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：語料庫為本的計量研究
   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)。

The Sixth Asia Information Retrieval Societies Conference
2010 亞洲資訊檢索研討會(AIRS 2010)

Call for Participation

http://irlab.csie.ntu.edu.tw/airs2010/

會議時間：2010 年 12 月 1~3 日

會議地點：台北市台灣大學應用力學所國際會議廳

AIRS (Asia Information Retrieval Societies Conference)為亞洲地區資訊檢索領域最主要的會議，該會由 2004 年開始，至今已舉辦過五屆，舉辦地區包含大陸、韓國、新加坡及日本，歷屆 AIRS 研討會不但出席踴躍，同時也都相當成功，亞洲區資訊檢索領域之重要學者專家都將參加，與會者將有機會了解資訊檢索研究領域中最爲重要且尖端的研究課題、最新的技術及研究成果。

AIRS 2010 亞洲資訊檢索研討會訂於 12 月 1~3 日在台灣大學舉行，今年一般論文接受率約爲 22%，其主題涵蓋各種不同的資訊檢索技術和應用，包含資訊檢索理論模型、效能評估和驗證、資料分類及分群、多媒體資訊檢索、自然語言及機器學習在資訊檢索的應用等。

爲促進國內外學者專家的討論及交流，AIRS 2010 將結合「中華民國計算語言學學會 IR Workshop」共同舉行，***「註冊 AIRS 即可免費參加 IR Workshop***」，IR Workshop 主要邀請國內外相關學者專家演講，係繼 2002 年「資訊自動分類技術研討會」、2003 年「資訊檢索與電腦輔助語言教學研討會」、2004 年「文件探勘技術研討會」、2005 年「網路資訊檢索技術與趨勢研討會」、2006 年「網路探勘技術與趨勢研討會」、2007 年「Web 2.0 技術與應用研討會」、2008 年「網路社群服務計算暨探勘技術研討會」以及 2009 年「行動資訊檢索暨行動定位服務技術研討會」之後續的年度會議活動。

歡迎各界人士踴躍參加 AIRS 2010。

線上註冊網址: http://irlab.csie.ntu.edu.tw/airs2010/reg.php

Program:

Wednesday, December 1, 2010
Invited Talk
Session 1: Machine Learning **Multi-viewpoint based similarity measure and optimality criteria for document clustering** *Thang D. Nguyen, Lihui Chen, Keong C. Chan* Nanyang Technological University, Singapore **A Text Classifier with Domain Adaptation for Sentiment Classification** *Wei Chen and Jingyu Zhou* Shanghai Jiao Tong University

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

Connecting qualitative and quantitative analysis of Web search process: Analysis using Search Units
Hitomi Saito, Masao Takaku, Yuka Egusa, Hitoshi Terai, Makiko Miwa, Noriko Kando
Aichi University of Education, National Institute for Materials Science,
National Institute for Educational Policy Research, Tokyo Denki University,
The Open University of Japan, and National Institute of Informatics

Transliteration Retrieval Model for Cross Lingual Information Retrieval
Ea-Ee Jan, Shih-Hsiang Lin, Berlin Chen
IBM T.J. Watson Research Center and National Taiwan Normal University

The Role of Lexical Ontology in Expanding the Semantic Textual Content of On-Line News Images
Shahrul Azman Noah and Datul Aida Ali
University Kebangsaan Malaysia

Order Preserved Cost-sensitive Listwise Approach in Learning to Rank
Min Lu, MaoQiang Xie, Yang Wang, Jie Liu, YaLou Huang
Nankai University, Tianjin, China

Pseudo-Relevance Feedback Based on mRMR Criteria
Yuanbin Wu, Qi Zhang, Yaqian Zhou, Xuanjiang Huang
Fudan University

An Integrated Deterministic and Nondeterministic Inference Algorithm for Sequential Labeling
Yu-Chieh Wu, Yue-Shi Lee, Jie-Chi Yang, Show-Jane Yen
National Central University, Ming-Chuan University

