# 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.**

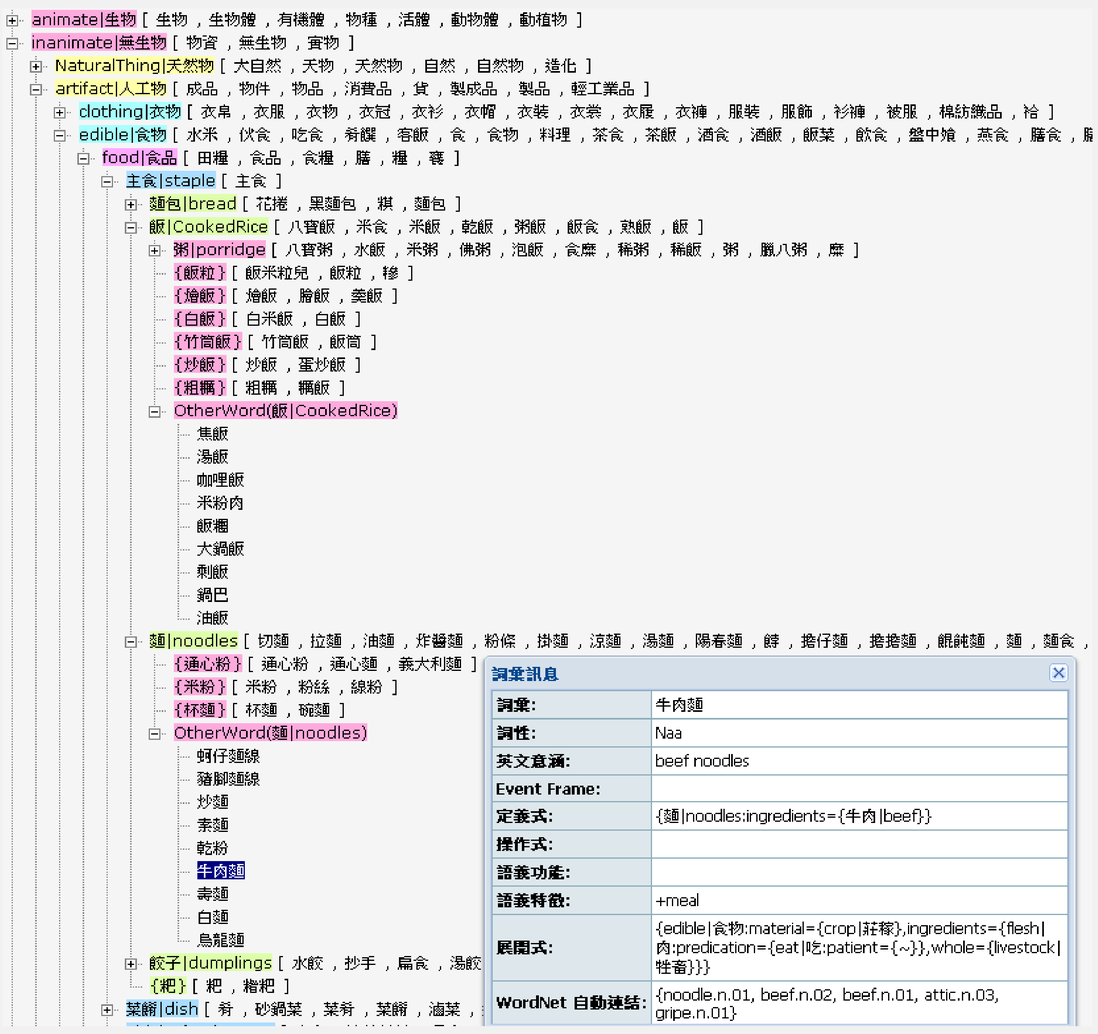
## ➔Model←

## ConceptSeq-to-WordSeq Model

### Step1: Concept Prediction

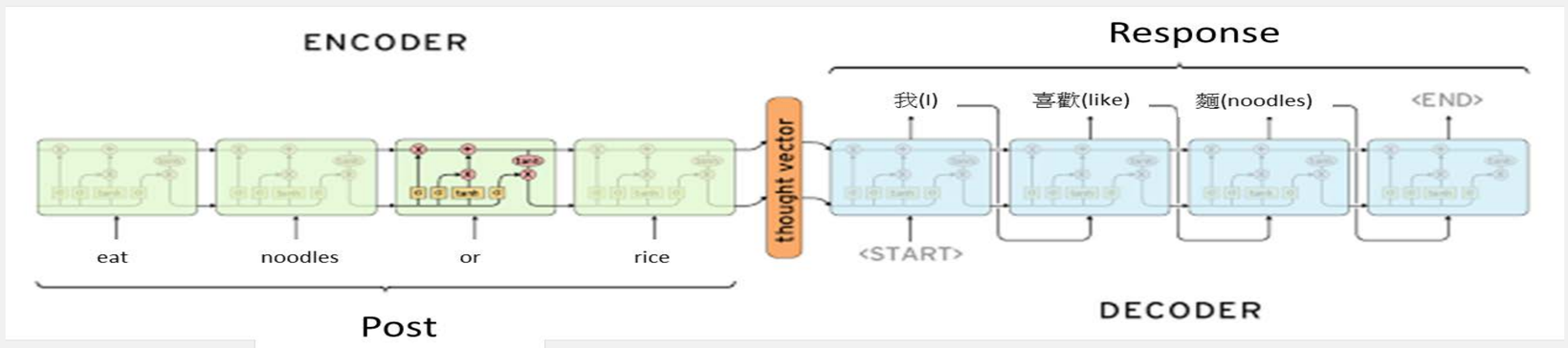
* To predict the concept for each word of the given word sequence, we first need to predict the sense for each word in ENowNet
* The challenge is there is no annotated corpus using sense definition of EHowNet available. To address this issue, we utilize the comprehensive part of speech (POS) defined 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.




## ➔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-plus | 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。 |