中華民國計算語言學學會
通訊 ACLCLP
發行人：吳宗憲
主　編：王新民
執行編輯：黃　琪
The Association for Computational Linguistics and Chinese Language Processing 第二十一卷第四期
E-Mail：aclclp@hp.iis.sinica.edu.tw
Website : http://www.aclclp.org.tw
傳 眞：(02)2788-1638
地 址：台北市研究院路二段128號中研院資訊所
電 話：(02)2788-3799 ext:1502
劃 撥：19166251

本期要目

第十屆博碩士論文得獎名單

博士論文獎

優等獎一名：獲獎金二萬元及獎狀

得獎姓名：闕壯華 (成功大學資訊工程所)
中文題目：強健性語言模型於語音辨識之研究
英文題目：Flexible Language Models for Speech Recognition
指導教授：簡仁宗 教授

佳作獎一名：從缺

碩士論文獎

優等獎一名：從缺

佳作獎三名：獲獎金伍仟元及獎狀

1. 得獎姓名：潘靜芬 (臺灣師範大學英語學系)
中文題目：漢語動詞語意特指之量度：語料庫爲本的計量研究
英文題目：Measuring the Semantic Specificity in Mandarin Verbs: A Corpus-based Quantitative Survey
指導教授：謝舒凱 教授
2. 得獎姓名：蔡財祿 (交通大學電信工程所)
中文題目：國客雙語語音辨認
英文題目：A study on Mixed Hakka-Mandarin Chinese Bilingual Speech Recognition
指導教授：陳信宏 教授
3. 得獎姓名：林信宏 (成功大學外國語文學系)
中文題目：從語料庫語言學探究當代英文專利：專利範圍獨立項數的語言特徵
英文題目：Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents
指導教授：謝菁玉 教授

ROCLING-2010

由國立暨南國際大學資訊工程學系、電機工程學系、及本會共同主辦的「第二十二屆自然語言與語音處理研討會」已於 99/9/2 在南投縣埔里鎮暨南國際大學科技學院第一演講廳順利圓滿結束，參與此次盛會的人士分別來自新加坡及台灣，與會人數多達 160 人次。本次會議共收錄了 17 篇口頭報告論文及 10 篇海報論文。蔡佩珊小姐、沈涵平先生、及吳宗憲教授共同著作之「發音事件驗證於多語辨識發音變異模型之產生」獲得最佳論文獎，會議閉幕式中，分別獲頒獎狀乙紙，並共同獲頒獎金伍仟元。會議論文已建置在 ACL Anthology(http://aclweb.org/anthology-new/) 及本會網站(http://www.aclclp.org.tw/pub_proce_c.php)。

Lunch
Session 2: IR Models **Relevance Ranking using Kernels** *Jun Xu1, Hang Li, Chaoliang Zhong* Microsoft Research Asia **Mining YouTube to Discover Hate Videos, Users and Hidden Communities** *Ashish Sureka, Ponnurangam Kumaraguru, Atul Goya, Sidharth Chhabra* Indraprastha Institute of Information Technology, Delhi (IIIT-D), and Delhi Technological University (DTU) **Title-based Product Search - Exemplified in a Chinese E-commerce Portal** *Chien-Wen Chen and Pu-Jen Cheng* National Taiwan University **Relevance Model Revisited: With Multiple Document Representations** *Ruey-Cheng Chen, Chiung-Min Tsai, Jieh Hsiang* National Taiwan University
Session 3: User Studies and Evaluation **Effective Time Ratio: A measure for Web search engine with document snippet** *Jing He, Baihan Shu, Xiaoming Li, Hongfei Yan* Peking University **Investigating Characteristics of Non-click Behavior Using Query Logs** *Ting Yao, Min Zhang, Yiqun Liu, Shaoping Ma, Yongfeng Zhang, Liyun Ru* Department of C.S.T, Tsinghua University **Score Estimation, Incomplete Judgments, and Significance Testing in IR Evaluation** *Sri Devi Ravana and Alistair Moffat* University of Melbourne and University of Malaya
Reception and Poster Session **Multi-Search: A Meta-Search Engine Based on Multiple Ontologies** *Mohammed Maree, Saadat Alhashmi, Hidayat Hidayat, Bashar Tahayna* Monash University **Co-HITS-Ranking Based Query-Focused Multi-Document Summarization** *Po Hu, Donghong Ji, Chong Teng* Wuhan University, Huazhong Normal University, and Wuhan University **Advanced Training Set Construction for Retrieval in Historic Documents** *Andrea Ernst-Gerlach and Norbert Fuhr* University of Duisburg-Essen **Ontology Driven Semantic Digital Library** *Shahrul Azman Noah, Nor Afni Raziah Alias, Nurul Aida Osman, Zuraidah Abdullah, Nazlia Omar, Yazrina Yahya, Maryati Mohd Yusof* University Kebangsaan Malaysia **Revisiting Rocchio's Relevance Feedback Algorithm for Probabilistic Models** *Zheng Ye, Ben He, Xiangji Huang, Hongfei Lin* York University, Dalian University of Technology **When Two is Better than One: A Study of Ranking Paradigms and Their Integrations for Subtopic Retrieval** *Teerapong Leelanupab, Guido Zuccon, Joemon M. Jose* University of Glasgow

