## 自然語言實作 第12週:利用來回翻譯產生重述片語

國立清華大學 資工系計算語言學教授

張俊盛

### 同義詞與重述

- 同義詞=同詞性,以單字為主
- 自動重述=不限定詞性、結構,以片語為主
- 作法 1: 同一本書的不同翻譯(互相對齊)
- 作法 2 : 利用片語式統計機器翻譯
  - 例如,將輸入英文翻譯成漢語,再翻譯回英文
  - 取英漢翻譯機率 x 漢英翻譯機率最大值的最佳結果
  - 或用英漢翻譯機率 x 漢英翻譯機率,來排序一組可能的結果
- 相關研究 重述資料庫 PPDB

#### Sources:

- 1. <a href="https://en.wikipedia.org/wiki/Paraphrasing\_(computational\_linguistics)">https://en.wikipedia.org/wiki/Paraphrasing\_(computational\_linguistics)</a>)
- 2. <a href="http://paraphrase.org/#/">http://paraphrase.org/#/</a> (<a href="http://paraphrase.org/#/">http://paraphrase.org/#/</a>)
- 3. <a href="http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz">http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz</a> (<a href="ht

#### In [15]:

```
from IPython.display import HTML
url = "https://en.wikipedia.org/wiki/Paraphrasing_(computational_linguistics)"
iframe = "<iframe src=%s width=800 height=350></iframe>"%url
HTML(iframe)
```

Out[15]:

### WikipediA

# Paraphrasing (computational linguistics)

**Paraphrase** or **Paraphrasing** in computational linguistics is the natural language processing tast and generating paraphrases. Applications of paraphrasing are varied including information retrievanswering, text summarization, and plagiarism detection. Paraphrasing is also useful in the machine translation, as well as semantic parsing and generation of new samples to expression.

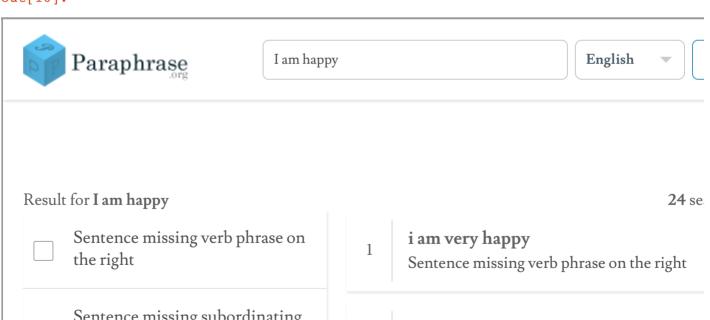
### **Contents**

Paraphrase generation

### In [16]:

```
from IPython.display import HTML
url = "http://paraphrase.org/#/search?q=I%20am%20happy&filter=&lang=en"
iframe = "<iframe src=%s width=800 height=350></iframe>"%url
HTML(iframe)
```

#### Out[16]:



- 利用 consistent blocks 的概念,由詞對詞翻譯,導出片語到片語翻譯
  - 第五章講義 9-15 頁
  - 利用 state space search 的概念,搜尋最佳翻譯
    - 由○翻譯(起始狀態,到整句翻譯(結束狀態)
    - 下一步:考慮翻譯新片語(連續或跳躍)
    - 下一步:計算所有可能的翻譯片語
    - 下一步:計算翻譯機率+語言模型機率
    - 第六章講義 7-15 頁
  - 每次作兩次翻譯英語到漢語,漢語到英語

#### Sources:

- 1. <a href="http://www.statmt.org/book/slides/05-phrase-based-models.pdf">http://www.statmt.org/book/slides/05-phrase-based-models.pdf</a> (<a href="http://www.statmt.org/book/slides/05-phrase-based-models.pdf">http://www.statmt.org/book/slides/05-phrase-based-mode
- 2. <a href="http://www.statmt.org/book/slides/06-decoding.pdf">http://www.statmt.org/book/slides/06-decoding.pdf</a> (http://www.statmt.org/book/slides/06-decoding.pdf)
- 3. <a href="http://www.statmt.org/book/slides/07-language-models.pdf">http://www.statmt.org/book/slides/07-language-models.pdf</a> (<a href="http://www.statmt.org/book/slides/07-language-models.pdf">http://www.statmt.org/book/slides/07-lang

### 本週任務

Level A: 用詞到詞模式,自動產生片語的重述語 (80分)

● 不考慮結構變化,不涉及跳躍式翻譯,使用 docode.py

Level B: 用片語到片語模式,自動產生片語的重述語 (100分)

● 不考慮結構變化,不涉及跳躍式翻譯,使用consistent\_block.py 產生片語翻譯表

Level C: 用片語到片語模式,自動產生片語的重述語 (期末專題)

● 考慮結構變化,涉及跳躍取原文片語,作連續性翻譯書處 修改 decode.py