

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 \times 漢英翻譯機率最大值的最佳結果
 - 或用英漢翻譯機率 \times 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. [\(https://en.wikipedia.org/wiki/Paraphrasing_\(computational_linguistics\)\)](https://en.wikipedia.org/wiki/Paraphrasing_(computational_linguistics))
([\(https://en.wikipedia.org/wiki/Paraphrasing_\(computational_linguistics\)\)](https://en.wikipedia.org/wiki/Paraphrasing_(computational_linguistics)))
2. <http://paraphrase.org/#/> (<http://paraphrase.org/#/>)
3. [\(http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz\)](http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz) ([\(http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz\)](http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz))

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 task of generating paraphrases. Applications of paraphrasing are varied including information retrieval, answering, text summarization, and plagiarism detection.^[1] Paraphrasing is also useful in the machine translation,^[2] as well as semantic parsing^[3] and generation of new samples to expand corpora.^[4]

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]:



Paraphrase.org

I am happy

English

Result for I am happy

24 se



Sentence missing verb phrase on the right

1

i am very happy

Sentence missing verb phrase on the right

Sentence missing subordinating

利用片語式統計機器翻譯作自動重述

● 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. <http://www.statmt.org/book/slides/05-phrase-based-models.pdf> (<http://www.statmt.org/book/slides/05-phrase-based-models.pdf>)
2. <http://www.statmt.org/book/slides/06-decoding.pdf> (<http://www.statmt.org/book/slides/06-decoding.pdf>)
3. <http://www.statmt.org/book/slides/07-language-models.pdf> (<http://www.statmt.org/book/slides/07-language-models.pdf>)

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`