Top-down and Bottom-up: A Combined Approach to Slot Filling
Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Marissa Passantino, Heng Ji
City University of New York

Relation Extraction between Related Concepts by Combining Wikipedia and Web Information for Japanese Language
Masumi Shirakawa, Kotaro Nakayama, Eiji Aramaki, Takahiro Hara, Shojiro Nishio
Osaka University, The University of Tokyo

A Chinese Sentence Compression Method for Opinion Mining
Shi Feng, Daling Wang, Ge Yu, Binyang Li, Kam-Fai Wong
Northeastern University, China and The Chinese University of Hong Kong

Relation Extraction in Vietnamese Text using Conditional Random Fields
Rathany Chan Sam, Huong Thanh Le, Thuy Thanh Nguyen, The Minh Trinh
School of Information and Communication Technology Hanoi University of Technology, Vietnam, and Center for Training of Excellent Students Hanoi University of Technology, Vietnam

A Sparse L2-Regularized Support Vector Machines for Large-scale Natural Language Learning
Yu-Chieh Wu, Yue-Shi Lee, Jie-Chi Yang, Show-Jane Yen
Ming Chuan University, National Central University

An Empirical Comparative Study of Manual Rule-based and Statistical Question Classifiers on Heterogeneous Unseen Data
Cheng-Wei Lee, Min-Yuh Day, Wen-Lian Hsu
Institute of Information Science, Academia Sinica, Taiwan

Constructing Blog Entry Classifiers using Blog-level Topic Labels
Ken Hagiwara, Hiroya Takamura, Manabu Okumura
Tokyo Institute of Technology

Finding Hard Questions by Knowledge Gap Analysis in Question Answer Communities
Ying-Liang Chen and Hung-Yu Kao
National Cheng Kung University

Exploring the Visual Annotatability of Query Concepts for Interactive Cross-Language Information Retrieval
Yoshihiko Hayashi, Masaaki Nagata, Bora Savas
Osaka University, NTT Communication Science Laboratories

A Diary Study Based Evaluation Framework for Mobile Information Retrieval
Ourdia Bouidghaghen, Lynda Tamine, Mohand Boughanem
IRIT-University Paul Sabatier, Toulouse

Dynamics of Genre and Domain Intents
Shanu Sushmita, Benjamin Piwowarski, Mounia Lalmas
University of Glasgow

Query Recommendation Considering Search Performance of Related Queries
Yufei Xue, Yiqun Liu, Tong Zhu, Min Zhang, Shaoping Ma, Liyun Ru
Tsinghua University

Friday, December 3, 2010

Session 8:

Mining parallel documents across Web sites
Pham Ngoc Khanh and Ho Tu Bao
Japan Advanced Institute of Science and Technology

A Revised SimRank Approach for Query Expansion
Yunlong Ma, Hongfei Lin, Song Jin
Dalian University of Technology, Dalian , China

Improving Web-Based OOV Translation Mining for Query Translation
Yun Dong Ge, Yu Hong, Jian Min Yao, Qiao Ming Zhu
Soochow University

On a Combination of Probabilistic and Boolean IR Models for Question Answering
Masaharu Yoshioka
Hokkaido University

Session 9: NLP for IR

A Two-Stage Algorithm for Domain Adaptation with Application to Sentiment Transfer Problems
Qiong Wu, Songbo Tan, Miyi Duan, Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences, China