FolkDiffusion: A Graph-based Tag Suggestion Method for Folksonomies
Zhiyuan Liu, Chuan Shi, Maosong Sun
Tsinghua University

Effectively Leveraging Entropy and Relevance for Summarization
Wenjuan Luo, Fuzhen Zhuang, Qing He, Zhongzhi Shi
Institute of Computing Technology, Chinese Academy of Sciences

Machine Learning Approaches for Modeling Spammer Behavior
Md. Saiful Islam, Abdullah Al Mahmud, Md. Rafiqul Islam
University of Dhaka, Bangladesh, Ahsanullah University of Science and Technology, Bangladesh, and Deakin University, Australia

Research of Sentiment Block Identification for Customer Reviews Based on Conditional Random Fields
Lei Jiang, Yuanchao Liu, Bingquan Liu, Chengjie Sun, Xiaolong Wang
School of Computer Science and Technology, Harbin Institute of Technology

Semantic Relation Extraction Based on Semi-supervised Learning
Haibo Li, Yutaka Matsuo, Mitsuru Ishizuka
University of Tokyo

Corpus-based Arabic Stemming using N-grams
Abdelaziz Zitouni, Asma Damankesh, Foroogh Barakati, Maha Atari,
Mohamed Watfa, Farhad Oroumchian
University of Wollongong in Dubai

Analysis and Algorithms for Stemming Inversion
Ingo Feinerer
Vienna University of Technology

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

Thursday, December 2, 2010

Invited Talk

Session 4: NLP for IR

A Local Generative Model for Chinese Word Segmentation
Kaixu Zhang and Maosong Sun
Department of Computer Sci. & Tech., Tsinghua University

Re-ranking Summaries Based on Cross-document Information Extraction
Heng Ji, Juan Liu, Benoit Favre, Dan Gillick, Dilek Hakkani-Tur
City University of New York, LIUM, Université du Maine, University of California, Berkeley

Lunch

Session 5: Machine Learning 1

Learning to Rank with Supplementary Data
Wenkui Ding, Tao Qin, Xu-Dong Zhang
Tsinghua University, Microsoft Research Asia

Event Recognition from News Webpages through Latent Ingredients Extraction
Rui Yan, Yu Li, Yan Zhang, Xiaoming Li
Peking University

Tuning Machine-Learning Algorithms for Battery-Operated Portable Devices
Ziheng Lin, Yan Gu, Samarjit Chakraborty
National University of Singapore

Session 6: Multimedia

Emotion Tag Based Music Retrieval Algorithm
Jing Li, Hongfei Lin, Lijuan Zhou
Dalian University of Technology

An Aesthetic-based Approach to Rank Web Images
Shao Hang Kao, Wei-Yen Day, Pu-Jen Cheng
National Taiwan University

Session 7: IR Models 2

A Unified Iterative Optimization Algorithm for Query Model and Ranking Refinement
Yunping Huang, Le Sun, Jian-Yun Nie
IS, Chinese Academy of Sciences, and Universite' de Montre'al

A Study of Document Weight Smoothness in Pseudo Relevance Feedback
Peng Zhang, Dawei Song, Xiaochao Zhao, Yuexian Hou
Robert Gordon University, UK, and Tianjin University, China

Modeling Variable Dependencies between Characters in Chinese Information Retrieval
Lixin Shi and Jian-Yun Nie
University of Montreal

Banquet

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

Machine Translation: A Score Years Ago

Chia-Ping Chen

Abstract

In this article, I will review a classic paper on 5 statistical models, also known as the IBM Models, of machine translation. These models are presented in the order of complexity. In this way, a reader can clearly see the incremental improvements, by understanding the critical issues in the old models that the new models try to address. Although the paper was written almost twenty years ago, to me the joy of reading it has not faded over the years.

Index Terms

machine translation, IBM models

I. Introduction

The methodology for treating the machine translation problem in the paper by Brown et al. [1] is a statistical one. Therein, the fundamental equation of machine translation is given by

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \quad Pr(\mathbf{e})Pr(\mathbf{f}|\mathbf{e}), \tag{1}$$

where $\mathbf{f}$ is a sentence in French, and $\mathbf{e}$ is a candidate sentence in English. $Pr(\mathbf{e})$ is called the language model, and $Pr(\mathbf{f}|\mathbf{e})$ is called the translation model. It is important to note that the direction of translation is from French to English in (1). The translation in the opposite direction is an entirely different problem.

