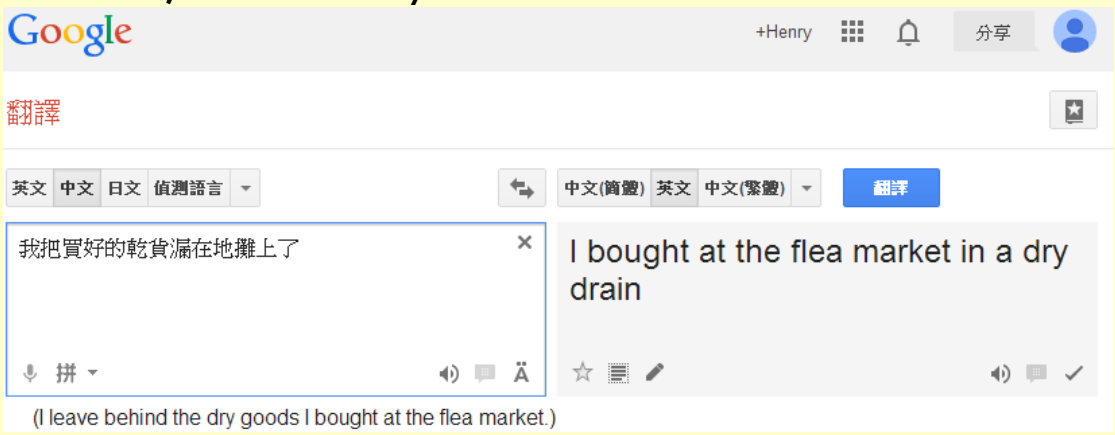
Making Machines More Human:
AI Models for Learning to Understand and Translate Human Language

Lie Ming Hing
Dekai Wu

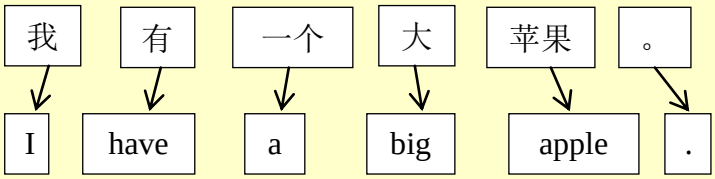
Introduction

Accurate translation between Chinese and English involves making machines understand the true meaning of a sentence, which is very hard for a machine to do.



Even Google cannot accurately translate every sentence.

One reason is the structure of Chinese sentences is different from European languages. A Chinese sentence is merely composed of Chinese characters. They do not have spaces (word boundaries) between one word and another. The problem can be reduced if the Chinese sentences are segmented. For example, a Chinese sentence 我有一个大苹果 'I have a big apple' is segmented into



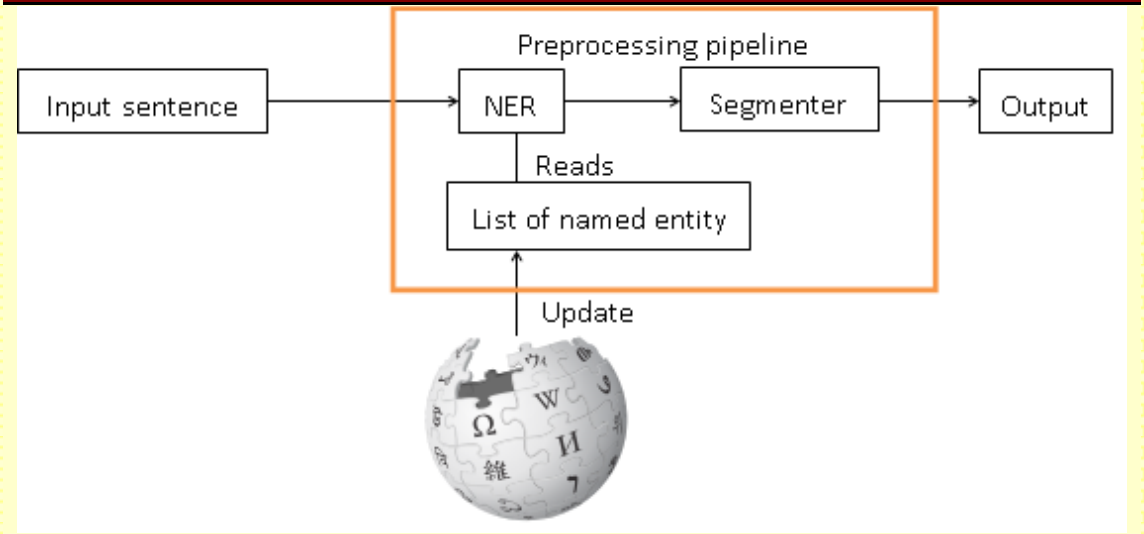
Then the translation is easy.