Doamin-Specific Term Rankings Using Topic Models
Zhiyuan Liu and Maosong Sun
Tsinghua University

Learning Chinese Polarity Lexicons by Integration of Graph Models and Morphological Features
Bin Lu, Yan Song, Xing Zhang, Benjamin K. Tsou
City University of Hong Kong

Lunch & Closing Session

ACLCLP IR Workshop

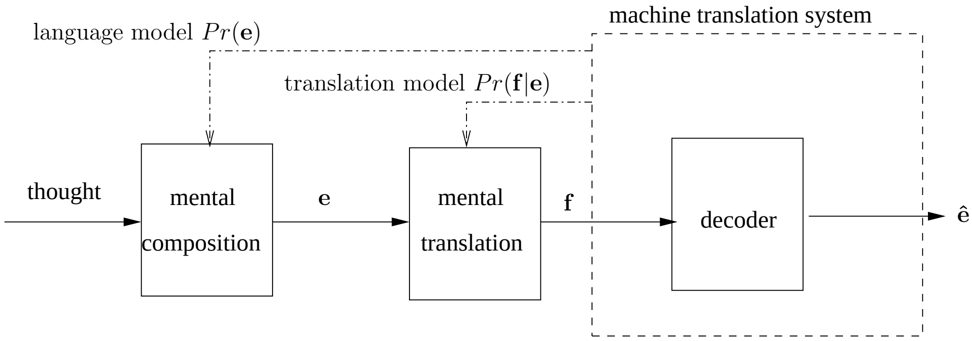

Fig. 1. Imaginative scheme for machine translation. A person's thought is mentally composed in English, and translated to French. The decoder is a machine translation system designed to minimize the probability of error $Pr(\hat{\mathbf{e}} \neq \mathbf{e})$.

- to propose adequate models for $Pr(\mathbf{e})$ and $Pr(\mathbf{f}|\mathbf{e})$;
- to estimate the parameters in the proposed models;
- to search for the optimal candidate $\hat{\mathbf{e}}$.

The IBM models are special cases of translation models $Pr(\mathbf{f}|\mathbf{e})$. Note it is not important for $Pr(\mathbf{f}|\mathbf{e})$ to concentrate on well-formed French sentences, as a well-formed $\mathbf{f}$ will always be given in a translation from French to English. That is why we are going to see a few strangely constructed $\mathbf{f}$ in the development of the theory.

II. Alignment

Assuming certain readers are familiar with the automatic speech recognition (ASR), I am going to draw an analogy*. In ASR, the training data for the acoustic model comes in pairs, with each pair consisting of a waveform and a phoneme (or word) sequence. It is not unusual that the phoneme boundary times in the

*An alerted reader has probably already noticed that (1) has the same form as the fundamental equation of ASR

$$\hat{W} = \arg\max_{W} \quad Pr(W)Pr(A|W),$$

where $Pr(\mathbf{e})$ is replaced by the language model $Pr(W)$, and $Pr(\mathbf{f}|\mathbf{e})$ is replaced by the acoustic model $Pr(A|W)$. In fact, both equations are instances of the noisy-channel communication scenario. In speech recognition, a speaker (source) has some text in mind, then he generates speech waveform for the text. The recognizer has to decode the hidden text based on the observed waveform. In machine translation, a person (source) thinks in English, but he generates French for the thought in English. The translator has to decode the hidden English based on the seen French. Fred Jelinek was the leader of the IBM research group at the times these models are proposed. He did his Ph.D. thesis in information theory under Robert Fano in MIT. It is not coincidental that such a information-theoretic thinking plays fundamental roles in modern statistical language and speech processing.

