行政院國家科學委員會專題研究計畫　期中進度報告

資訊萃取技術在生物醫學文獻上的應用與探討(1/2)

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-009-074-
執行期間：93年08月01日至94年07月31日
執行單位：國立交通大學資訊科學學系(所)

計畫主持人：梁婷

計畫參與人員：吳典松、施並格、林裕祥、王怡嘉、龔自良、黃立泓、蘇傳堯、施曉茹

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94年5月20日

行政院國家科學委員會補助專題研究計畫 □ 成果報告
ˇ期中進度報告

資訊萃取技術在生物醫學文獻上的應用與探討 (1/2)

計畫類別：ˇ個別型計畫 □ 整合型計畫
計畫編號：NSC 93－2213－E－009 －074 －
執行期間： 93 年 8 月 1 日至 94 年 7 月 31 日

計畫主持人：梁 婷 副教授
共同主持人：
計畫參與人員：吳典松、施並格、林裕祥、王怡嘉、龔自良、黃立泓、蘇傳堯、施曉茹

成果報告類型(依經費核定清單規定繳交)：ˇ精簡報告 □完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學資訊科學學系

中 華 民 國 94 年 5 月 20 日

資訊萃取技術在生物醫學文獻上的應用與探討 (1/2)

在本計畫中我們開發有效實用的自然語言處理技術和文件探勘技術，進而建製一個可應用在生物文獻的自動資訊萃取系統。主要的工作包括生物實體名稱辨識、名稱指代處理、關係的辨識與萃取。我們結合法則式和統計式的方法來強化實體名稱辨識的效能。此外我們利用文件探勘技術來解決語句中指式型指代問題。同時我們也探討生物訊息和非生物訊息在實體關係的辨識和強度計算上的影響力，並利用探勘技術建立關聯法則以處理存在於語句中的實体關係的語言問題。

關鍵詞：自然語言處理、資訊萃取、文件探勘、實體名稱、指代處理、關係辨識

英文摘要

Information Extraction In Biomedical Domain (1/2)

We propose to develop an efficient information extraction system useful for biomedical literature by using natural language processing and textual mining techniques. This system will mainly address the tasks such as named entity identification, anaphora resolution, relation identification and extraction. We will employ both statistical and linguistic models for named entities identification. We will use textual mining to deal with those sortal anaphora problems. Meanwhile, the proposed relation recognition mechanism will take into account both the biomedical information encoded in the existing databases as well as the information directly mined from the literature. Besides the problems associated with the linguistic varieties will be tackled by using the proposed association rules.

Keywords: natural language processing, textual mining, information extraction, named entity identification, anaphora resolution, relation identification.

資訊萃取技術在生物醫學文獻上的應用與探討

一、前言

近年生物醫學研究蓬勃發展，相關文獻快速累積。例如以果蠅資料庫參考文獻而言在近一百年間 (1900-2000)几乎是呈指數型的增長。如此增長的速度對從事研究者而言，要能在浩瀚的資料中全備追蹤掌握相關研究資訊是一項不容易的事情。另一方面，多數的生物資料庫如 Protein Information Resource (PIR), SWISS-PROT, Database of Interacting Proteins (DIP), Molecular INTeraction database (MINT)...等多仰賴生物醫學專家閱讀論文，再將其中重要研究發現和結果，萃取、整理、儲存到結構化資料庫中。然而毫無疑問地，這種人工精心打造的知識庫，其建立、更新與資訊正確性(integrity)的檢查，實在是一件耗時耗力的工作。因此極需資訊萃取工具的開發來協助生物專家，以加速生物知識的萃取和管理。這種資訊萃取系統的建立無疑地將可促進資訊的整合、交流和更新，甚至帶來生物醫學技術的突破。

二、研究目的

本計劃將探討兩個議題分別是萃取技術的研發和問答系統的製作將分兩年來進行。在本年度我們將開發有效實用的自然語言處理技術和文件探勘技術，進而建製一個可應用在生物文獻的自動資訊萃取系統。主要的工作將包括生物實體名稱和關係的辨識與擷取。我們相信此計畫的執行不僅有助於生物學家的知識擷取和整理，進而促進生物研究的新發現，同時亦有益於實用的資訊萃取技術的發展，以應用於其它領域的知識庫建構的自動化。

三、文獻探討

近年資訊萃取主要的議題分別在生物實體名稱和實體之間的關係辨識與擷取。在實體名稱的辨識上如同新聞語料中所面臨的挑戰包括詞界、新詞、命名的不規則與不一致性、語義的多樣性、省略詞彙、縮寫、指代現象處理等問題。由於名稱的組成往往包含了兩個以上的詞， 是以詞界辨識的問題在名稱分類前需先予以解決。

目前名稱的辨識有專注於單類實體如蛋白質名稱到多種類實體的辨識。使用的技術可分為兩種。第一種是利用實體名稱的組成成分以人工歸納的法則作為辨識基礎。 一般而言這種方法相較於統計法可以達到較高的正確率，然而手建的規則需要專家知識的輔助，故缺乏擴充性(scalability)和可移植性(portability)。目前這方面系統有蛋白質名稱的辨識工具 KeX [Fukuda et al., ‘98] 和Yapex [Olsson et al., ‘02]。在[Hou and Chen, ‘03]的文章中 Hou and Chen 則交叉利用這兩個系統結果和篩選法則來提高蛋白質名稱的辨識率。

第二種方法是應用統計模組進行辨識，如 Hidden Markov Model [Collier et al., ’00; Shen et al., ‘03], Maximum Estimation [Nobata et al., ’99; Kazama et al., ’01; Chieu and Ng, ‘03], Support Vector Machine [Kazama et al., ’02; Takeuchi and Collier, ’03; Yamanoto et al., ‘03], Naïve Bayes [Tsuruoka and Tsujii, ‘03]等等.。然而此種機器學習為主的辨識需要大量的標記好的語料以達到可接受的成效。依據 IdentiFinder System 的結果分析顯示新聞語料中其名稱辨識結果與訓練語料量成對數(log)的增長。因此對機器學習的方法首要的挑戰之一包括如何簡易地產生足夠量的訓練語料。目前廣為所用的已標記語料有 Bio1，它包含有 100 篇標記好的 Medline 摘要以及所用的 taxonomy 是由 Tateishi et al. 於 2000 年所建的和 GENIA project 的 GENIA corpus。