In order to understand (1), it may help to follow an imaginative scheme: Believe it or not, the creator of a French text thinks in English! He first mentally composes the English text, denoted by $\mathbf{e}$, for his thought. Then he mentally translate the English text to French, denoted by $\mathbf{f}$. The task of machine translation is to come up with methods to decide $\hat{\mathbf{e}}$ based on $\mathbf{f}$ such that the probability that $\hat{\mathbf{e}} \neq \mathbf{e}$ is minimized. This is illustrated in Fig. 1.

We can see from (1) that there are three core problems in this formulation as follows:

Chia-Ping Chen is with the Department of Computer Science and Engineering, National Sun Yat-Sen University. Address: 70 Lien-Hai Road, Kaohsiung, Taiwan 804; Phone: +886.7.525.2000; Fax: +886.7.525.4301; Email: cpchen@mail.cse.nsysu.edu.tw

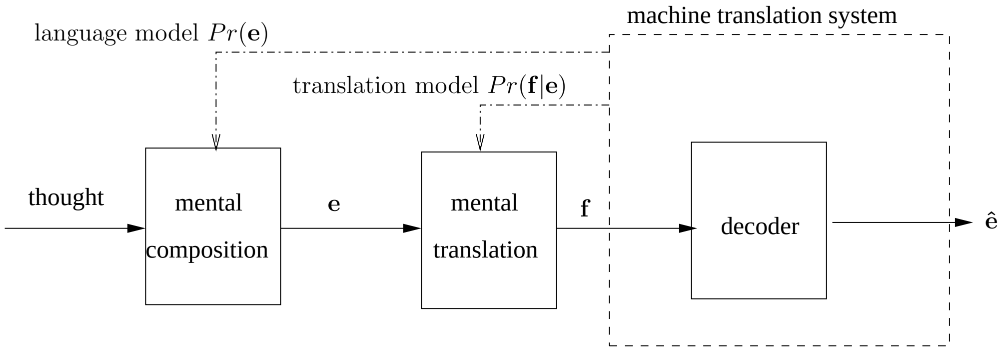

Fig. 1. Imaginative scheme for machine translation. A person's thought is mentally composed in English, and translated to French. The decoder is a machine translation system designed to minimize the probability of error $Pr(\hat{\mathbf{e}} \neq \mathbf{e})$.

- 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.

waveform are left unspecified, and somehow we need to decide the detailed correspondence between the waveform segments and the phonemes. This detailed correspondence is known as the "alignment", and we have the operation known as "forced alignment" to estimate the correspondence. In machine translation (MT), the training data for the translation model also comes in pairs, with each pair consisting of a sentence $\mathbf{f}$ in French and a sentence $\mathbf{e}$ in English. Therefore, for each word e in $\mathbf{e}$, we would like to know the corresponding words in $\mathbf{f}$. This correspondence essentially manifests the same idea as the alignment in ASR.

The alignment in MT for the translation model is slightly more complicated than the alignment in ASR for the acoustic model. In ASR, the alignment is almost always left-to-right. In MT, on the other hand, the correspondence are often out-of-order, and the words corresponding to the same word may be non-contingent. Therefore, MT necessarily requires a more complicated scheme of alignment than ASR.

"Words" may appear to be natural enough to be the labeling units for sentences. However, in the later development of machine translation, the "phrase-based" approaches have been proposed [2]. The "phrases" are actually "alignment templates" derived from the alignment between words of parallel sentences. That is the core technology of the Google translator, and would be an interesting subject, but we will not pursuit it in this article.