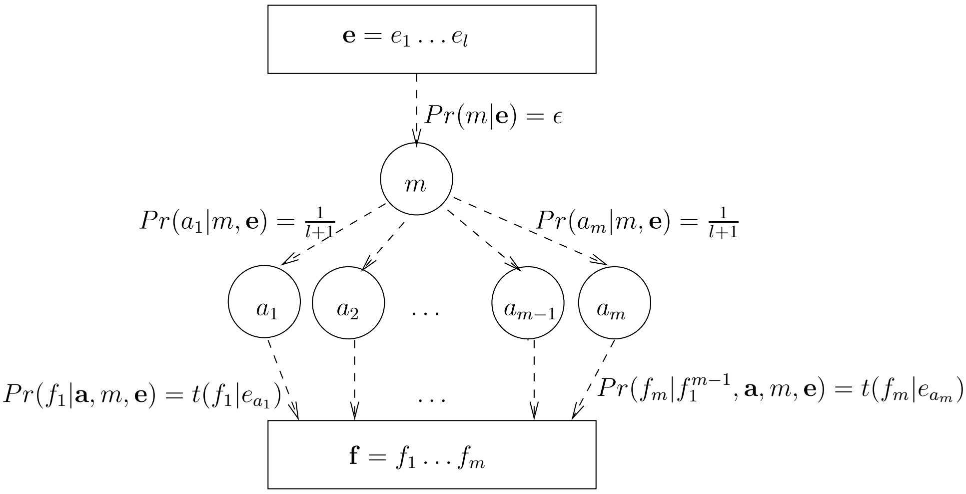

Fig. 2. The generating process of Model 1.

III. Model 1

Referring to the general probability factorization (3), in Model 1 it is assumed that

- $\epsilon \triangleq Pr(m|\mathbf{e})$ is independent of m and $\mathbf{e}$;
- $Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$ depends only on l, and consequently must be $(l+1)^{-1}$;
- $Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e})$ depends only on f_j and e_{a_j}, thus defining a *translation probability*

$$t(f_j|e_{a_j}) \triangleq Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}). \tag{6}$$

With these assumptions, (3) becomes

$$Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}), \tag{7}$$

and the "likelihood" of the parallel sentences $(\mathbf{f}|\mathbf{e})$ is given by

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}). \tag{8}$$

The translation probabilities $t(f|e)$ are estimated to maximize $Pr(\mathbf{f}|\mathbf{e})$ subject to the constraints that

$$\sum_{f} t(f|e) = 1, \quad \forall e. \tag{9}$$

The generating process is depicted in Fig. 2.

An iterative algorithm can be used to estimate $t(f|e)$, given an initial estimate and a training set of parallel sentences. The basic idea of iteration is as follows.

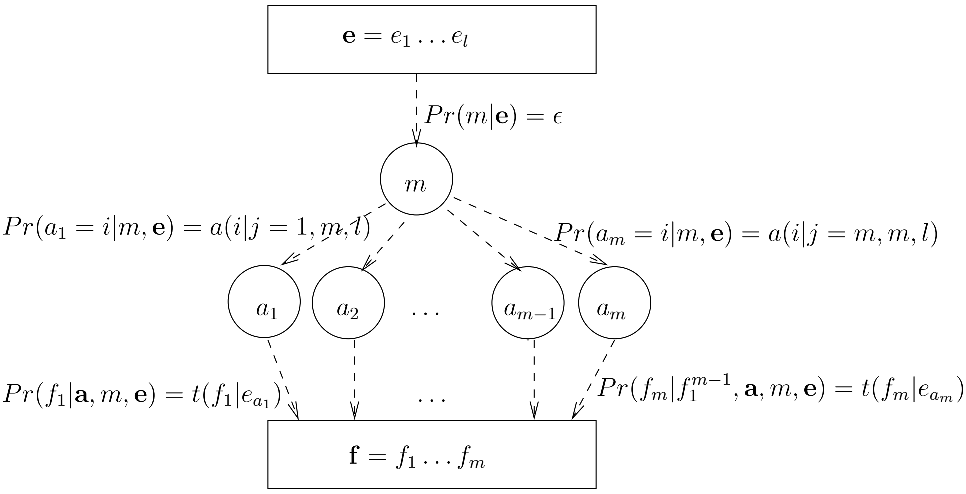

Fig. 3. The generating process of Model 2. Compared to Model 1, the alignment probability is modified.

The aforementioned iterative algorithm to estimate $t(f|e)$ can be adapted to estimate $t(f|e)$ and $a(i|j, m, l)$ jointly.

Note that Model 1 is a special case of Model 2, so the parameters of Model 2 can be initialized by the parameters of Model 1. Specifically, one can compute the alignment probability by Model 1 with $t(f|e)$, and then collect the required counts to initialize $a(i|j, m, l)$ of Model 2.