在生物實體名稱邊界的判定上有以辭典作比對或使用 BIO(Beginning/Inside/Outside of a named entity) (或其變異體如 BIO1, BIO2, IOE1, IOE2)的表示方法，將名稱邊界辨識問題轉換成分類的問題。此外多數的統計式辨識方法的成效(SVM 方法較與特徵無關)亦有賴於特徵的挑選[Kazama et al., ‘02]。使用的特徵包括 *Part-Of-Speech, Surface, Cue Word, Morphological, Contextual features*。目前的結果在實體名稱的分類中，對單一和多種實體名稱分類上，以 GENIA 3.0 Corpus 所做的實驗而言，大約可分別達到 70%和 66%左右的 F-Score [Shen et al., ’03; Tsuruoka and Tsujii et al., ‘03]。這樣的結果相較於一般語料(如新聞語料中)的實體名稱辨識率可達 90%以上的 F-score，生物醫學的實體名稱辨識技術仍有努力的空間。

至於生物實體之間關係的辨識與抽取的挑戰性在於句型語意表示方法的多樣性和關係存在的複雜度，諸如肯定關係、否定關係、未定關係、隱藏關係、歧異關係的確認。再者生物文件中，如 Genia Corpus，單句所含的生物實體平均數有 5.28 個。因此實體之間的單一或多重關係的處理需要進一步的辨識。

在關係辨識部分，目前多著重在兩個生物實體間關係的存在與否。因此可利用統計式的方法，來進行大量的辨識處理。例如 Carven 與 Kumlien [‘99] 使用 Navie Bayes classification 技術將關係辨識轉為單一句子的分類工作以進行蛋白質的子細胞位置及其子細胞結構之自動辨識。Stephens et al. [‘01] 利用傳統的詞頻加權技術，來計算基因之間關係的強度，做為判讀關係的存在與否。此外在[Ding et al. ‘02]的文章中也討論到關係的存在與兩實體在文章中距離有關，距離越近，關係辨識的正確率越高，但召回率越低。一般而言，統計式

方法因缺少語法的分析，所以無法界定實體在關係之間的角色，同時統計式方法的辨識成效也有賴於語料庫收集的完備與否。

目前多數關係萃取取系統多倚賴語法剖析器的協助和主要動詞為主的語法模型比對來進行。如早期 Blaschke et al. [‘99] 使用些許手動建立的規則，半自動的辨識兩蛋白質間的交互作用。Sekimizu et al. [‘98] 利用在 Medline 摘要中的常見動詞，辨識基因與基因產物間的交互作用。Proux et al. [‘00] 以 finite-state machine 為基礎的語言工具和知識概念圖，自動地從 Flybase 的 1200 個句子中，萃取出基因的交互作用。Yakushiji et al. [‘00] 使用 full parser，以輔助生物事件萃取。Ono et al. [‘01] 使用蛋白質詞彙、提示詞、以及簡單的詞性標記，萃取出兩蛋白質間的交互作用。雖然這個方法得到高於 80%的召回率、精確率，但交互作用只限於少許的關鍵詞。

至於跨語句的關係萃取則有賴於指代現象的處理機制。在[Castano et al. ‘02] 考量名詞片語的相似性，語意類別，語法角色做為先行詞的挑選依據。Hahn et al. [‘02] 則定義所謂的 ”Center Lists”，將每一個名詞依據一些特徵往前找尋最相關名詞。[Gaizauskas et al. ‘03] 使用事先定義的語意法則，藉以串聯語句之間的生物實體，處理同指現象。上述大部分的方法都只在少量的文章作實驗且多侷限於簡單關係的萃取。對於單、多句子隱含的多重關係，事實上，仍需進一步的探勘處理技術，方能有效的解析出來。

二十一世紀可以說是資訊網路與生物科技產業的世紀，其中生物科技又被譽為希望工程，許多學術研究機構莫不積極發展。近幾年中，我們也在本校的跨系所的重點計劃推動下，應用資訊擷取技術建立一個整合型的網路微生物文獻自動化處理和查詢系統[Liang, et al. 03]。此外我們也初步探討資訊萃取在實體名稱和關係辨識上的應用，包括利用機器學習方法製作一個無需辭典協助的生物實體名稱辨識與分類工具 Bio-tagger Version 1[Chen, ‘03]。此標記工具可處理省略詞彙還原，並可進行多種生物實體名稱辨識(Protein、DNA、RNA、 Source 和其他生物體名稱)，在 GENIA 3.01 Corpus 的實驗上整體名稱辨識和分類的 F-Score 分別達到 69%和 60%，與一般 knowledge-poor 的方法相當。此外我們也提出一個權重式 Navie Bayes 分類模型對單句中肯定、否定、未定關係進行辨識，再以 pattern rules 作多重關係的萃取 [You, 2003]。然而如同多數前述所提到的資訊萃取系統，在這些初步的辨識和萃取方法設計上，我們在使用生物方面的領域知識的研究上仍有待努力。因此在本計畫中我們將加強這方面的能力，希望結合文件探勘技術探勘出存在於資料庫中的生物訊息，以強化並驗證我們所提的資訊萃取系統和生物知識問答系統。

四、研究方法

在本年度我們開發有效實用的自然語言處理技術和文件探勘技術，進而建製一個可應用在生物文獻的自動資訊萃取系統。主要的工作將包括生物實體名稱和關係的辨識與擷取及實體名稱指代處理。

有別於多人使用的 GENIA 3.01 語料，其所涵蓋的範圍較廣泛，在本計畫中，我們以 SWISS-Prot 資料庫的參考文獻收集成為訓練語料。由於這是蛋白質相關資訊的語料，對於蛋白質名稱的辨識和關係的資訊探勘上將有較丰富的資訊以達到較高正確的辨識率。另外我們也收集、整合各生物實體名稱，及蛋白質之間的關係，做為日後的實体與關係的比對、檢驗用。文章前置處理包括斷句處理和斷詞處理。我們用 Sentence Splitter 來處理斷句部分。斷詞處理是使用 Penn Treebank tokenization。我們修改這兩個前置處理工具以適合所使用的語料。

