

利用單語翻譯 進行英文文法改錯

教科書網站：`www.statmt.org/book/`

參考課程網站：`mt-class.org/jhu/syllabus.html`

10/23, 2018

文獻回顧

- Brockett, Chris, William B. Dolan, and Michael Gamon. "Correcting ESL errors using phrasal SMT techniques." *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 2006.
- Grundkiewicz, Marcin Junczys-Dowmunt Roman. "The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by Data-Intensive and Feature-Rich Statistical Machine Translation." *CoNLL-2014* (2014): 25.
- Felice, Mariano, et al. "Grammatical error correction using hybrid systems and type filtering." *CoNLL-2014* (2014): 15.

語料庫：錯與對的對應英文句 **Parallel Corpora**

- NUCLE (National Singapore University Learner Corpus, 1,450 essays)
- FCE (Cambridge Learner Corpus, First Certificate English Test, 1,124 samples)
- CLEC (Chinese Learner English Corpus, 1 百萬詞)
- WikEd Error Corpus (Wikipedia revision log, 1千2百萬句)
 - `romang.home.amu.edu.pl/wiked/wiked.html`
- EFCAMDAT Learner Corpus (2百萬句)
 - 17萬學生1百萬篇作文8千3百萬詞，程度A1-C2.
 - `corpus.mml.cam.ac.uk/efcamdat2/public_html/`

EFCAMDAT 舉例

Hi, Mr.Blight,

There are four [-candidated//D-] properties as {+per//MW+} your requirement. The first one I would like to recommend is a property that you will see anywhere in the world.

It has [{-intergrated+integrated//SP}] functional rooms, bedroom, living room, kitchen and bathroom, which needs a new roof renovation. The [{-size land+land size//W0}] is 288.45 sq m and cottage is 54.15 sq with {+an//AR+} extension {+of//PR+} up to 150 sqm [{-permittable+permissible//SP}] with {+a//AR+} price {+of//PR+} \$200,000.

If you are interested in another property with history, I can recommend one owned by Lady Elizabeth Hamilton. It is much larger than {+the//AR+} first one, {+the//AR+} land

[{-of+is//x>>y}] 1200 sqm approx and {+the//AR+} house
[{-of+is//x>>y}] 224.76 sq m upstairs and down stairs, including
more rooms. This house cannot be demolished and the price is
\$1.5M.

[{-Another+Other//x>>y}] new apartments, 67 sqm / 78 sqm
can be introduced to you. The price is \$160,000 each, including
the fittings you choose. Now there are only 3x2 bedroom ones
[{-available+available//SP}] .

The last one is a luxurious property for investing.
{+Word Limit//I(x)+} It located in a quiet and traditional corner
of the twon with bay, village and mountain scapes surrounding it.
Only a few minutes walk to the centre and close to surrounding
beaches. The land size is 453.20 sq m and house size is 111.78 sq m.

All related facilities are provided with the price of \$450.000.

Whenever you have the decision, you can call me for the following events.

Best regards, XXX

CLC-FCE 舉例

Dear Madam ,

I 'm writing to [-You-]{+you+} in order to express my feelings about the International Arts Festival .

I spent two days there , and I think it was the best Arts Festival I have ever been [-at-]{+to+} .

I also hope [-,-] that the festival could be even better next year . The great idea of making it " international " was unfortunately not brought to life , because the artists ([-mentioned-]{+said+} to be " from around the world ") were from only six countries .

Maybe the [-eason-]{+reason+} for that is very simple : lack of money ?

It would explain [-,-] why some concert halls were simply too small - you just could n't afford bigger ones .

On the other hand , the wide range of plays and films to choose
{+from+} impressed me {+the+} most .

I thank you for organizing it , and hope to see more next year .

I also thank you for the excellent idea of {+having+} one ticket
for all events , which allowed me to see all the things I was
interested in .

