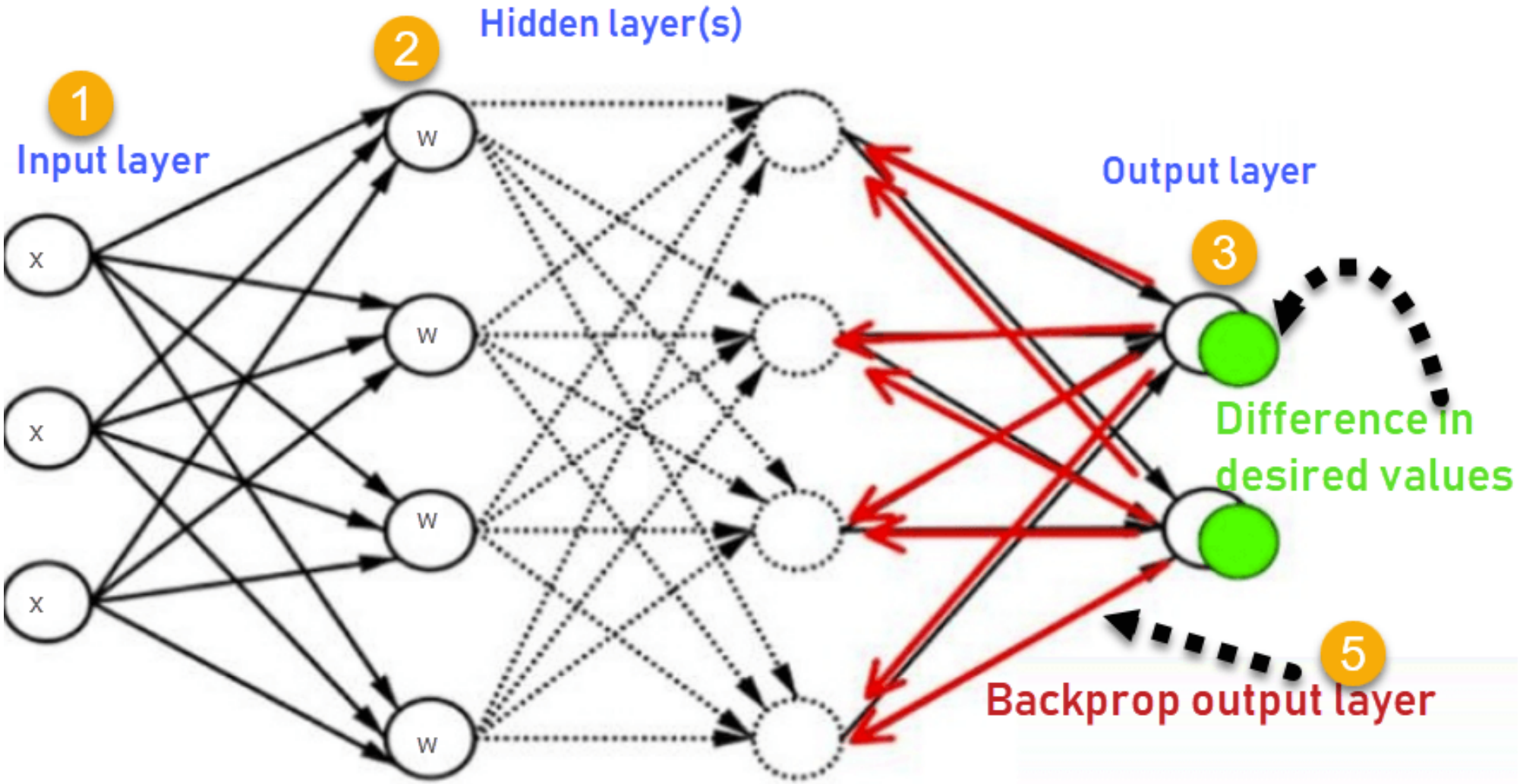


Word Embedding

Word embedding is the collective name for a set of language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers. The technique is primarily used with Neural Network Models.