實體名稱辨識處理

有別多數實體名稱辨識系統中几乎都沒有處理到詞彙省略和變異現象，我們發現這個問題在建構一個實際有效的名稱辨識器上仍有其探討

的必要性。以 GENIA 3.01 Corpus 而言，人工標記的省略詞就有 1,595 個，略可分成四種不同的型式。因此在這本計畫中我們改良先前建構的省略回復處理器，利用詞群技術以導引自動機的途徑以提升其正確率。就文獻探討，我們是第一位有提出解決 coordination variants 現象的研究。我們改良我們之前所設計的方法加入了詞群的機制以導引辨識器途徑，F-score 可提升 12%。此外針對常見的縮寫現象，我們利用法則並以 Park and Byrd [‘01] 所提的策略來決定 window size，以辨識縮寫和原型，F-score 可達 93%。

另一方面從前述的文獻研究中，我們也發現法則式或統計式的辨識處理都各有其優缺點。因此在本計劃中我們以混合式的策略來建構實體名詞辨識器。其中規則式的部分我們使用先前所開發適用於生物醫學文獻的 HMM-based POS tagger (目前此 tagger 在 GENIA 3.02p 對 65 種標記可達到 94%的正確率) 來做詞性標記，並利用文件探勘技術探勘出文獻庫中實体名稱的組成詞段模型，以產生候選實體名稱。在統計式的部分，我們改良之前所設計的 HMM-based 的辨識器，加入從 SWISS-Prot corpus 中所探勘出耒的蛋白質各生物屬性之重要資訊，以得到候選實體名稱。我們將利用收集到的實体名稱建立所需的標記訓練語料耒訓練所建的辨識器。我們再合併從法則式和統計式辨識中所得到的候選實體名稱，並藉由篩選機制進行最後的檢驗。在沒有 dictionary 輔助下，我們的 F-score 可達 76% 與現今有 dictionary 輔助的結果接近。

實體名稱指代處理

文句中的指代現象處理是自動辨識實體關係的重要一環。因此除了加以修改先前所設計的虛詞和代名詞的指代處理模組 [Liang and Wu, ‘03]，使之適用於生物醫學語料處理。在本計畫當中，我們還將加強指示型指代處理模組，以建構一個概念式的指代還原器，有效處理跨語句的關係萃取。此指示型指代處理模組將利用指代詞和先行詞之間的語意標記訊息和詞彙共現等特徵的一致性作挑選依據。同時我們也進一步藉助探勘技術來萃取出指代詞的概念功能詞 (如酵素, receptor 之於蛋白質)，以有效找出先行詞和指代詞的配對，進而達到 UMLS 中生物概念的分類自動擴充。這方面困難度在於有此標記的訓練語料很少。目前我們借助專家手工標記的 Medtract Corpus 整理出的先行詞和指代詞的配對及從 PubMed 所得的 patterns 建立關連法則。所提的方法在指代詞指代消解，F-score 可達 92%；在名詞指代消解，F-score 可達 78%。

實體關係辨識與萃取

如前所述大多數的實體關係辨識到目前為止都僅限於少量的語句測驗並且很少用到生物實体屬性訊息。因此在本計劃中我們嚐試從已知的部份實體關係 (以 DIP 資料庫中有關果蠅的部份)，利用 SWISS-Prot 所存的生物語意資訊(如是否出自同一個生物、具有相近的功能表現、‥等)，以強化我們所提的辨識機置。我們以蛋白質的交互關係為主要處理項目，再擴展到其它關連的辨識。利用部份具關係的蛋白質配對做為搜尋詞組進行對所整理出的 SWISS-Prot 語料做文件探勘，以取出可用的關係辨識訊息並藉此產生可用的標記訓練語料。在關係確認的處理上，除了利用生物資料庫如 SWISS-Prot database 所探勘到的生物語意資訊，我們也考量耒自文獻中所探勘到的非生物訊息(如距離、頻率、共現率‥等)和語言的訊息 (如否定訊息、語法角色、與動詞關係、相鄰詞‥等)。對於同義詞組我們將事先予以群集以增強關係的辨識程度。我們分別以統計式的模組和權重計算耒探討不同的特徵在已知的關係辨識上的影響力，最後再另拿部份的關連做測試以檢驗我們的辨識機置。

計劃成果自評

本計畫中，我們開發有效實用的自然語言處理技術和文件探勘技術，進而建製一個可應用在生物文獻的自動資訊萃取系統。目前完成的工作項目及成果如下：

1．相關語料和相關資料庫整合

2．訓練語料標記程序建構：我們建立一個新的標記語料庫SRC較之一般所用的 GENIA Corpus（東京大學所建）更適合作為 Protein 文獻探勘技術的 evaluation corpus.

3．適用於生物文獻的詞類標記器建構:正確率可達 94%.