V. Fertility and Permutation

Another generating process from given $\mathbf{e}$ to $\mathbf{f}$ is as follows. The number of words the word e_i in $\mathbf{e}$ generates is called the **fertility** of e_i, denoted by Φ_{e_i}, and sometimes abbreviated by Φ_i when there is no ambiguity. The list of words for e_i is denoted by T_i, called the **tablet** of e_i. The k-th word in T_i is denoted by T_{ik}. The collection of T_i is denoted by $\mathbf{T}$, called the **tableau** of $\mathbf{e}$. The words in a tableau are permuted to produce $\mathbf{f}$. The **permutation** is denoted by $\mathbf{\Pi}$, in which the position of the word T_{ik} is denoted by Π_{ik}. Note that from instantiations of tableau $\mathbf{T} = \tau$ and permutation $\mathbf{\Pi} = \pi$, the corresponding instantiations of alignment $\mathbf{a}$ and French string[†] $\mathbf{f}$ are determined.

According to this generating process, the conditional probability of $T = \tau, \Pi = \pi$ given $\mathbf{e}$ can be

[†]Note we say "string" instead of "sentence" for reasons to be stated later.

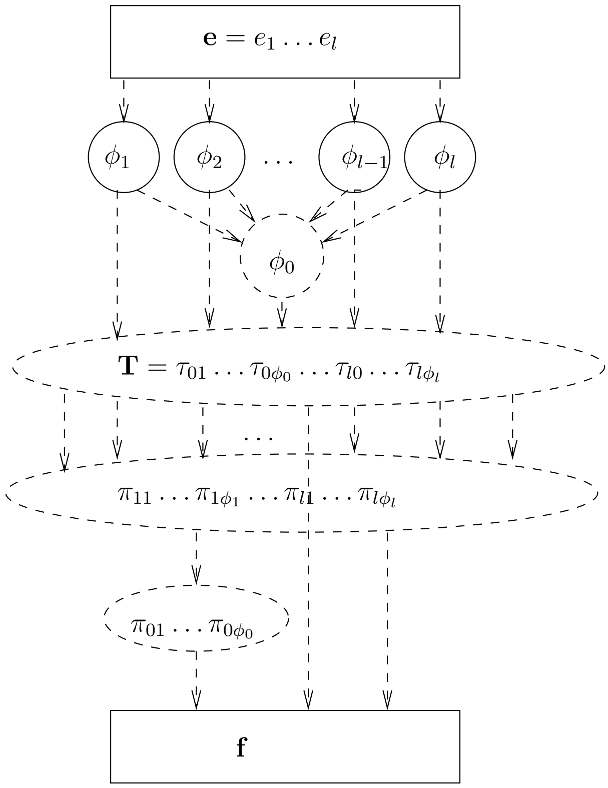

Fig. 4. The generating process based on fertility and permutation. This is the basis for Models $3-5$.

A pair of instances of tableau and permutation $(\mathbf{T}=\tau, \mathbf{\Pi}=\pi)$ correspond to a unique pair of string and alignment $(\mathbf{f}, \mathbf{a})$. With the assumed probability functions, (15) becomes

$$\begin{aligned} Pr(\tau,\pi|\mathbf{e}) = &\prod_{i=1}^{l} n(\phi_i|e_i) \binom{\phi_1+\cdots+\phi_l}{\phi_0} p_0^{\phi_1+\cdots+\phi_l-\phi_0} p_1^{\phi_0} \times \\ &\prod_{j=1}^{m} t(f_j|e_{a_j}) \times \\ &\prod_{j=1}^{m} d(j|a_j,m,l) \times \\ &\frac{1}{\phi_0!}, \end{aligned} \tag{19}$$

where f_j is the French word in the j-th position of $\mathbf{f}$, a_j is the position of the English word that f_j is aligned to, and m is the length of $\mathbf{f}$. The display of (19) purposely parallels (15) for the readers to follow the correspondence.

- $\mathcal{N}(\mathbf{a})$ is the set of all neighbors of $\mathbf{a}$;
- $b^{\infty}_{i\leftarrow j}(\mathbf{a})$: the alignment of convergence in the series $b^{k+1}_{i\leftarrow j}(\mathbf{a}) \triangleq b_{i\leftarrow j}(b^{k}_{i\leftarrow j}(\mathbf{a}))$, where $b_{i\leftarrow j}(\mathbf{a})$ is the neighbor of $\mathbf{a}$ with the maximum posterior probability and ij is pegged;

VII. Deficiency

The probability factorization for $Pr(\tau, \pi|\mathbf{e})$ as shown in (19) enables us to quickly compute the posterior probabilities of the neighbors of an alignment, which is crucial in the approximation for the parameter estimation of Model 3.