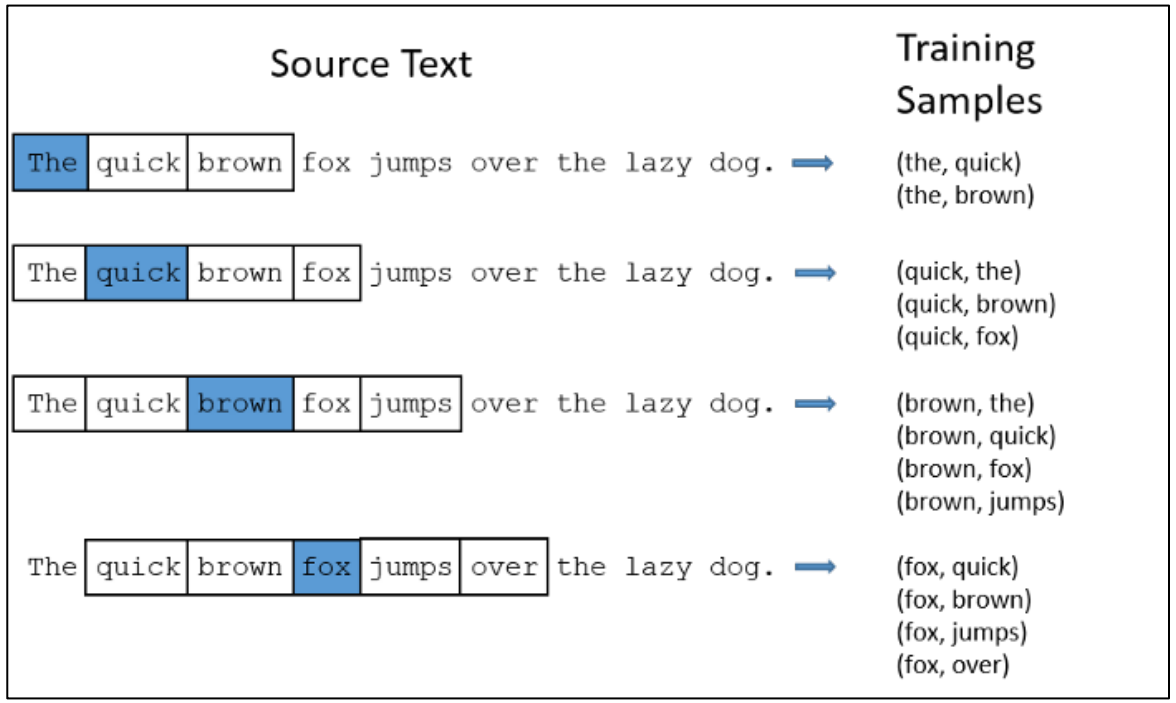


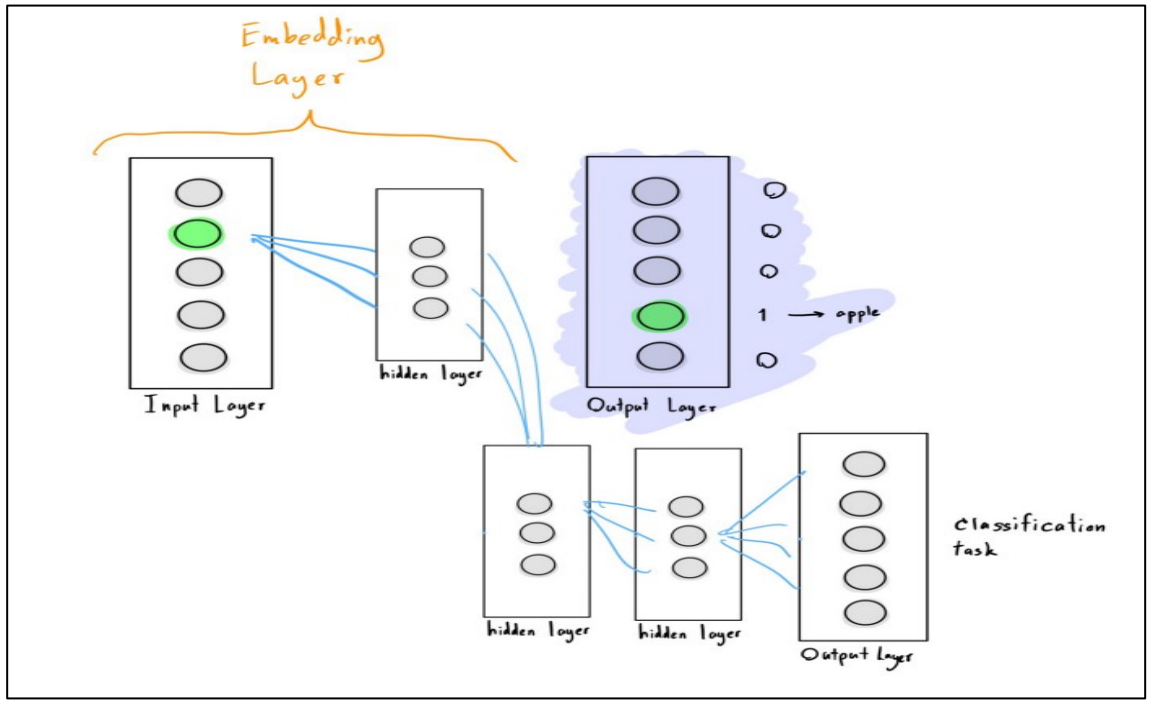
Neural Network Model



Word Embedding

skip-gram

