自然語言實作 第12週

自然語言實作：利用排名比率擷取名稱、術語、專名

國立清華大學 資工系計算語言學教授

張俊盛

如何在大型語料庫中，發現重要的名稱

- Names 或 Terms 都是辭典條目或術語庫條目
- 指涉世界上的實體、關係、概念的字詞（一字或多字）
- 非組合性詞語，不能拆解（如 hot dog 熱狗)
- 比字面更特定的領域術語（如 CPU)
- 專有名詞，人事時地物（如 John McCarthy)

Sources:

1. https://en.wikipedia.org/wiki/Terminology (https://en.wikipedia.org/wiki/Terminology)
2. https://zh.wikipedia.org/wiki/術語學 (https://zh.wikipedia.org/wiki/%E8%A1%93%E8%AA%9E%E5%AD%B8)

相關研究

- Church and Hank (1989) 提出相互資訊可以發現英文名稱詞
- Sproat and Shih (1990) 利用相互資訊發現中文二字詞
- Deane (2005) 提出 RankRatio 比相互資訊更有效

Sources:

1. Church, Kenneth, and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography." 27th Annual Meeting of the Association for Computational Linguistics. 1989. https://www.aclweb.org/anthology/P89-1010.pdf (https://www.aclweb.org/anthology/P89-1010.pdf)
2. Sproat, Richard, and Chilin Shih. "A statistical method for finding word boundaries in Chinese text." Computer Processing of Chinese and Oriental Languages 4.4 (1990): 336-351. https://rws.xoba.com/newindex/cpcol.pdf (https://rws.xoba.com/newindex/cpcol.pdf)
3. Deane, Paul. "A nonparametric method for extraction of candidate phrasal terms." Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). 2005. https://www.aclweb.org/anthology/P05-1075.pdf (https://www.aclweb.org/anthology/P05-1075.pdf)

Church and Hank (1989)

1. Church, Kenneth, and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography." 27th Annual Meeting of the Association for Computational Linguistics. 1989. https://www.aclweb.org/anthology/P89-1010.pdf (https://www.aclweb.org/anthology/P89-1010.pdf)

In [88]:

```
from IPython.display import IFrame
from IPython.core.display import display
url = "https://aclweb.org/anthology/P89-1010.pdf"
#display(IFrame(url, '199%', '600px'))
IFrame(url, width=800, height=950)
```

Out[88]:

Sproat and Shih (1990)

1. Sproat, Richard, and Chilin Shih. "A statistical method for finding word boundaries in Chinese text." Computer Processing of Chinese and Oriental Languages 4.4 (1990): 336-351. https://rws.xoba.com/newindex/cpcol.pdf (https://rws.xoba.com/newindex/cpcol.pdf)

In [89]:

```
from IPython.display import IFrame
from IPython.core.display import display
url = "https://rws.xoba.com/newindex/cpcol.pdf"
#display(IFrame(url, '199%', '600px'))
IFrame(url, width=800, height=950)
```

Out[89]:

Deane (2005)

1. Deane, Paul. "A nonparametric method for extraction of candidate phrasal terms." Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). 2005. https://www.aclweb.org/anthology/P05-1075.pdf (https://www.aclweb.org/anthology/P05-1075.pdf)

In [90]:

```
from IPython.display import IFrame
from IPython.core.display import display
url = "https://www.aclweb.org/anthology/P05-1075.pdf"
#display(IFrame(url, '199%', '600px'))
IFrame(url, width=800, height=950)
```

Out[90]:

名稱術語的頻率與排名特徵

- 片語頻率相對性提高（相對於個別字的頻率）
  - mutal information I(hot, dog) = P(hot dog) / P(hot) P(dog)
  - Dice(hot, dog) = 2 x Count(hot dog) / (Count(hot) + Count(dog))
  - Rank Ratio(hot, dog) = sqrt(ExpRank(_ dog) ExpRank(hot _) / Rank(hot dog)
- Rank Ratio = 片語排名提前的程度（相對性）
- 以 hot dog 為例
- 我們計算 hot dog 的排名 (actual rank, AR)
- 我們計算 __ dog 和 hot __ 期望排名的乘積 (expected rank, ER)
- Rank Ratio = sqrt(ER)/AR

計算 Rank Ratio

- 以 hot dog 為例
- 第 1 步 ：將 ngram 次數，轉為排名

In [91]:

```
! sort -k2nr -t $'\t' count_2w.txt | cut -f 1 | awk '{print NR "\t" $0}' > bigram.ra
! head bigram.rank.txt
! tail bigram.rank.txt
```

```
1       of the
2       in the
3       to the
4       on the
5       for the
6       and the
7       to be
8       is a
9       with the
10      from the
286349  to productivity
286350  University shall
286351  enough already
286352  Greece is
286353  capture this
286354  final week
286355  map maps
286356  some estimates
286357  winning service
286358  also helping
```

計算 Rank Ratio

- 第 2 步：計算 _ dog 的排名

計算 Rank Ratio

- 第 2 步：計算 __ dog 和 hot __

In [92]:

```
! grep ' dog$' bigram.rank.txt | sed -n '6,10p;11q'
print ()
! egrep '\thot ' bigram.rank.txt | sed -n '20,25p;26q'
```

```
61991   and dog
64570   of dog
64920   sex dog
72919   hot dog
86540   The dog

60122   hot hot
65399   hot naked
68441   hot mature
70797   hot pussy
71488   hot chocolate
72919   hot dog
```

計算 Expected Rank

- 第 2 步：計算 __ dog 和 hot __ 的平均排名

In [93]:

```
! grep '\thot ' bigram.rank.txt | cut -f 1 | jq -s add/length
! grep ' dog$' bigram.rank.txt | cut -f 1 | jq -s add/length

import math
 print()
print ('ER =', math.sqrt(146988*175492))
```

```
  File "<ipython-input-93-745019207a0c>", line 5
    print()
    ^
IndentationError: unexpected indent
```

計算 Actual Rank

- 第 3 步：計算 hot dog 的排名（Actual Rank, AR)

In [ ]:

```
! grep 'hot dog$' bigram.rank.txt
```

計算簡化版的 Rank Ratio

- 第 4 步 RR = ER/AR

In [ ]:

```
print (160608/72919)
```

實驗資料

- Google Web 1T unigram (`count_1w.txt`)
- Google Web 1T bigram (`count_2w.txt`)
- 紅樓夢全書160 回 (`count_2w.txt`)

本週任務

Level A: 用 MutInfo 計算 Count_2w.txt 中的二字 ngrams 中的術語 (80分)

Level B: 用 RankRatio 計算 Count_2w.txt 中的二字 ngrams 中的術語 (100分)

Level C: 用 RankRatio 來做非督導性紅樓夢斷詞 (加分題、期末專題)

- 計算紅樓夢全書的連續二字的 RankRatio 以及 Mutual Information
- 製作 Sproat and Shih (1990, page 339) 進行全書的斷詞
- 統計紅樓夢全書的詞頻統計表
- 加分題：考慮辭典（例如 光頭 ， 赤腳)、構詞規則（例如 光著頭 ， 赤著腳)

In [ ]:

```

```