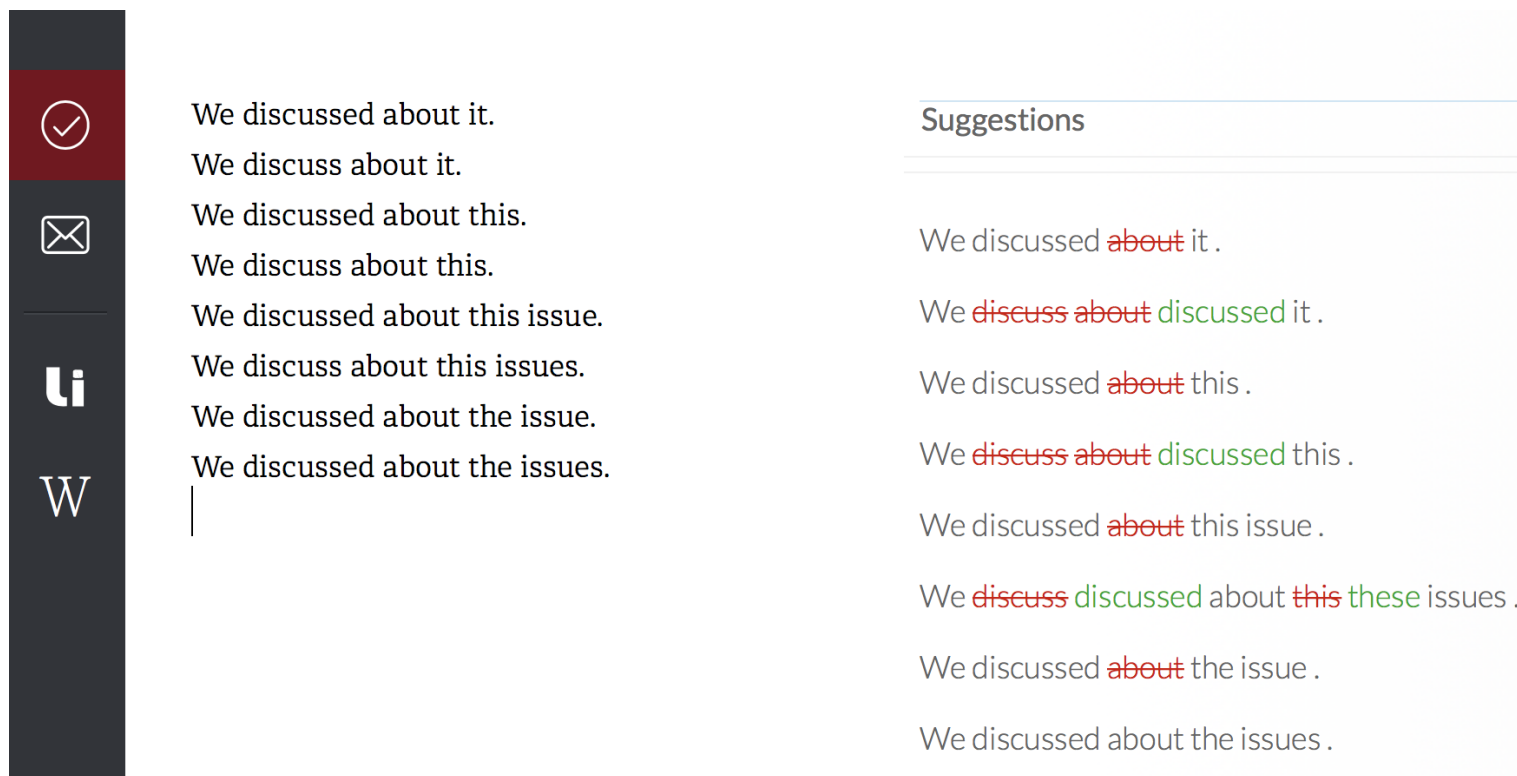
My suggestion for next year 's festival is simple - please make
it a few days longer !

Yours faithfully ,

XXX

本次實作的目的

- `nlp-ultron.cs.nthu.edu.tw/gec/`
- 用 EFCAMDAT 實作的類神經文法改錯系統



	Suggestions
We discussed about it.	We discussed about it .
We discuss about it.	We discuss about discussed it .
We discussed about this.	We discussed about this .
We discuss about this.	We discuss about discussed this .
We discussed about this issue.	We discussed about this issue .
We discuss about this issues.	We discuss discussed about this these issues .
We discussed about the issue.	We discussed about the issue .
We discussed about the issues.	We discussed about the issues .

本次實作的目的

- 用片語為本的統計式機器翻譯，進行文法改錯
 - **discuss about | discuss**
 - **comment | comment on**
- 以後的實作（或期末專題），再轉換成為文法為本的作法
 - **discuss: V about n | V n**
 - **comment: V n | V on n**
- 可以從 *WikEd* 和 *EFCAMDAT*
 - ... cope [-up-] with the disaster ...
 - ... discuss [-about-] the issue ...
 - ... media comment {+on+} some issues ...

起始程式和研究資料

- 程式：
 - `gen.ngram.py`
 - `gen.ngram.edit.py`
 - `gen.ngram.no.edit.py`
 - `decode.py`
 - `models.py`
- 資料：
 - `fce.diff.tokenized.txt`, `fce.diff.about.txt`
 - `ef.diff.simple.simple.txt.zip`, `ef.diff.difficult.txt`
 - `count_1w.txt`, `count_2w.txt` (計算語言模型之用，總數 $N = 10^{12}$)
 - `lidoce.problem.words.nvar.txt`
 - `tm` (參考格式)
 - `bnc.prune.bin`, `bnc.prune.arpa`

展示程式和資料 `gen.ngram.edit.py`

```
$ time cat ef.diff.simple.simple.txt | grep discuss |  
    python gen.ngram.edit.py | sort | uniq -c | grep discuss |  
    sort -k1nr | head -40
```

```
87 discuss [-about-]  
55 discussed [-about-]  
26 discuss [-about-] the  
23 discuss {+the+}  
22 [-for>>to+] discuss  
21 {+to+} discuss  
20 discussed [-about-] the  
14 [-we>>We+] discussed  
14 discuss {+it+}  
13 discussing [-about-]  
12 discuss [-about-] this
```