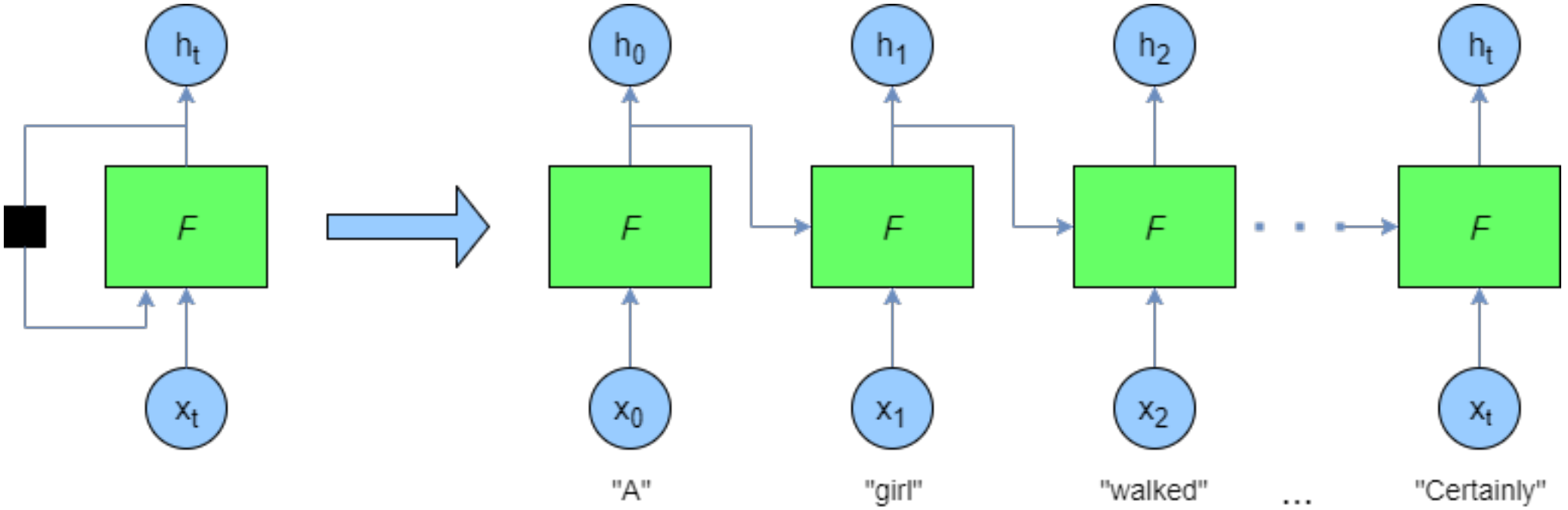




- **Skip-Gram** — ใช้ Context 1 คำ เพื่อหา next word หลายคำ

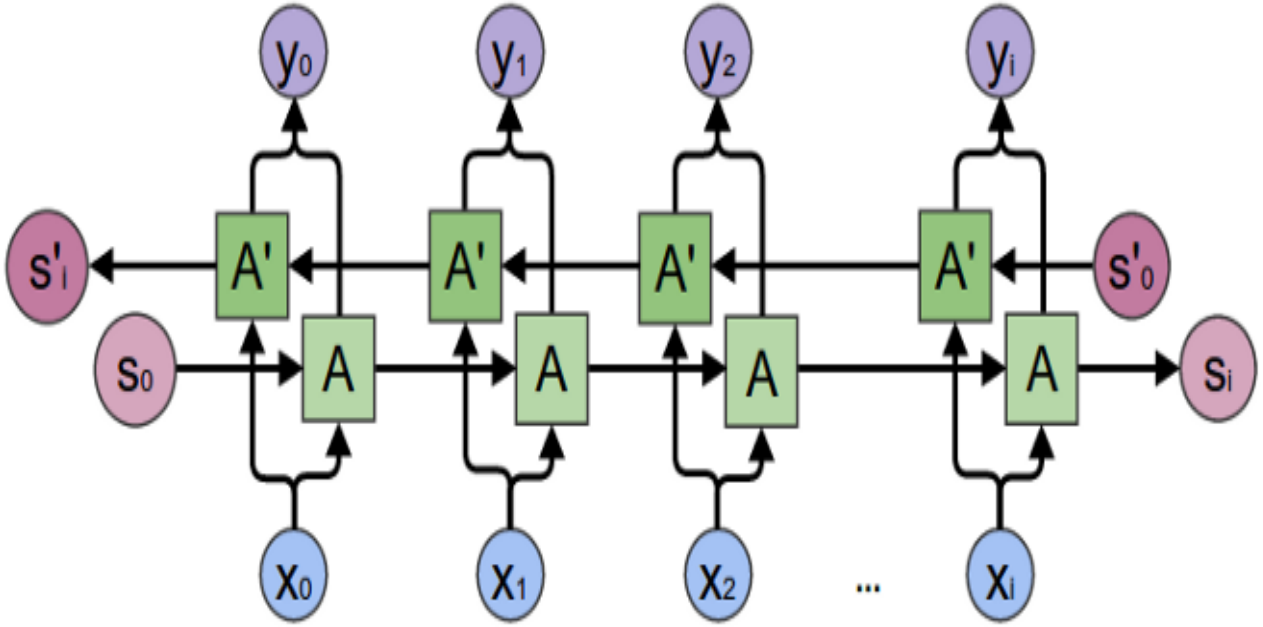
Recurrent Neural Networks (RNN)