Treating the sentences $\mathbf{f}, \mathbf{e}$ and the alignment, denoted by $\mathbf{a}$, as random variables, we can write

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}). \tag{2}$$

Assuming $\mathbf{e}$ has l words and $\mathbf{f}$ has m words, without loss of generality, we can factorize the joint probability $Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$ by

$$Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = Pr(m|\mathbf{e}) \prod_{j=1}^{m} Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) Pr(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}), \tag{3}$$

where a_j is the position of the English word that f_j is aligned to, i.e.,

$$e_{a_j} \leftarrow f_j. \tag{4}$$

In (3), it is implicitly assumed that each French word is aligned to at most one English word. Those French words not aligned to any English word is said to be aligned to the "null word", denoted by e_0. From the perspective of an English word e_i, it can be aligned to 0 or multiple French words, which happens if

$$a_j \neq i \quad \forall j, \quad \text{or} \quad a_j = a_{j'} \quad \exists j \neq j'. \tag{5}$$

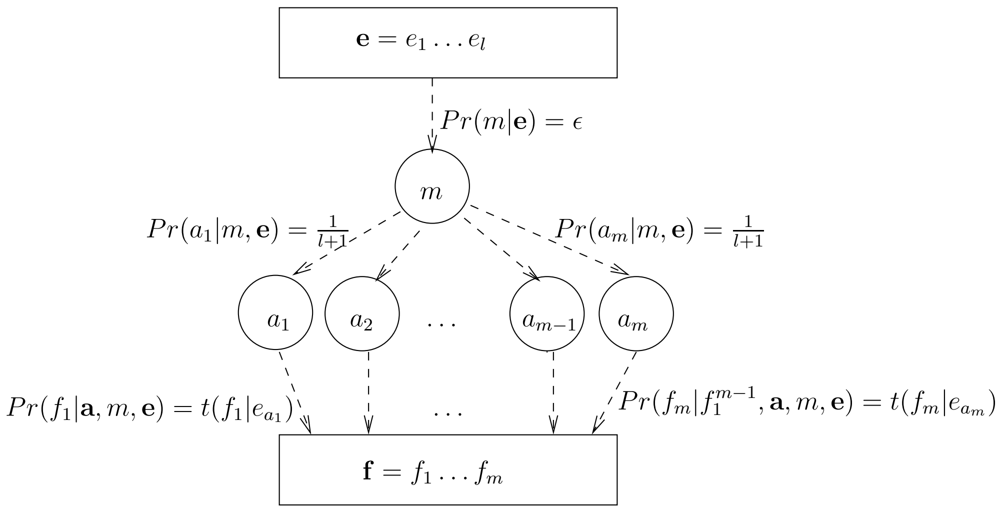

Fig. 2. The generating process of Model 1.

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.

- The word-pair count, denoted by $c(f|e;\mathbf{f},\mathbf{e})$, is accumulated over the set of training parallel sentences, based on the number of co-occurrences of (f,e) and the current estimate of $t(f|e)$;
- These counts are renormalized to update the estimate of $t(f|e)$.

The count of an instance of co-occurrence of e,f is weighted by the posterior probability of an alignment $\mathbf{a}$ in which f is aligned to e. The non-integral count of $Pr(\mathbf{a}|\mathbf{f},\mathbf{e})$ is also known as the "probability count" or the "soft count". From the definition of posterior probability, we have

$$Pr(\mathbf{a}|\mathbf{f},\mathbf{e}) = \frac{Pr(\mathbf{f},\mathbf{a}|\mathbf{e})}{Pr(\mathbf{f}|\mathbf{e})}. \tag{10}$$

In (10), the numerator, the joint probability $Pr(\mathbf{f},\mathbf{a}|\mathbf{e})$, can be straightforwardly computed. For the denominator, the data-likelihood $Pr(\mathbf{f}|\mathbf{e})$, it turns out the summation in (8) can be re-written as

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i). \tag{11}$$

It turns out that (11) makes the computation for the count $c(f|e;\mathbf{f},\mathbf{e})$ exact and efficient, which remains the same way in Model 2.

