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.

nimate 生物[生物 ,生物體 ,有機體 ,物種 ,活體 ,動物體 ,動植物] animate 無少物 [物溶 無少物 寅物]
animate 無生物[物資 ,無生物 ,實物]
· <mark>NaturalThing 夭然物</mark> [大自然 ,夭物 ,夭然物 ,自然 ,自然物 ,造化] ————————————————————————————————————
· <mark>artifact 人工物</mark> [成品 ,物件 ,物品 ,消費品 ,貨 ,製成品 ,製品 ,輕工業品]
□ 由 <mark>clothing 衣物[</mark> 衣帛 ,衣服 ,衣物 ,衣冠 ,衣衫 ,衣帽 ,衣裝 ,衣裳 ,衣履 ,衣褲 ,服裝 ,服飾 ,衫褲 ,被服 ,棉紡織品 ,袷]
□□□ <mark>□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□</mark>
□ <mark>food 食品</mark> [田糧,食品,食糧,膳,糧,饔]
□· <mark>主食 staple</mark> [主食]
直··· <mark>麵包 bread</mark> [花捲 ,黑麵包 ,粸 ,麵包]
□ <mark>飯 CookedRice</mark> [八寶飯 ,米食 ,米飯 ,乾飯 ,粥飯 ,飯食 ,熟飯 ,飯]
────────────────────────────────────
— <mark>{ 燴飯 }</mark> [燴飯 , 膾飯 , 羹飯]
·····································
⊡ OtherWord(飯 CookedRice)
大鍋飯 剰飯 鍋巴

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>麵 noodles</mark> [切麵,拉麵,油麵,炸醬麵,>	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	
	洞棄訊息 ×		
— <mark>{米粉}</mark> [米粉 ,粉絲 ,線粉] — <mark>{杯麵}</mark> [杯麵 ,碗麵]	詞彙:	牛肉麵	
⊡ OtherWord(麵 noodles)	詞性:	Naa	
蚵仔麵線	英文意涵:	beef noodles	
豬腳麵線 炒麵	Event Frame:		
~~~~~ 素麵	定義式:	{麵 noodles:ingredients={牛肉 beef}}	
	操作式:		
	語義功能:		
	語義特徵:	+meal	
── 烏龍麵 ⊡─ <mark>餃子 dumplings</mark> [水餃 ,抄手 ,扁食 ,湯餃	展開式:	{edible 食物:material={crop 莊稼},ingredients={flesh  肉:predication={eat 吃:patient={~}},whole={livestock  牲畜}}}	
── <mark>{粑}</mark> [粑,糌粑] ⊡─ <mark>菜餚 dish</mark> [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}	



→ Experiment ←

### **Experimental Settings**

	LSTM Seq-to- Seq 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.

	Mean MSnDCG@0001	Mean P-nius	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

**Results** 

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	<b>#、</b> 继续 <b>个冬至了</b>
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。