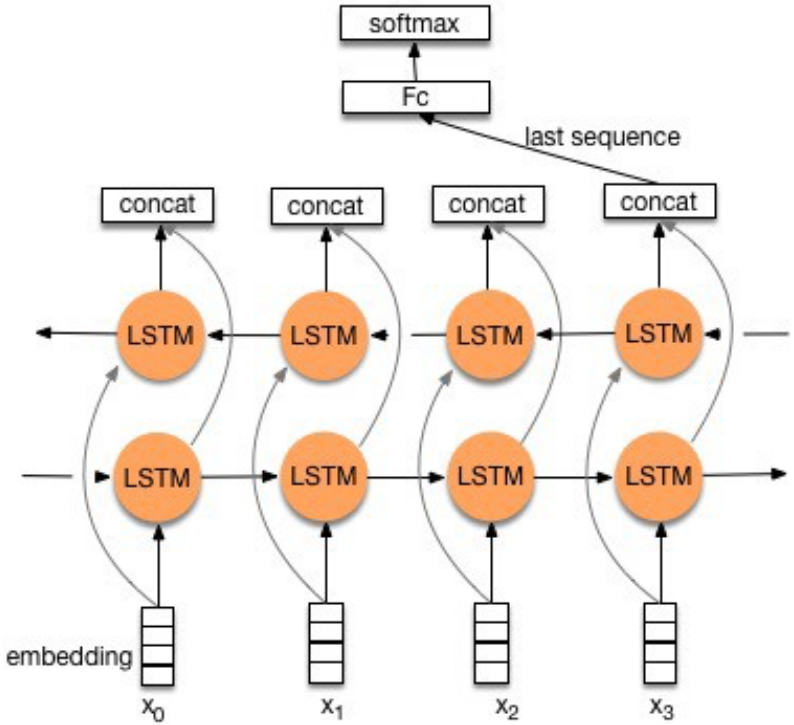
Conceptually they differ from a standard neural network as the standard input in a RNN is a word instead of the entire sample. This gives the flexibility for the network to work with varying lengths of sentences. It also provides an additional advantage of sharing features learned across different positions of text which can not be obtained in a standard neural network.



Bidirectional RNN

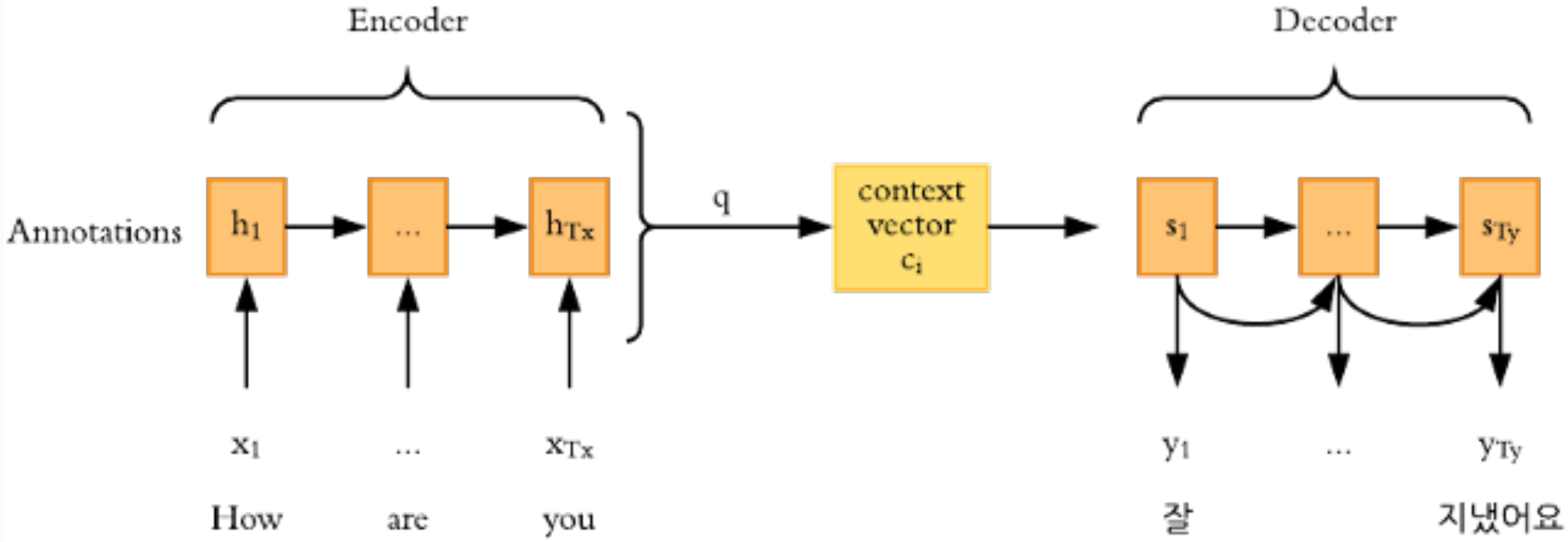
Bidirectional recurrent neural networks(RNN) are really just putting two independent RNNs together. The input sequence is fed in normal time order for one network, and in reverse time order for another. The outputs of the two networks are usually concatenated at each time step. This structure allows the networks to have both backward and forward information about the sequence at every time step.