4．省略回復處理器建構：就文獻探討，我們是第一位有提出解決 coordination variants 現象的研究。我們改良我們之前所設計的方法加

入了詞群的機制以導引辨識器途徑，F-score 可提升 12%

5．混合式實體名稱辨識器建構：在沒有 dictionary 輔助下，我們的 F-score 可達 76% 與現今有 dictionary 輔助的結果接近。

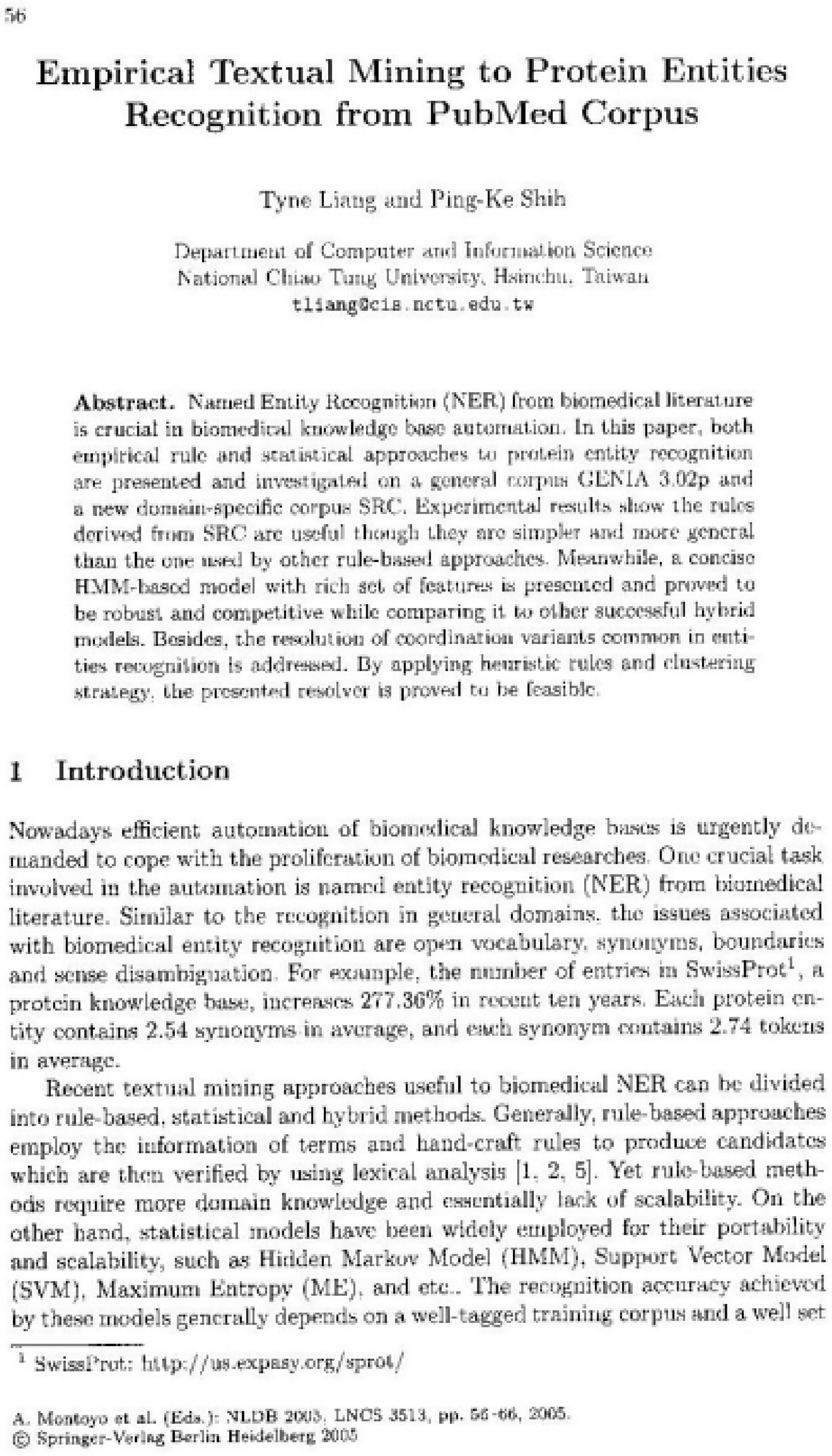
6．實體名稱指代處理器建構：我們處理虛詞、代名詞和指示型指代消解及縮寫。所提的方法較之以往的方法在相同的測試語料下可達較高的 F-score.

7．實體關係探勘器和辨識器建構：生物語意特徵和生物語意特徵探勘。尚未完成的部份為關係處理上關係強度權重計算分析，預計七月底前應可完成。

目前完成計劃的成果包括完成三篇相關碩士論文（參考文獻 1,2,3)，和三篇會議論文(4,5,6)。其中一篇如附件將於六月在西班牙第 10 屆 International Conference on Application of Natural Languages to Database Systems 發表。另兩篇亦在 16th 計算語言學會議發表。

參考文獻

1. Ping-ke ShiH, 2004, “Automatic Protein Entities Recognition from PubMed Corpus”, Master Thesis, National Chiao Tung University.
2. Yu-Hsiang Lin, 2004, “Coreference Resolution in Biomedical Literature,” Master Thesis, National Chiao Tung University.
3. Yi-Chia Wang, 2004,“Web-based Unsupervised Learning to Query Formulation for Question Answering,” Master Thesis, National Chiao Tung University.
4. Tyne Liang and Ping-ke Shih, 2005, “Empirical Textual Mining to Protein Entities Recognition From PubMed Corpus,” NLDB 2005, Lecture Notes in Computer Science, 3513, pp. 56-66.
5. Yu-Hsiang Lin and Tyne Liang, 2004, *Pronominal and Sortal Anaphora Resolution for Biomedical Literature*, Proceedings of ROCLING XVI, Taipei, Taiwan, pp. 101-110.
6. Yi-Chia Wang, Jian-Cheng Wu, Tyne Liang, and Jason S. Chang, 2004, *Using the Web as Corpus for Unsupervised Learning in Question Answering*, Proceedings of ROCLING XVI, Taipei, Taiwan, pp. 191-198.



Empirical Textual Mining to Protein Entities Recognition from PubMed Corpus

Tyne Liang and Ping-Ke Shih

Department of Computer and Information Science
National Chiao Tung University, Hsinchu, Taiwan
tliang@cis.nctu.edu.tw

Abstract. Named Entity Recognition (NER) from biomedical literature is crucial in biomedical knowledge base automation. In this paper, both empirical rule and statistical approaches to protein entity recognition are presented and investigated on a general corpus GENIA 3.02p and a new domain-specific corpus SRC. Experimental results show the rules derived from SRC are useful though they are simpler and more general than the one used by other rule-based approaches. Meanwhile, a concise HMM-basod model with rich set of features is presented and proved to be robust and competitive while comparing it to other successful hybrid models. Besides, the resolution of coordination variants common in entities recognition is addressed. By applying heuristic rules and clustering strategy, the presented resolver is proved to be feasible.

1 Introduction

Nowadays efficient automation of biomedical knowledge bases is urgently demanded to cope with the proliferation of biomedical researches. One crucial task involved in the automation is named entity recognition (NER) from biomedical literature. Similar to the recognition in general domains, the issues associated with biomedical entity recognition are open vocabulary, synonyms, boundaries and sense disambiguation. For example, the number of entries in SwissProt[1], a protein knowledge base, increases 277.36% in recent ten years. Each protein entity contains 2.54 synonyms in average, and each synonym contains 2.74 tokens in average.

Recent textual mining approaches useful to biomedical NER can be divided into rule-based, statistical and hybrid methods. Generally, rule-based approaches employ the information of terms and hand-craft rules to produce candidates which are then verified by using lexical analysis [1, 2, 5]. Yet rule-based methods require more domain knowledge and essentially lack of scalability. On the other hand, statistical models have been widely employed for their portability and scalability, such as Hidden Markov Model (HMM), Support Vector Model (SVM), Maximum Entropy (ME), and etc.. The recognition accuracy achieved by these models generally depends on a well-tagged training corpus and a well set

[1] SwissProt: http://us.expasy.org/sprot/

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 56–66, 2005.

of features [3, 6, 7, 9, 10]. Recently, hybrid approaches are proposed by combining coded rules, statistical model and dictionaries [4, 9]. As pointed in [10], it is expected that systems on a specified evaluation corpus with help of dictionaries tend to perform better than the general ones without help of any dictionaries. For example, the recognition performance is significantly improved when dictionary and rules are applied at post-processing together with a ME-based recognition mechanism in [4].