As is pointed out in Section VI, one issue about Model 3 is that it is **deficient**. In Model 3, part of the probability mass is assigned to the generalized French strings. In fact, Models $1-2$ assign probability to sentences that are not well-formed, so they are also deficient in a different sense.

Note that deficiency is merely an “issue” rather than a “problem”, (or a “warning” but not a “bug”), as in the current translation direction from French to English, a well-formed French sentence $\mathbf{f}$ will always be given. Under the circumstances, probabilities computed using Models $1-3$ are proportional to the conditional probabilities that $\mathbf{f}$ is a well-formed sentence, so it is not a problem.

VIII. Model 4

It is noted that in Model 3, the movement of a long phrase will incur large *distortion penalty* (i.e. low probability) as each word in the phrase is treated the same way as moving independently. However, it is common sense (to linguists, at least) that the words constituting a phrase tend to move around a sentence jointly, rather than independently. Therefore, in Model 4, the probability model for distortion is modified to allow easier phrase movements than in Model 3.

In Model 3, an English word, say e_i, generates a tablet of ϕ_i words, $\tau_{i1}, \ldots, \tau_{i\phi_i}$. If $\phi_i > 0$, e_i is an one-word **cept**[‖] and the corresponding ϕ_i words aligned to e_i constitute a phrase in a loose sense.

In Model 4, two sets of probability are introduced to make the joint movement of the French words corresponding to a one-word cept easier:

- the probability to place the first word, called the head word, in the one-word cept;
- the probability to place the remaining words, if any;

For the head word, the probability for placing the head word of the i-th one-word cept is

$$Pr(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \triangleq d_{=1}(j - \Theta_{i-1}|\mathcal{A}(e_{[i]-1}), \mathcal{B}(f_j)), \quad [i] > 0. \tag{23}$$

[‖]A **cept** is a fraction of a **con-cept**.

For the non-head words, the probability for placing the k-th word of the i-th one-word cept is

$$\begin{aligned} & Pr(\Pi_{[i]k} = j | \pi_{[i]1}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \\ & \triangleq d_{>1}(v_j - v_{\pi_{[i]k-1}} | \mathcal{B}(f_j), v_m - v_{\pi_{[i]k-1}} - \phi_{[i]} + k)(1 - \delta(v_j, v_{j-1})), \quad [i] > 0, k > 1. \end{aligned} \tag{27}$$

A set based on and trimmed from the set defined by (25) is used to gather the counts required for the parameter estimation in Model 5.

Both Models 3 and 4 are deficient. From (26) and (27), we make sure that at any point of the generating process from $\mathbf{e}$ to $\mathbf{f}$, the word to be placed must occupy a vacant position. Thus Model 5 is no longer deficient.

X. Conclusion

In this article, I try to convince the readers that machine translation is an interesting problem, by going through the classic paper by Brown et al. I hope the readers can enjoy the mathematical treatment as much as I did when I first came across it a decade ago. I was truly thrilled to see that mathematics, statistics, and engineering can be combined so beautifully to tackle the real problem of machine translation.

Peter Brown and Bob Mercer left IBM and joined the Renaissance Technologies, which stands today as the richest hedge fund investment company, shortly after they published this paper. They are co-CEOs as of the year of 2010. For another example for the variety of achievements by the people working on machine translation, I will add that Krzysztof Jassem [3][4] from Poland, is a world life master in the game of bridge.

XI. Epilogue

While writing this article, I heard about the sad news that Fred Jelinek passed away (18 November 1932 - 14 September 2010). Professor Jelinek was a critical character in applying statistical approaches for machine translation [5]. According to himself, he actually stumbled upon speech and language processing. Nonetheless, I believe he is one of the greatest founders of modern automatic speech recognition and machine translation with the statistical methodology. I have the impression that he has ways to explain statistical automatic speech recognition clearly [6].

References

[1] P. F. Brown, V. J. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.

[2] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.