Sequence-to-sequence model

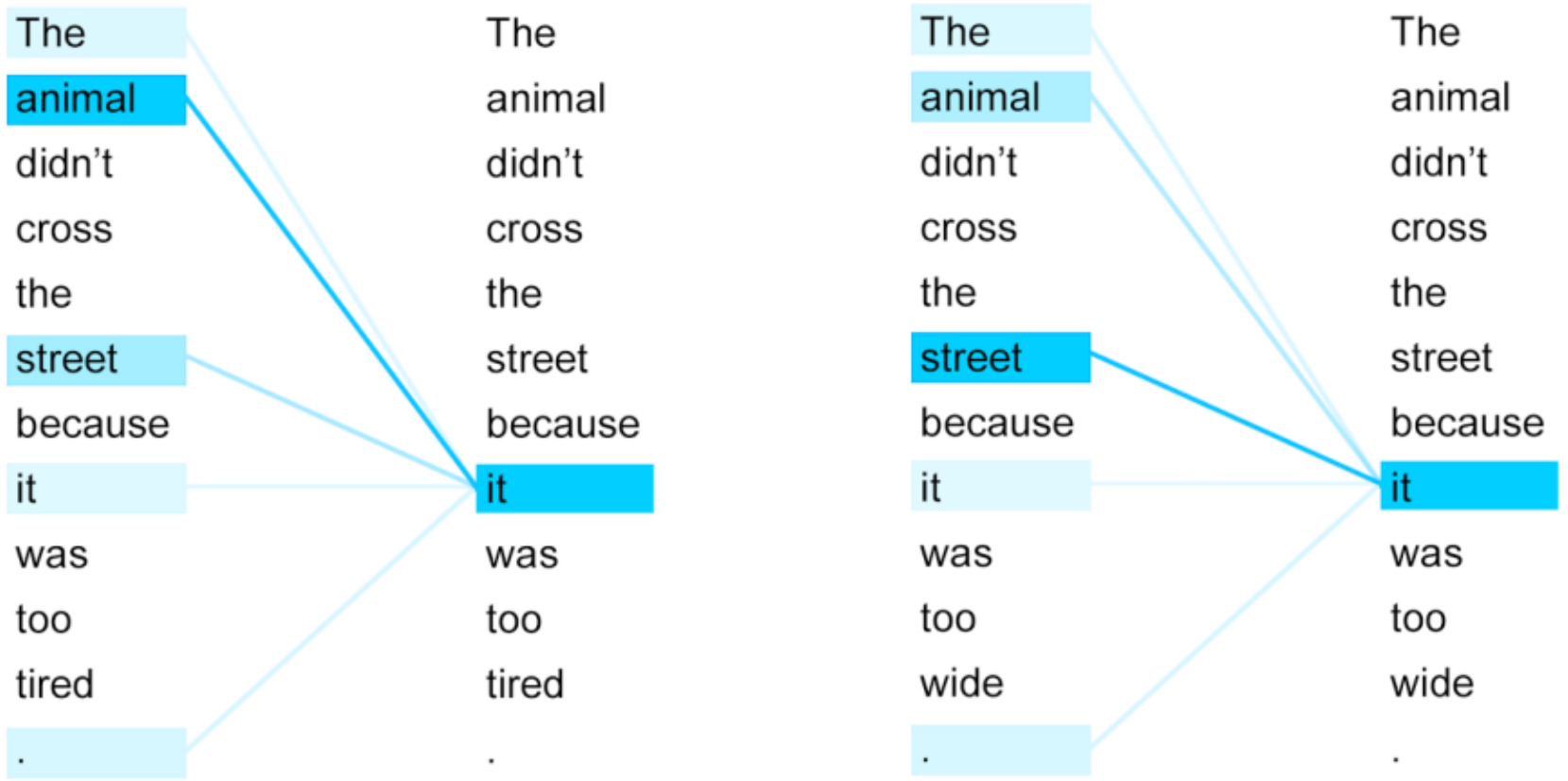
sequence-to-sequence model (seq2seq) หรือเรียกอีกชื่อว่า RNN Encoder-Decoder ซึ่งโมเดลนี้จะแบ่งเป็นสองส่วน เรียกว่า encoder กับ decoder โดยส่วน encoder จะรับ input เข้ามาทีละหน่วยผ่านทาง RNN และเก็บสะสม information ที่จำเป็นไว้ จากนั้นจะผ่าน information นี้ไปยังส่วน decoder ซึ่งก็จะเป็น RNN อีกตัวนึงที่ให้ output ออกมาทีละหน่วย โดยดูจาก information ที่ได้รับมา และ output ตัวก่อนหน้า



โมเดล seq2seq จะมีปัญหาคอขวดเกิดขึ้น นั่นคือการส่ง information เป็นทอดๆ ตามสายยาวแบบนี้ อาจจะมี information ที่จำเป็นบางอย่างสูญหายไประหว่างทางได้ จะดีกว่าไหมถ้าเราให้กระบวนการสร้าง output สามารถโฟกัสไปที่ input ส่วนใดส่วนหนึ่งได้โดยตรง และนี่คือที่มาของ Attention นั่นเอง