12 {+the+} discussion
8 [-have-] discussed
8 discuss [-about-] it
8 {+will+} discuss
7 discussed [-about-] our
7 discussed {+the+}
6 [-was>>were+] discussed
6 discuss [-about-] my
6 discuss [-about-] our
6 discuss [-with-]
6 discuss {+it+} with
6 discussed {+about+}
6 discussing [-about-] the
6 {+a+} discussion

展示程式和資料 `gen.ngram.py`

```
$ time cat ef.diff.simple.simple.txt | grep discuss |  
    python gen.ngram.py | sort | uniq -c | grep discuss |  
    sort -k1nr | head -40
```

```
789 to discuss  
222 discuss the  
173 discuss about  
169 we discussed  
130 and discuss  
125 to discuss the  
98 to discuss about  
94 discuss with  
87 discuss [-about-]  
82 meeting to discuss  
80 discussed about
```

76 We discussed
72 discussed the
68 discuss this
65 can discuss
62 discussion about
61 the discussion
55 discussed [-about-]
53 discuss about the
50 a discussion
50 discussion ,
49 to discuss with
46 to discuss [-about-]
45 will discuss
44 discussed in
41 discuss it

41 to discuss about the
40 to discuss this
37 need to discuss
37 we discuss
35 be discussed
35 discussion with
33 a meeting to discuss
33 discuss some
32 discussed with
30 another meeting to discuss
30 discussed about the
29 of discussion
28 I discussed
28 discuss our
28 should discuss

28 we discussed about
27 discussing the
26 discuss [-about-] the
26 was discussed
24 and discussed
24 discussed and
23 discuss and
23 discuss them
23 discuss +the+
23 discussion about the
23 have discussed
23 we discussed the

實作任務

- 寫一個程式 **gen.phrase.table.py** 來產生 phrase-based SMT 的片語翻譯表
 - 格式：<錯誤片語> ||| <正確片語> ||| <機率>
- 改寫 `decode.py` 進行文法改錯 (今天無法完成，下一次實作繼續做)
 - `decode.py` 不太需要修改，因為文法改錯處理局部錯誤，不做大幅度重排
 - 翻譯模型 `tm` 變成英文到英文的轉換
 - 「錯誤片語」轉成「正確片語」

任務 1：產生片語翻譯表

- 將 ngram 轉為 phrase translation table
- 例子
 - discuss [-about-] 87 → discuss about ||| discuss
 - discuss about 173 → discuss about ||| discuss about
- 把相同片語（如 discuss about）不同翻譯（如 discuss about 和 discuss）放在同一組
 - discuss [-about-] 87 → discuss about ||| discuss
 - discuss about 173 → discuss about ||| discuss about

任務 2：計算機率值

- 例子

- discuss [-about-] 87 → discuss about ||| discuss ||| <機率值 P_1 >
- discuss about 173 → discuss about ||| discuss about ||| <機率值 P_2 >

- 用 MLE 計算翻譯機率值

- $P_1 = 87/(87+173) = 0.33$
- $P_2 = 173/(87+173) = 0.67$

- 用 $\log P$ 表示翻譯機率值

- $\log P_1 = \log 0.33 = -0.48$
- $\log P_2 = \log 0.67 = -0.18$

計算機率值的問題

- 有些資料有明顯的文法錯誤，句判斷應該是老師忽略沒有修改
- 如何減低這種標示錯誤（false negative）的影響
- 用語言模型衡量翻譯（修改後）的機率
 - 例如 $P(\text{'discuss about'})$ 和 $P(\text{'discuss'}) - P(\text{'discuss about'})$
 - 或採取 $P(\text{'discuss X'})$ $X \neq \text{'about'}$ 的平均值
(使用 count_1w 和 count_2w)
- 用修改後機率計算一項新的機率值，再正規化（加總為 1）
- 次數算出來的機率和標註信心機率，兩項相乘（對數相加）
- 最後。得到每一則片語翻譯（如 discuss about ||| discuss）的單項機率值

資料集資料可靠度模型

- 另外，一個作法是統計可靠度模型裡面的兩項機率
 - $P_1(True \mid \text{conflicting phrase with edit, for all conflicting phrase, edit})$
 - $P_2(True \mid \text{conflicting phrase with no edit for all conflicting phrase})$
- 取樣本 S_1 （有編輯有衝突）和 S_2 （無編輯有衝突）
- 人工判斷編 S_1 和 S_2 每筆資料的正確性
- 用相對次數，估計 P_1 和 P_2
- 不同語料庫 P_1 和 P_2 不一樣

語言模型的使用範例

- <https://github.com/kpu/kenlm#python-module>
- 安裝
 - `pip install https://github.com/kpu/kenlm/archive/master.zip`
- 基本用法

```
import kenlm
model = kenlm.Model('lm/test.arpa')
print(model.score('this is a sentence .', bos = True, eos = True))
```