IV. Model 2

Referring to the general probability factorization (3), in Model 2 it is assumed that

- $\epsilon \triangleq Pr(m|\mathbf{e})$ is independent of m and $\mathbf{e}$ (the same as Model 1);
- $Pr(a_j|a_1^{j-1},f_1^{j-1},m,\mathbf{e})$ depends only on j, a_j, and m, as well as on l, thus defining an *alignment probability*

$$a(a_j|j,m,l) \triangleq Pr(a_j|a_1^{j-1},f_1^{j-1},m,\mathbf{e}); \tag{12}$$

- $Pr(f_j|a_1^{j},f_1^{j-1},m,\mathbf{e})$ depends only on f_j and e_{a_j}, which is modeled by a translation probability $t(f|e)$ (the same as Model 1).

The generating process with the new probability is depicted in Fig. 3. With these assumptions, (3) is reduced to

$$Pr(\mathbf{f}|\mathbf{e}) = \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) a(a_j|j,m,l). \tag{13}$$

Along with the translation probabilities $t(f|e)$, the alignment probabilities $a(a_j|j,m,l)$ are jointly estimated to maximize $Pr(\mathbf{f}|\mathbf{e})$ subject to the constraints that

$$\sum_{i=0}^{l} a(a_j=i|j,m,l) = 1, \quad \forall j,m,l. \tag{14}$$

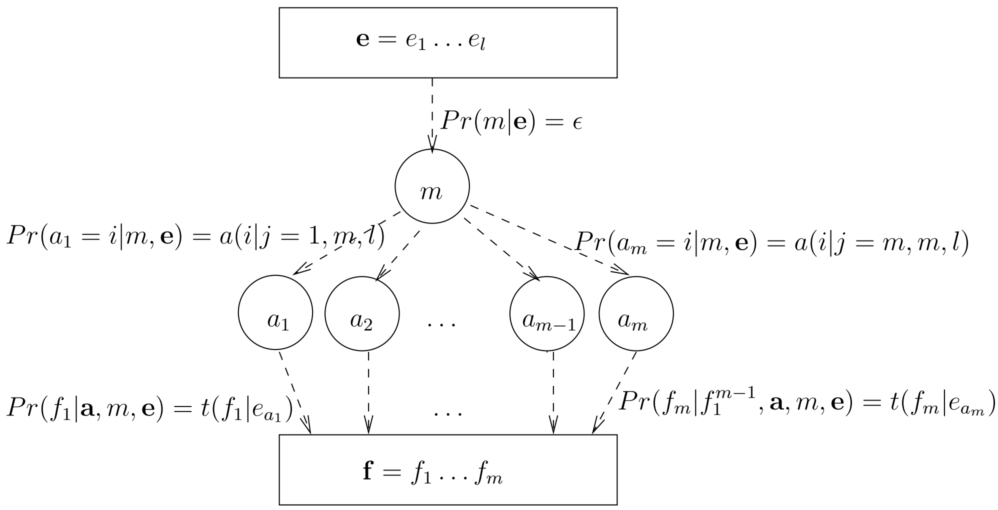

Fig. 3. The generating process of Model 2. Compared to Model 1, the alignment probability is modified.

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.

factorized as

$$
\begin{aligned}
Pr(\tau,\pi|\mathbf{e}) = &\prod_{i=1}^{l} Pr(\phi_i|\phi_1^{i-1},\mathbf{e}) \;\times\; Pr(\phi_0|\phi_1^l,\mathbf{e}) \;\times \\
&\prod_{i=0}^{l}\prod_{k=1}^{\phi_i} Pr(\tau_{ik}|\tau_{i1}^{k-1},\tau_0^{i-1},\phi_0^l,\mathbf{e}) \;\times \\
&\prod_{i=1}^{l}\prod_{k=1}^{\phi_i} Pr(\pi_{ik}|\pi_{i1}^{k-1},\pi_1^{i-1},\tau_0^l,\phi_0^l,\mathbf{e}) \;\times \\
&\prod_{k=1}^{\phi_0} Pr(\pi_{0k}|\pi_{01}^{k-1},\pi_1^l,\tau_0^l,\phi_0^l,\mathbf{e}).
\end{aligned}
\tag{15}
$$

The generating process is depicted in Fig. 4.