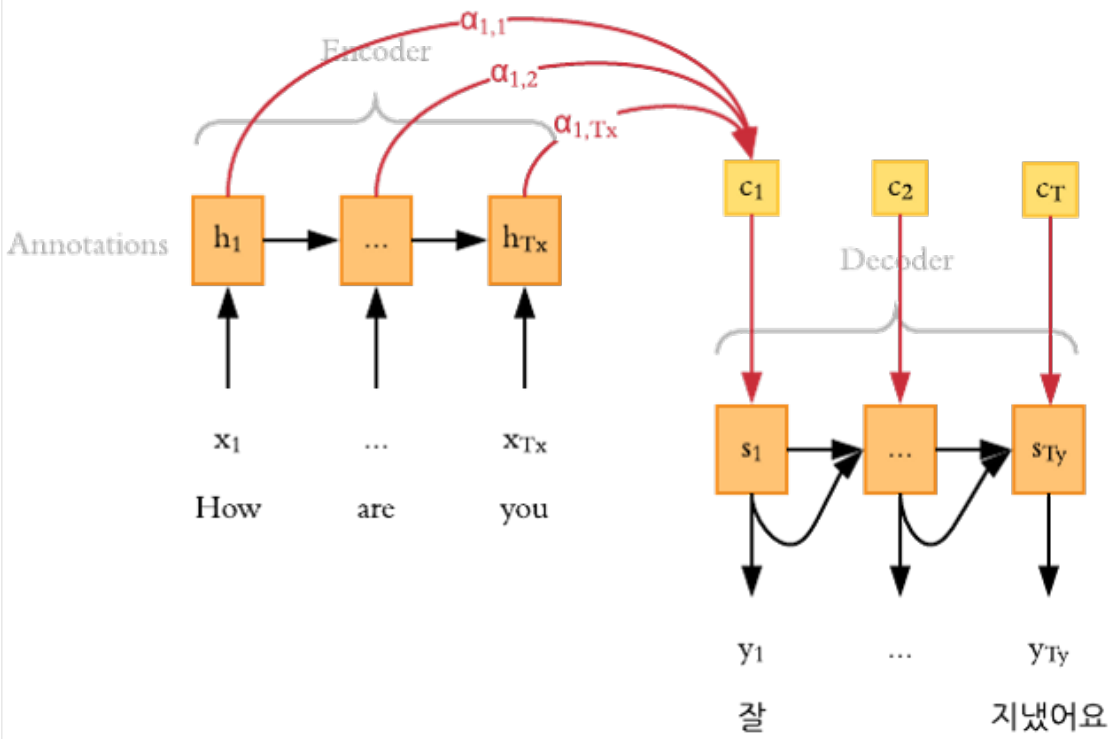
Transformer model : Attention Is All You Need

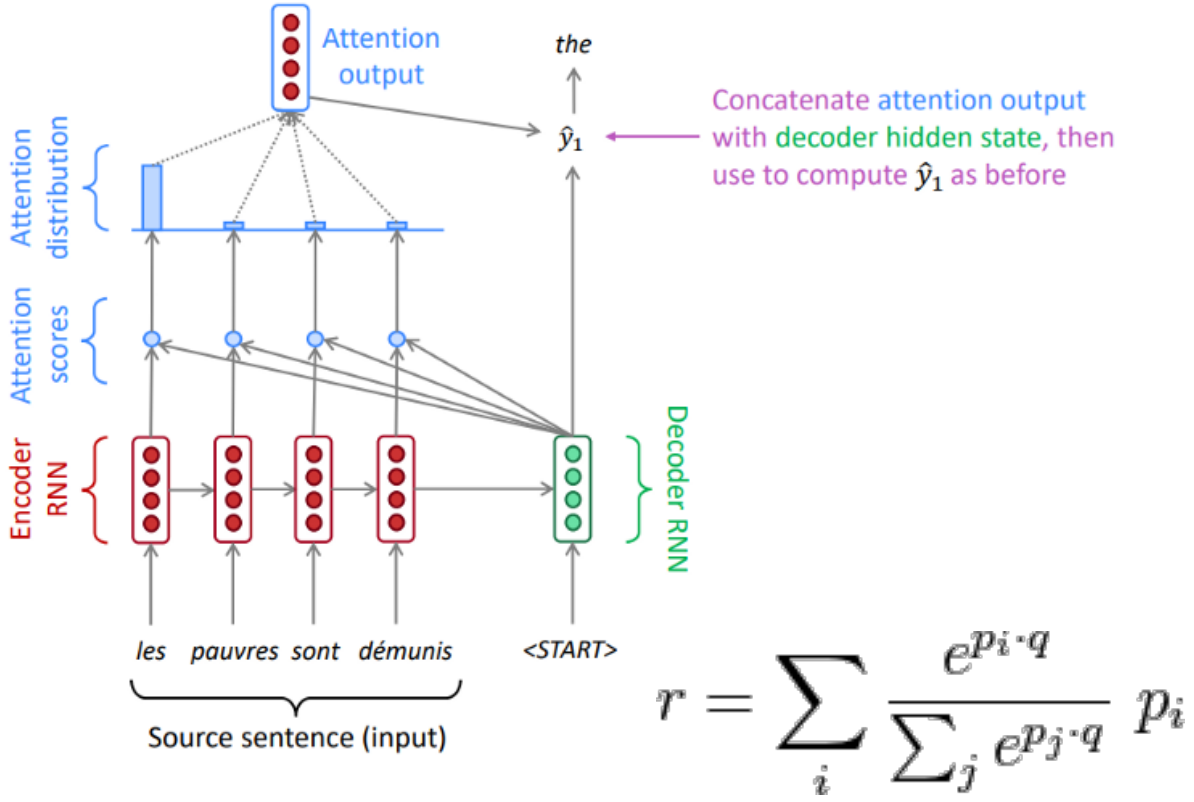
The **idea here is to learn a context vector** (say *U*), which gives us global level information on all the inputs and tells us about the most important information (this could be done by taking a **cosine similarity** of this context vector U the input hidden states from the fully connected layer. We do this for each input *x_i* and thus obtain a *theta_i* (attention weights).