In this paper, recognition for protein entities from PubMed[2] corpus is addressed so as to facilitate the automation of protein interaction databases construction. In order to mine more features relevant to protein entities, we assembled a domain-specific protein corpus SRC (SwissProt Reference Corpus) which were extracted from SwissProt reference articles and we tagged it by simply matching SwissProt entry collection. Experimental results show that this new domain corpus is indeed helpful in generating informative patterns used in both rule-based and statistical models. It is also found that though the derived rules are fewer and less complicated than the ones used in the rule-based systems Kex [1] or Yapex [5], the presented model outperforms these two systems in terms of higher F-scores on a general corpus like GENIA 3.02p [3] and the domain-specific SRC.

On the other hand, a concise HMM-based model is presented with a back-off strategy to overcome data sparseness. With a rich set of features, the presented approaches could achieve promising results, by showing 76-77% F-scores on both GENIA corpus and SRC. Compared to the results achieved by some successful systems (the best 78% F-score for protein instances in [9]) which employ dictionaries or semantic lexicon lists, our results are competitive for three reasons. First, the recognition is done without any help of dictionaries or predefined lexicon lists. Second, the presented concise HMM is easily implemented and robust for different corpora. Third, our results are evaluated with strict annotation and enetities with the longest annotation are adopted in case they are in the nested forms.

Besides, this paper addresses the issue of coordination variants while we tackle with NER problems in written texts. To resolve such term variants, a method based on heuristic rules and clustering strategy is presented. Experimental results on GENIA corpus 3.0 proved its feasibility by achieving 88.51% recall and 57.04% precision on a test of 1850 sentences, including 174 variants.

2 Corpus Preparation

In order to boost protein entities recognition by mining more relevant information, we assembled a domain-specific corpus 'SwissProt Ref Corpus' ('SRC' for short), other than the widely-used tagged corpus like GENIA 3.02p. The new corpus was processed by employing Sentence Splitter[4] and Penn Treebank

[2] PubMed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Pubmed

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[4] Sentence Splitter: http://l2r.cs.uiuc.edu/~cogcomp/

Tokenizer[5] for sentence segmentation and tokenization respectively. The POS-tagging is processed by a HMM-based POS tagger which was developed in our lab. By using GENIA 3.02p as training set, our POS-tagger could yield 95% F-score. For the sake of saving human efforts, annotating SRC with all the target entities was simply implemented with the following steps:

1. Tokens are split by space and hyphen.
2. Each token is converted to lower case except its initial character.
3. Entity is recognized if it matches an entity from SwissProt version 42.0.

The final specific SRC corpus is composed of 2,894 abstracts, which were particularly selected from SWISSPORT 82,740 reference articles in such a way that each of them contains at least six target entities. Table 1 lists the basic statistics for SRC and GENIA 3.02p.

Table 1. The statistics of SRC corpus and GENIA corpus 3.02p.

	SRC		GENIA	
	count	average	count	average
Abstract (a)	2,894		1,999	
Sentence (s)	28,154	9.73 (s/a)	18,572	9.29 (s/a)
Token (t)	740,001	255.70 (s/a) 26.28 (t/s)	490,469	245.36 (t/a) 26.41 (t/s)
Protein (p) Entity	31,977	11.05 (p/a) 1.14 (p/s)	32,525	11.05 (p/a) 1.14 (p/a)
Entity Token (t)	57,878	1.81(t/p)	58,200	1.79 (t/p)

3 Coordination Variants Resolution

Coordination variants are one common type of variants in general written texts like MEDLINE records. For example there are 1598 coordination variants in GENIA 3.02p corpus and each variant contains 2.1 entities in average. Table 2lists three types of the regular expressions generalized from the GENIA 3.02p training corpus of 16,684 sentences (in which 1421 coordination variants are distributed in 1329 sentences). There #, H, T, and R indicate core, head, tail, and coordinate terms respectively. For example, in the coordination '91 and 84 kDa proteins', '91' and '84' are the core terms, 'kDa proteins' is the tail term, and 'and' is the coordinate term.

The variant resolution was implemented with finite state machines (FSM) which are verified by a test set of 1850 sentences in which 174 variants are distributed in 165 sentences. Experimental results showed that this approach yielded 91.38% recall and 42.06% precision (indicated as baseline approachïn Table 3). In practice, the precision can be improved by presenting more number of FSMs so as to cover all possible variant patterns, yet it will slow down the resolving throughput. In order to increase the sensitivity of coordination identification, a simple term clustering is employed. Suppose terms t_i, t_j co-occur

[5] http://www.cis.upenn.edu/˜treebank/tokenization.html

Table 2. Original patterns, expanded patterns, and examples.

	Regular Expression		Example
Type 1	Original	$H\#(R\#)^+$	human chromosomes 11p15 and 11p13
	Expanded	$(H\#R)^+H\#$	humman chromosomes 11p15 and human chromosome 11p13
Type 2	Original	$\#(R\#)^+T$	c-fos, c-jun, and EGR2 mRNA
	Expanded	$\#T(R\#)^+T$	c-fos mRNA, c-jun mRNA, and EGR2 mRNA
Type 3	Original	$H\#(R\#)^+T$	human T and B lymphocytes
	Expanded	$\#T(R\#)^+T$	human T lymphocytes and human B lymphocytes

in one coordination variant, and terms t_i, t_k co-occur in another one. Then we put t_i, t_j and t_k into one cluster. The clustering procedure was implemented recursively. With such term clustering strategy (indicated as 'unlimited-distance' in Table 3), the resolution precision is increased by 4%. This showed that the clustering approach is helpful to restrict the path movement in FSMs. To distinguish the closeness of the terms in the same cluster, we furthermore applied the Floyd-Warshall algorithm to cluster sets. That is, if terms t_i, t_j co-occur in a sentence and terms t_i, t_k co-occur in another one but t_j, t_k do not co-occur in any sentence, then the $dist(t_j, t_k) = 2$. With this clustering strategy, the precision became 57.04% (increasing 15% with respect to the baseline method) at the expense of lower recall.