Objective

A Chinese pre-processing pipeline that includes

- A Chinese segmenter
- A named entity recognizer
  (Named entities means proper nouns)

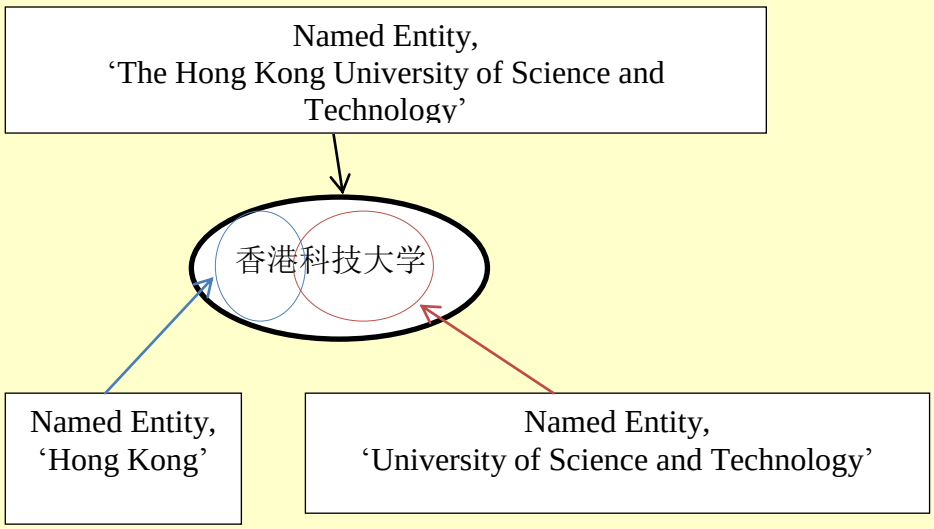
Design



The design is simple. The sentences input will be passed through a Named Entities Recognizer (NER) that the library uses the data from Wikipedia. Then the sentences will be segmented by a chosen segmenter and be output.

NER Algorithm

The main part of the NER is a word-by-word search. When the entity is ambiguous to be recognized, the longest one is pick by the NER.



(In this case only 'The Hong Kong University of Science and Technology' will be recognized.)

NER Performance

The NER is implemented and working as expected, a input-output example is as follow:

INPUT	OUTPUT
香港科技大学是香港一所知名的大学	<namex translation = "The Hong Kong University of Science and Technology">香港科技大学</namex>是<namex translation = "Hong Kong">香港</namex>一所知名的大学

Result

The processing pipeline increased the performance of a baseline translation across all evaluation metrics.

MT system	MEANT	BLEU	NIST	METEOR	TER	WER	PER	CDER
Baseline	32.50	14.56	5.01	43.11	69.69	73.19	57.04	68.89
Baseline with New Chinese preprocessing	32.93	15.19	5.27	45.51	68.70	72.57	55.42	67.55

(MEANT, BLEU, NIST, METEOR are scoring metrics where higher scores in these metrics indicate better translation output. TER, WER, PER, CDER are error rate metrics where lower error rates in these metrics indicate better translation output.)