Attention

Attention สำหรับ seq2seq สามารถแสดงได้ดังรูปด้านล่างนี้ โดยเมื่อต้องการจะคำนวณ output ที่ตำแหน่งนึง ก็จะนำ vector ของ decoder (*q*) ณ ตำแหน่งนั้น มาใช้หา attention score กับ vector ของ encoder (*p*) ในทุกตำแหน่ง ซึ่งถ้า score ที่ encoder ตำแหน่งไหนสูง หมายความว่าเราจะให้ความสำคัญ หรือใส่ใจกับตำแหน่งนั้นมาก การคำนวณค่านี้ก็ทำได้หลายวิธี โดยวิธีที่ง่ายที่สุดก็คือการทำ dot product กันตรงๆระหว่าง *p* กับ *q* เลย ซึ่งหมายความว่าเราจะใส่ใจกับตำแหน่งที่มีค่า *p* ใกล้เคียงกับค่า *q* และเมื่อได้ค่า score ออกมาแล้ว ก็จะเอาเข้าฟังก์ชัน *softmax* เพื่อแปลงเป็นค่าความน่าจะเป็น



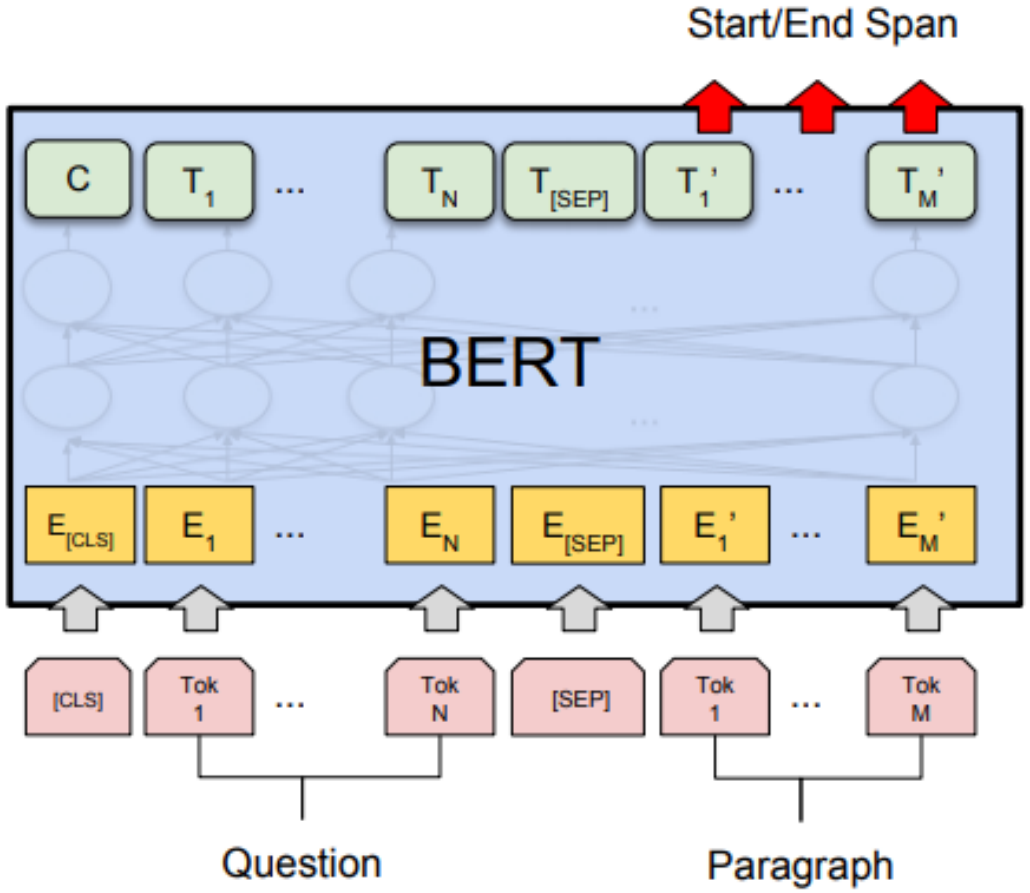


$$r = \sum_i \frac{e^{p_i \cdot q}}{\sum_j e^{p_j \cdot q}} p_i$$

BERT

(Bidirectional Encoder Representations from Transformers)

BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The paper's results show that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models.



(c) Question Answering Tasks:

PDF File

การศึกษาเปรียบเทียบภาษาไทยและภาษาอังกฤษในประเทศไทย
A Contrastive Study of Thai and English in Thailand

พิสุทธิ์พงศ์ เอ็นดู[1]
Pisutpong Endoo[1]