Table 3. Accuracy of coordination variants identification in GENIA 3.02p.

	dist.	Variants	tp+fp	tp	Recall	Precision	F-Score
Baseline	N/A	174	378	159	91.38%	42.06%	57.61%
Term Clustering	unlimited	174	338	158	90.80%	46.75%	61.72%
	1	174	270	154	88.51%	57.04%	69.37%

4 Protein Entity Recognition

In this paper, protein entity recognition is approached and investigated by both rule-based and HMM models. The performance verification is implemented by using both SRC and GENIA 3.02p corpora in such a way that the corpora are divided into 90% for training phase and 10% for testing phase.

4.1 Rule-Based Approach

The rule-based recognition is implemented by employing the patterns of the protein nomenclature mined from SRC and GENIA corpora. The patterns are formed in terms of core, function or predefined terms. Core terms show the closest resemblance to regular proper names. Function terms describe the functions or characteristics of a protein. Table 4 shows the frequent regular expressions which 'C' indicates core term, 'F' indicates function term, and 'P' indicates predefined term, namely specifier, amino acid and unit.

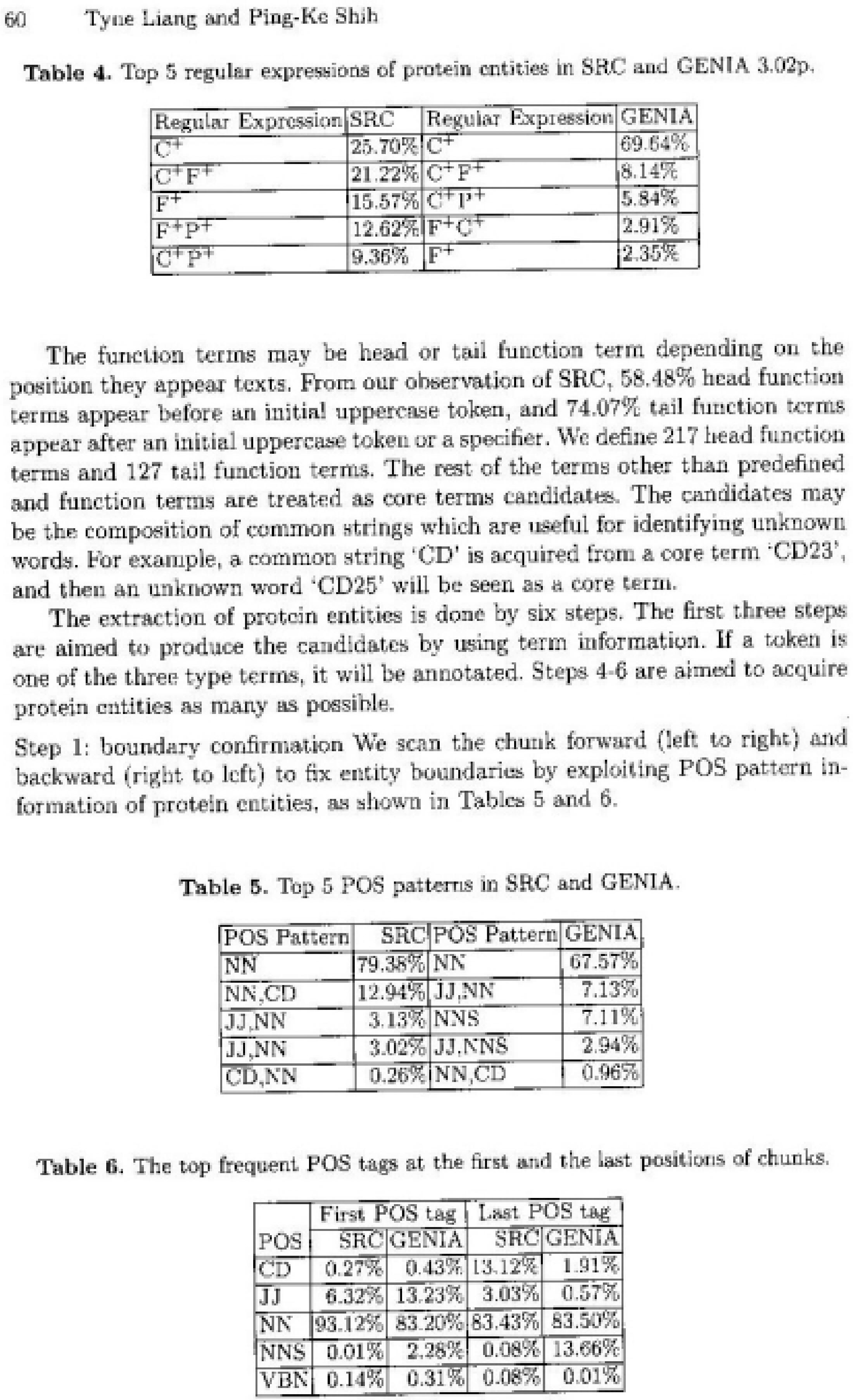


Table 4. Top 5 regular expressions of protein entities in SRC and GENIA 3.02p.

Regular Expression	SRC	Regular Expression	GENIA
C^+	25.70%	C^+	69.64%
C^+F^+	21.22%	C^+F^+	8.14%
F^+	15.57%	C^+P^+	5.84%
F^+P^+	12.62%	F^+C^+	2.91%
C^+P^+	9.36%	F^+	2.35%

The function terms may be head or tail function term depending on the position they appear texts. From our observation of SRC, 58.48% head function terms appear before an initial uppercase token, and 74.07% tail function terms appear after an initial uppercase token or a specifier. We define 217 head function terms and 127 tail function terms. The rest of the terms other than predefined and function terms are treated as core terms candidates. The candidates may be the composition of common strings which are useful for identifying unknown words. For example, a common string 'CD' is acquired from a core term 'CD23', and then an unknown word 'CD25' will be seen as a core term.

The extraction of protein entities is done by six steps. The first three steps are aimed to produce the candidates by using term information. If a token is one of the three type terms, it will be annotated. Steps 4-6 are aimed to acquire protein entities as many as possible.

