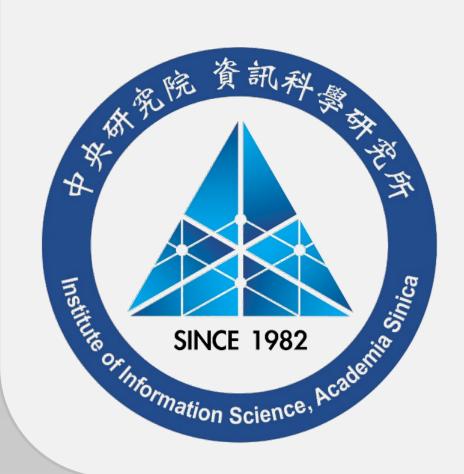
CKIP at the NTCIR-13 STC-2 Task



Wei-Yun Ma ma@iis.sinica.edu.tw Academia Sinica, Taiwan

Chien-Hui Tseng r05725004@ntu.edu.tw National Taiwan University Yu-Sheng Li b03902086@ntu.edu.tw National Taiwan University



→Introduction ←

Motivation

- In recent years, encoder-decoder mechanism like Sequence-to-Sequence Model has been applied successfully in many fields, including short text conversation and machine translation. The inputs and outputs of the models are usually word sequences, named as WordSeq-to-WordSeq Model
- However, for a fixed-size training corpus, data sparseness problem could be an obstacle.

Main Idea

- To address the problem, through this task, we propose the idea of ConceptSeq-to-WordSeq Model
- That is, given input word sequence, we first predict the concept for each word of the word sequence and thus form a concept sequence as the input of the LSTM model. The output remains the form of word sequence.

→ Model ←

ConceptSeq-to-WordSeq Model

Step1: Concept Prediction

- * To predict the concept for each word of the given word sequence, we first need to predict the sense for each word in ENowNet
- * The challenge is there is no annotated corpus using sense definition of EHowNet available. To address this issue, we utilize the comprehensive part of speech (POS) defined in EHowNet and a Chinese corpus with annotations of simplified POS to achieve the effect of WSD.
- * The approach is based on our two observations:
- 1. For almost all Chinese words, once a word's simplified POS is identified, its comprehensive POS can be referred.
- 2. For most cases in Ehownet, a pair of word and its comprehensive POS represents a unique sense.

Input: 吃 牛肉麵 還是 炒飯?

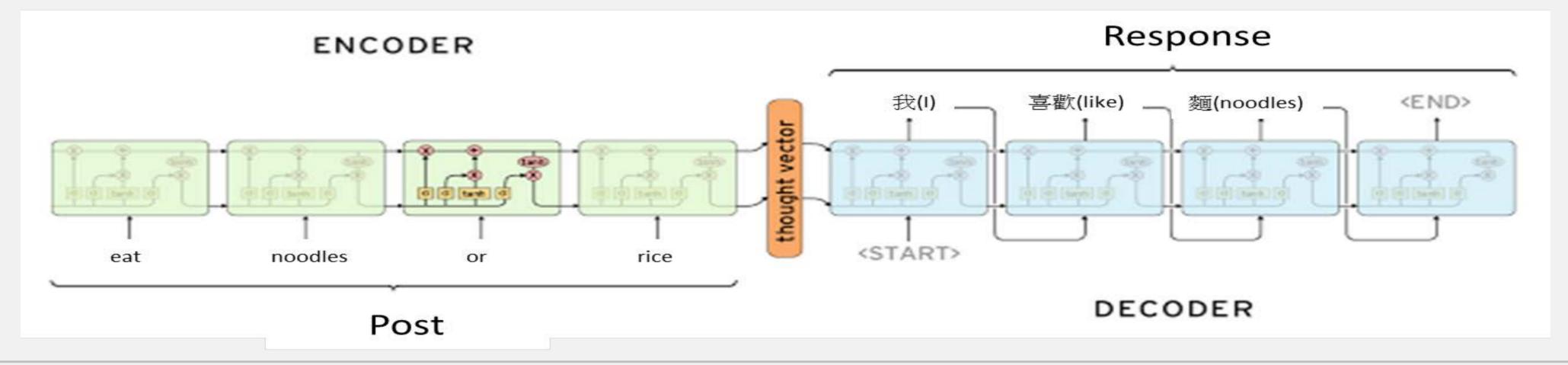
After Sense Prediction: 吃_VC31 牛肉麵_Naa 還是_Caa 炒飯_ Nab? After Concept Prediction: eat noodles or rice?