It is important to recognize e_0 as the null English word. We use e_0 for those French words not aligned to any English words in Models 1 and 2. It has the same function in the current generating process. In the current generating process, it is used to make the numbers of the words in the tableau sum to m, i.e.,

$$
\Phi_{e_0} = m - \sum_{i=1}^{l} \Phi_{e_i}, \quad \text{or} \quad \phi_0 = m - \sum_{i=1}^{l} \phi_i. \tag{16}
$$

VI. Model 3

Referring to the factorization (15) based on the generation process of fertility and permutation, in Model 3 it is assumed that

- $Pr(\phi_i|\phi_1^{i-1},\mathbf{e})$ for $i = 1,\ldots,l$ depends only on e_i and ϕ_i;
- $Pr(\tau_{ik}|\tau_{i1}^{k-1},\tau_0^{i-1},\phi_0^l,\mathbf{e})$ for $i = 0,\ldots,l$ depends only on τ_{ik} and e_i;
- $Pr(\pi_{ik}|\pi_{i1}^{k-1},\pi_1^{i-1},\tau_0^l,\phi_0^l,\mathbf{e})$ for $i = 1,\ldots,l$ depends only on π_{ik}, i, m, and l;

The corresponding probability functions in Model 3 are

- $n(\phi|e_i) \triangleq Pr(\Phi_{e_i} = \phi|\phi_1^{i-1},\mathbf{e})$ is called the *fertility probability*;
- $t(f|e_i) \triangleq Pr(T_{ik} = f|\tau_{i1}^{k-1},\tau_0^{i-1},\phi_0^l,\mathbf{e})$ is the *translation probability*, the same as in Models $1-2$;
- $d(j|i,m,l) \triangleq Pr(\Pi_{ik} = j|\pi_{i1}^{k-1},\pi_1^{i-1},\tau_0^l,\phi_0^l,\mathbf{e})$ is called the *distortion probability*;
- For the fertility Φ_{e_0}, the probability function is

$$
Pr(\Phi_{e_0} = \phi_0|\phi_1^l,\mathbf{e}) = \binom{\phi_1+\cdots+\phi_l}{\phi_0} p_0^{\phi_1+\cdots+\phi_l-\phi_0} p_1^{\phi_0}, \quad \text{where} \quad p_0 + p_1 = 1. \tag{17}
$$

- For the permutation Π_{0k}, the probability function is

$$
Pr(\Pi_{0k} = j|\pi_{01}^{k-1},\pi_1^l,\tau_0^l,\phi_0^l,\mathbf{e}) = \begin{cases} \frac{1}{\phi_0-(k-1)}, & \text{if } j \text{ is vacant} \\ 0, & \text{otherwise} \end{cases} \tag{18}
$$

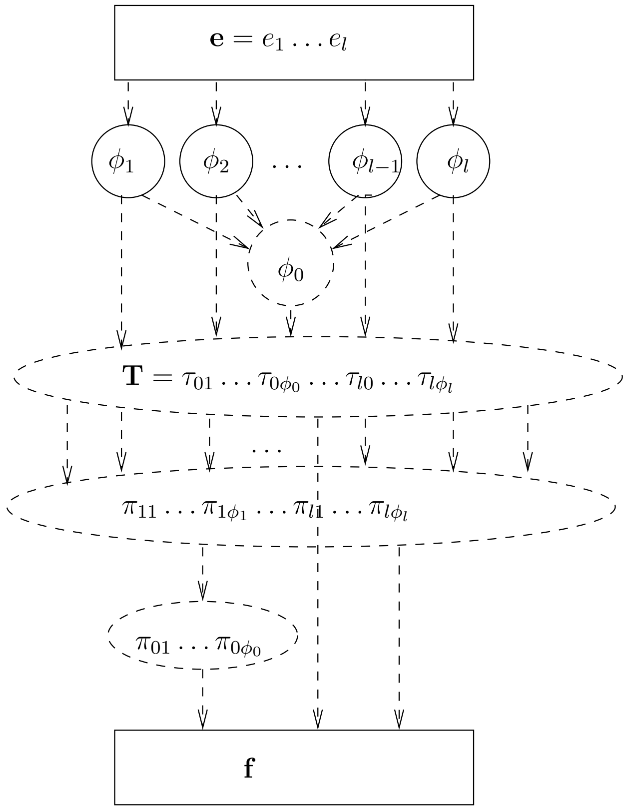

Fig. 4. The generating process based on fertility and permutation. This is the basis for Models $3-5$.