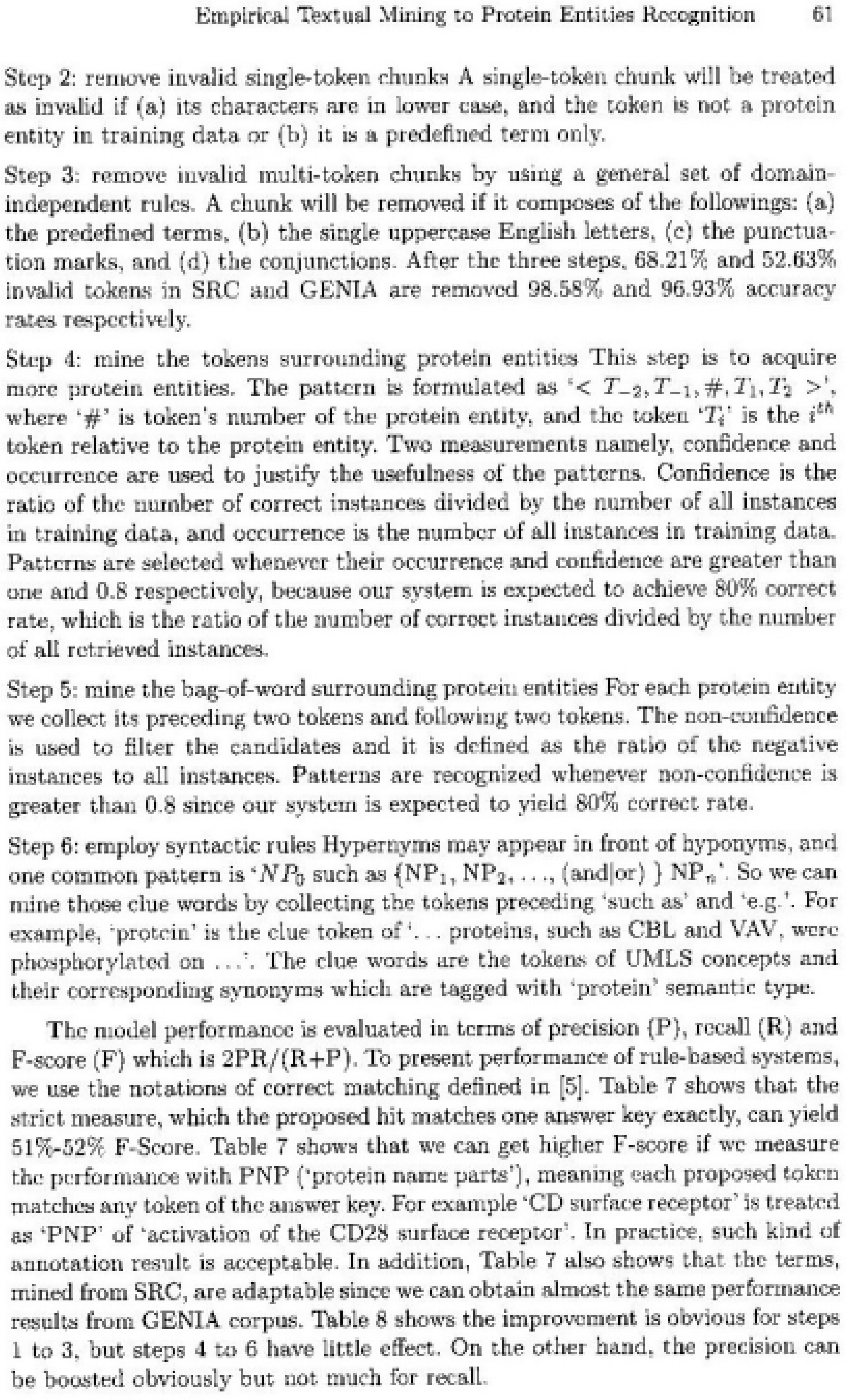
Step 1: boundary confirmation We scan the chunk forward (left to right) and backward (right to left) to fix entity boundaries by exploiting POS pattern information of protein entities, as shown in Tables 5 and 6.

Table 5. Top 5 POS patterns in SRC and GENIA.

POS Pattern	SRC	POS Pattern	GENIA
NN	79.38%	NN	67.57%
NN,CD	12.94%	JJ,NN	7.13%
JJ,NN	3.13%	NNS	7.11%
JJ,NN	3.02%	JJ,NNS	2.94%
CD,NN	0.26%	NN,CD	0.96%

Table 6. The top frequent POS tags at the first and the last positions of chunks.

	First POS tag		Last POS tag	
POS	SRC	GENIA	SRC	GENIA
CD	0.27%	0.43%	13.12%	1.91%
JJ	6.32%	13.23%	3.03%	0.57%
NN	93.12%	83.20%	83.43%	83.50%
NNS	0.01%	2.28%	0.08%	13.66%
VBN	0.14%	0.31%	0.08%	0.01%



Step 2: remove invalid single-token chunks A single-token chunk will be treated as invalid if (a) its characters are in lower case, and the token is not a protein entity in training data or (b) it is a predefined term only.

Step 3: remove invalid multi-token chunks by using a general set of domain-independent rules. A chunk will be removed if it composes of the followings: (a) the predefined terms, (b) the single uppercase English letters, (c) the punctuation marks, and (d) the conjunctions. After the three steps, 68.21% and 52.63% invalid tokens in SRC and GENIA are removed 98.58% and 96.93% accuracy rates respectively.

Step 4: mine the tokens surrounding protein entities This step is to acquire more protein entities. The pattern is formulated as '$< T_{-2}, T_{-1}, \#, T_1, T_2 >$', where '#' is token's number of the protein entity, and the token 'T_i' is the i^{th} token relative to the protein entity. Two measurements namely, confidence and occurrence are used to justify the usefulness of the patterns. Confidence is the ratio of the number of correct instances divided by the number of all instances in training data, and occurrence is the number of all instances in training data. Patterns are selected whenever their occurrence and confidence are greater than one and 0.8 respectively, because our system is expected to achieve 80% correct rate, which is the ratio of the number of correct instances divided by the number of all retrieved instances.

Step 5: mine the bag-of-word surrounding protein entities For each protein entity we collect its preceding two tokens and following two tokens. The non-confidence is used to filter the candidates and it is defined as the ratio of the negative instances to all instances. Patterns are recognized whenever non-confidence is greater than 0.8 since our system is expected to yield 80% correct rate.

Step 6: employ syntactic rules Hypernyms may appear in front of hyponyms, and one common pattern is 'NP_0 such as {NP$_1$, NP$_2$, ..., (and|or) } NP$_n$'. So we can mine those clue words by collecting the tokens preceding 'such as' and 'e.g.'. For example, 'protein' is the clue token of '... proteins, such as CBL and VAV, were phosphorylated on ...'. The clue words are the tokens of UMLS concepts and their corresponding synonyms which are tagged with 'protein' semantic type.