Step2: ConceptSeq-to-WordSeq Model

- * An LSTM-based encoder-decoder model
- * Input is concept sequence while output is word sequence.

□ <mark>inanimate 無生物</mark> [物資 ,無生物 ,實物]					
由· <mark>NaturalThing 天然物</mark> [大自然 ,天物 ,天然物 ,自然 ,自然物 ,造化]					
😑 artifact 人工物 [成品,物件,物品,消費品,貨,製成品,製品,輕工業品]					
由 clothing 衣物 [衣帛,衣服,衣物,衣冠,衣衫,衣帽,衣装,衣裳,衣履,衣褲,服裝,服飾,衫褲,被服,棉紡織品,袷]					
□ edible 食物 [水米,伙食,吃食,肴饌,客飯,食,食物,料理,茶食,茶飯,酒食,酒飯,飯菜,飲食,盤中飧,燕食,膳食,					
白 food 食品 [田糧 ,食品 ,食糧 ,膳 ,糧 ,饔]					
□ 主食 staple [主食]					
由 <mark>麵包 bread</mark> [花捲 ,黑麵包 ,粸 ,麵包]					
白 飯 CookedRice [八寶飯,米食,米飯,乾餅					
	由 <mark>粥 porridge</mark> [八寶粥,水飯,米粥,佛粥,泡飯,食糜,稀粥,稀飯,粥,臘八粥,糜]				
- 【燴飯】[燴飯 , 膾飯 , 羹飯] - 「白飯」[白雀飯 , ┣飯 , 羹飯]					
	·····································				
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・					
□ OtherWord(飯 CookedRice)					
焦飯					
湯飯					
- 咖里飯					
- 米粉肉					
飯糰 飯糰					
	··· 大鍋飯				
······ 油飯					
	粉條,掛麵,涼麵,淺	弱麵 ,陽春麵 ,餑 ,擔仔麵 ,擔擔麵 ,餛飩麵 ,麵 ,麵食 ,			
[
	詞彙:	牛肉麵			
⊟ OtherWord(麺 noodles)	詞性:	Naa			
蚵仔麵線 蚵仔麵線	英文意涵:	beef noodles			
豬腳麵線 炒麵	Event Frame:				
	定義式:	{麵 noodles:ingredients={牛肉 beef}}			
- 乾粉	操作式:				
一 	語義功能:				
 - 壽麺	語義特徵:	+meal			
白麵	#0 3X19 0X*	{edible 食物:material={crop 莊稼},ingredients={flesh			
由· 餃子 dumplings [水餃 , 抄手 , 扁食 , 湯餃	展開式:	内:predication={eat 吃:patient={~}},whole={livestock 牲畜}}}			
□ □ <mark>菜餚 dish</mark> [粑 , 糌粑] □ 菜餚 dish [肴 , 砂鍋菜 , 菜肴 , 菜餚 , 滷菜 , :	WordNet 自動連結:	{noodle.n.01, beef.n.02, beef.n.01, attic.n.03, gripe.n.01}			

ate|生物 [生物 ,生物體 ,有機體 ,物種 ,活體 ,動物體 ,動植物]



→Experiment ←

Experimental Settings

	LSTM Seq-to- Seq Type	Pretrain word embedding	Attention model type	N-gram on decoding
Run-G1	WS-to-WS	CBOW	general	bigram
Run-G2	WS-to-WS	no	general	bigram
Run-G3	WS-to-WS	CBOW	concat	trigram
Run-G4	CS-to-WS	CBOW	concat	bigram

- Pretrained word embedding by using CBOW of word2vec on ASBC Chinese corpus with size of 10 million words.
- Embedding dim: 300
- During training, we filter out the pairs which are labelled as high quality by all three annotators, leaving only 6276 pairs are used for training.

Results

	Mean MSnDCG@0001	Mean P-nine	Mean nERR@0010
Run-G1	0.0017	0.0029	0.0015
Run-G2	0.005	0.0086	0.0046
Run-G3	0.01	0.0171	0.0093
Run-G4	0.0083	0.0143	0.0077

Post	好喜欢小葡萄的画啊[太开心] 喜欢的赶紧来围观哦[围观]
Run-G1	#、继续个冬至了
Run-G2	没有上海、幸福。精彩
Run-G3	你也要吃了吗?他不要过去的吗?
Run-G4	长的人心疼的图片好漂亮的好漂亮爆了。