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.

It is interesting to note that in Model 3 the generated string $\mathbf{f}$ is allowed to skip word positions. Such a string is called a *generalized string*. Contrarily, the sentences we have been thinking about are called the *normal strings*, where each position is occupied by exactly one word. The assignment of non-zero probability to the non-normal strings brings up the issue of *deficiency*, which will be addressed in a later model.

The number of *indistinguishable* tableau-permutation pairs for $(\mathbf{f}, \mathbf{a})$ is

$$\prod_{i=0}^{l} \phi_i!. \tag{20}$$

That is, (20) is the total number of pairs of (τ, π) that result in the same $(\mathbf{f}, \mathbf{a})$. Using (20) and (16), we have

$$\begin{aligned} Pr(\mathbf{f}|\mathbf{e}) &= \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \\ &= \sum_{\mathbf{a}} \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} n(\phi_i|e_i) \times \prod_{j=1}^{m} t(f_j|e_{a_j}) d(j|a_j, m, l) \times \prod_{i=1}^{l} \phi_i!. \end{aligned} \tag{21}$$

Unlike Model 1 and Model 2, the counts we need in order to update the probabilities are no longer exactly and efficiently computable. Suffice to say that we fall back to certain approximate schemes to accumulate the counts. Specifically, the summation over the set of all alignments $\mathcal{A}(\mathbf{e}, \mathbf{f})$ between $\mathbf{e}$ and $\mathbf{f}$ is approximated by the summation over a subset $\mathcal{S}$ of $\mathcal{A}(\mathbf{e}, \mathbf{f})$ given by

$$\mathcal{S} = \mathcal{N}(b^{\infty}(V(\mathbf{e}|\mathbf{f}; \mathbf{2}))) \bigcup \cup_{ij} \mathcal{N}(b^{\infty}_{i\leftarrow j}(V_{i\leftarrow j}(\mathbf{e}|\mathbf{f}; \mathbf{2}))), \tag{22}$$

where the meanings of the notations are

- $V(\mathbf{e}|\mathbf{f}; \mathbf{2})$: the alignment $\mathbf{a}$ with the maximum $Pr(\mathbf{a}|\mathbf{e}, \mathbf{f})$ based on Model 2, also called the Viterbi alignment‡;
- $V_{i\leftarrow j}(\mathbf{e}|\mathbf{f}; \mathbf{2})$: the Viterbi alignment in the subset of $\mathcal{A}(\mathbf{e}, \mathbf{f})$ where ij is pegged§;
- $b^{\infty}(\mathbf{a})$: the alignment of convergence in the series $b^{k+1}(\mathbf{a}) \triangleq b(b^k(\mathbf{a}))$, where $b(\mathbf{a})$ is a neighbor¶ of $\mathbf{a}$ with the maximum posterior probability;

‡Instead of Model 3, Model 2 is used because the Viterbi alignment can be obtained efficiently.

§ij is said to be pegged in an alignment $\mathbf{a}$ if $a_j = i$.

¶By definition, two alignments $\mathbf{a}$ and $\mathbf{a}'$

- differ by a move if $a_j \neq a'_j$ for exactly one j;
- differ by a swap if there exist $j \neq j'$ such that $a_j = a'_{j'}, a_{j'} = a'_j$ and $a_k = a'_k$ for $k \neq j, j'$.

$\mathbf{a}'$ is a neighbor of $\mathbf{a}$ if $\mathbf{a}' = \mathbf{a}$, or they differ by a move, or they differ by a swap.

- $\mathcal{N}(\mathbf{a})$ 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**.