บทคัดย่อ
การศึกษาการเปรียบเทียบภาษาอังกฤษและภาษาไทยในประเทศไทย เริ่มต้นตั้งแต่ปี พ.ศ 2503 เป็นต้นมาหลังจากมีนักศึกษาไทยเดินทางไปศึกษาต่อในระดับปริญญาเอกและสำเร็จการศึกษากลับมา นักศึกษาไทยผู้นวกรุ่นนั้นได้แก่ อาจารย์พุ่งเพื่อง เครือตราชู อาจารย์เฉลา ไชยวัฒนะและอาจารย์พรรณิมี รุทรวณิชย์ นักวิชาการเหล่านี้ได้ศึกษาค้นคว้าและทำวิจัยเกี่ยวกับการศึกษาภาษาเชิงเปรียบเทียบภาษาไทยกับภาษาอังกฤษโดยมีการศึกษาเปรียบเทียบเรื่องคำ วลี การออกเสียงและประโยค เป็นต้น จากการริเริ่มการศึกษาจากนักวิชาการเหล่านี้เองต่อมามีนักภาษาศาสตร์ในประเทศไทยเป็นจำนวนมากให้ความสนใจเกี่ยวกับการศึกษาภาษาศาสตร์โดยเฉพาะอย่างยิ่งการศึกษาภาษาเชิงเปรียบเทียบและในปัจจุบันนี้รายวิชาดังกล่าวได้เปิดทำการเรียนการสอนในมหาวิทยาลัยต่าง ๆ ทั่วประเทศไทยโดยมีวัตถุประสงค์เพื่อศึกษาวิเคราะห์เปรียบเทียบและอธิบายความเหมือนและความแตกต่างของภาษาทั้งสองในประเด็นหัวข้อทางไวยากรณ์ที่สำคัญเช่น เรื่องคำ วลี ประโยค ความหมายในลักษณะรูปแบบต่อรูปแบบ ยกตัวอย่างเช่น การศึกษาเปรียบเทียบความแตกต่างของปัจจุบันกาลในภาษาอังกฤษและภาษาไทยทั้งในด้านความหมายและรูปภาษาศึกษาหรือการศึกษาวิเคราะห์การออกเสียงพยัญชนะควบกล้ำสองเสียงท้ายคำภาษาอังกฤษของนักเรียนไทยเป็นต้น เครื่องมือที่ใช้ในการศึกษาวิเคราะห์และรวบรวมข้อมูลประกอบไปด้วยเหล่าโครงสร้างของทั้งสองภาษา ทฤษฎีภาษาศาสตร์ แบบสอบถาม แบบฝึกภาษณ์ เครื่องอัดเสียง เป็นต้น สถิติที่ใช้สำหรับการศึกษาวิเคราะห์นั้นได้แก่ความถี่ ค่าร้อยละ ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน ทั้งนี้ขึ้นอยู่กับประเด็นและงานที่จะศึกษา ผลของการศึกษาการเปรียบเทียบภาษาศาสตร์และภาษาอังกฤษนั้นสามารถอธิบายถึงความแตกต่างและความเหมือนของทั้งสองภาษาทำให้ผู้เรียนภาษาที่สองเรียนรู้ง่ายมากยิ่งขึ้น

คำสำคัญ: การศึกษาเปรียบเทียบ, ภาษา, ภาษาศาสตร์

Abstract

A contrastive study of Thai and English in Thailand started in 1960 after Thai students graduated with Doctorate degrees from universities abroad. The Thai students were Phungphueng Kruetrachoo, Chalao Chairatana and Pannanee Rutanawanich. All three of them studied and con-

[1] อาจารย์ประจำสาขาวิชาภาษาศาสตร์ คณะเทคโนโลยีการจัดการมหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี วิทยาเขตสุรินทร์
[1] A lecturer in Linguistics Department, Faculty of Management Technology, Rajamangala University of Technology Isan, Surin Campus

OCR

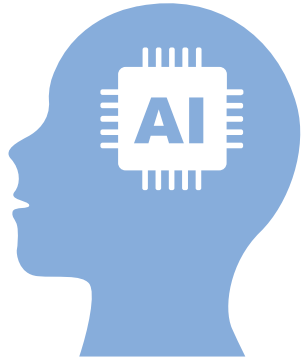
TEXT

\ufeff_____ .. . II . สัญญาเลขที่ TMV/CM212015-1575/BBB/ys Project : FTTX Fund Code : ..... Fund Center : สัญญาว่าจ้างสร้าง และ/หรือ ปรับปรุงข่ายสายสัญญาณโทรคมนาคม สัญญาฉบับนี้ทำขึ้นเมื่อวันที่ 0 1 .ค. 2558 ณบริษัท ทรู มูฟ จำกัด ระหว่าง (1) บริษัท ทรู มูฟ จำกัด โดย นายศุภชัย เจียรวนนท์ และ ศาสตราจารย์พิเศษ อสีก อัศวานันท์ กรรมการผู้มีอำนาจ สำนักงานตั้งอยู่เลขที่ 18 อาคารทรูทาวเวอร์ ถนนรัชดาภิเษก แขวง ห้วยขวาง เขตห้วยขวาง กรุงเทพมหานคร ซึ่งต่อไปนี้ จะเรียกว่า "ผู้ว่าจ้าง" ฝ่ายหนึ่ง กับ อพ (2) ห้างหุ้นส่วนจำกัด บรอดแบนด์ โดยนายปัญญาวุฒิ พลราช หุ้นส่วนผู้จัดการสำนักงานตั้งอยู่เลขที่ 154 หมู่ ที่ 12 ตำบลกุดลาด อำเภอเมืองอุบลราชธานี จังหวัด อุบลราชธานี ซึ่งต่อไปนี้จะเรียกว่า "ผู้รับจ้าง"อีกฝ่ายหนึ่ง โซกา นง CA2-007 หน้า 19 Standard Form 2015

BERT base model

Train

Trained BERT model



Predict

Result

ห้างหุ้นส่วนจำกัด หรือ บริษัทใดเป็นผู้รับจ้าง ?

ห้างหุ้นส่วนจำกัด บรอดแบนด์

Q&A