Note that in (23)

- $[i]$ denotes the position in the English sentence of the i-th one-word cept (note $[i] \geq i$, since $\phi_{i'}$ could be 0 for some English words $e_{i'}$);
- Θ_i is the center (ceiling of average) of the positions for the French words generated by e_i;
- $j - \Theta_{i-1}$ is called the displacement of cept i, measured from the previous cept;
- $\mathcal{A}(e)$ and $\mathcal{B}(f)$ are the word classes of the English word e and the French word f respectively.

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

$$Pr(\Pi_{[i]k} = j | \pi_{[i]1}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \triangleq d_{>1}(j - \pi_{[i]k-1} | \mathcal{B}(f_j)), \quad [i] > 0, k > 1. \tag{24}$$

Note that in (24), $d_{>1}(n|\mathcal{B}(f)) = 0$ for $n \leq 0$. That is, the condition $\pi_{[i]k} > \pi_{[i]k-1}$ is enforced, meaning the words $\tau_{[i]1}, \ldots, \tau_{[i]\phi_{[i]}}$ in a cept must be placed left-to-right in $\mathbf{f}$.

Again in Model 4, the counts we need in order to update the probabilities are not exactly and efficiently computable. Instead, the summation is over a subset $\mathcal{S}$ of $\mathcal{A}(\mathbf{e}, \mathbf{f})$ given by

$$\mathcal{S} = \mathcal{N}(\tilde{b}^{\infty}(V(\mathbf{e}|\mathbf{f}; \mathbf{2}))) \bigcup \underset{ij}{\cup} \mathcal{N}(\tilde{b}_{i \leftarrow j}^{\infty}(V_{i \leftarrow j}(\mathbf{e}|\mathbf{f}; \mathbf{2}))). \tag{25}$$

The difference between the set (25) used in Model 4 and the set (22) used in Model 3 is $\tilde{b}(\mathbf{a})$ and $b(\mathbf{a})$. Recall that $b(\mathbf{a})$ is the neighbor of the alignment $\mathbf{a}$ with the highest posterior probability $Pr(\cdot|\mathbf{f}, \mathbf{e}; \mathbf{3})$. Here, to find $\tilde{b}(\mathbf{a})$ requires us to firstly rank the neighbors of $\mathbf{a}$ by the posterior probability $Pr(\cdot|\mathbf{f}, \mathbf{e}; \mathbf{3})$, then to look for the highest-ranking neighbor $\mathbf{a}'$ with $Pr(\mathbf{a}'|\mathbf{f}, \mathbf{e}; \mathbf{4}) \geq Pr(\mathbf{a}|\mathbf{f}, \mathbf{e}; \mathbf{4})$, and set $\mathbf{a}' = \tilde{b}(\mathbf{a})$.

IX. Model 5

Model 5 is introduced to deal with the issue of deficiency. In Model 5, the probability for placing the head word of the i-th one-word cept is

$$Pr(\Pi_{[i]1} = j | \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \triangleq d_{=1}(v_j | \mathcal{B}(f_j), v_{\Theta_{i-1}}, v_m - \phi_{[i]} + 1)(1 - \delta(v_j, v_{j-1})), \tag{26}$$

where v_j is the number of vacancies up to and including position j just before we place $\tau_{[i]k}$ in $\mathbf{f}$. Note that

- $(1 - \delta(v_j, v_{j-1}))$ ensures that position j must be vacant if a head word is to be placed there;
- $v_m - \phi_{[i]} + 1$ is the number of vacancies pre-excluding those to be occupied by the remaining words of the i-th one-word cept;
- $v_{\Theta_{i-1}}$ is the number of vacancies up to and including the center of the previous one-word cept, i.e., position Θ_{i-1};

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 fellow in applying statistical approaches to 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.

[3] K. Jassem, “Semantic classification of adjectives on the basis of their syntactic features in Polish and English,” *Machine Translation*, vol. 17, no. 1, pp. 19–41, 2002.

[4] ——, *WJ05 - a modern version of Polish Club*. ISBN: 83-919009-1-6, 2004.

[5] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, “A statistical approach to machine translation,” *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[6] F. Jelinek, *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. The MIT Press, 1998.