The model performance is evaluated in terms of precision (P), recall (R) and F-score (F) which is 2PR/(R+P). To present performance of rule-based systems, we use the notations of correct matching defined in [5]. Table 7 shows that the strict measure, which the proposed hit matches one answer key exactly, can yield 51%-52% F-Score. Table 7 shows that we can get higher F-score if we measure the performance with PNP ('protein name parts'), meaning each proposed token matches any token of the answer key. For example 'CD surface receptor' is treated as 'PNP' of 'activation of the CD28 surface receptor'. In practice, such kind of annotation result is acceptable. In addition, Table 7 also shows that the terms, mined from SRC, are adaptable since we can obtain almost the same performance results from GENIA corpus. Table 8 shows the improvement is obvious for steps 1 to 3, but steps 4 to 6 have little effect. On the other hand, the precision can be boosted obviously but not much for recall.

Table 7. Experimental results by rule-based approach.

	Notation	tp+sn	tp+fp	tp	recall	precision	F-Score
SRC	SLOPPY	3234	4782	2987	92.36%	62.46%	74.53%
	PNP	3234	4782	2859	88.40%	59.79%	71.33%
	STRICT	3234	4782	2077	64.22%	43.43%	51.82%
	LEFT	3234	4782	2620	81.01%	54.79%	65.37%
	RIGHT	3234	4782	2363	73.07%	49.41%	58.96%
	LorR	3234	4782	2907	89.89%	60.79%	72.53%
GENIA	Notation	tp+sn	tp+fp	tp	recall	precision	F-Score
	SLOPPY	3451	4923	3010	87.22%	61.14%	71.89%
	PNP	3451	4923	2837	82.21%	57.63%	67.76%
	STRICT	3451	4923	2123	61.52%	43.12%	50.70%
	LEFT	3451	4923	2765	80.12%	56.16%	66.04%
	RIGHT	3451	4923	2296	66.53%	46.64%	54.84%
	LorR	3451	4923	2938	85.13%	59.68%	70.17%

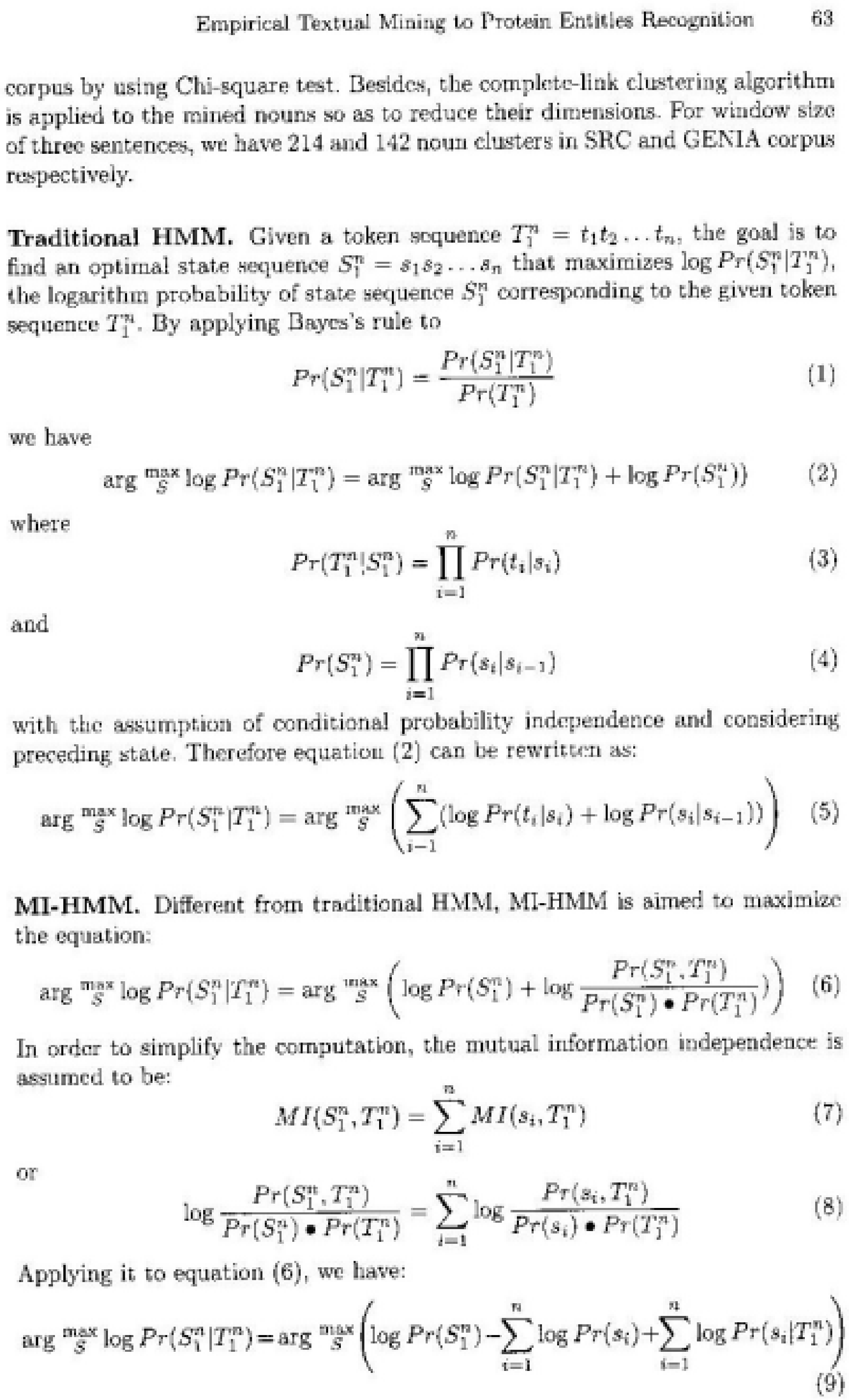
Table 8. The intermediate results of rule-based approach.

	Procedure	tp+sn	tp+fp	tp	recall	precision	F-Score
SRC	step1	3234	10480	2051	63.42%	19.57%	29.91%
	step1-2	3234	5493	2043	63.17%	37.19%	46.82%
	step1-3	3234	4911	2040	63.08%	41.54%	50.09%
	step1-4	3234	4977	2104	65.06%	42.27%	51.25%
	step1-5	3234	4781	2077	64.22%	43.33%	51.83%
	step1-6	3234	4782	2077	64.22%	43.43%	51.82%
GENIA	Procedure	tp+sn	tp+fp	tp	recall	precision	F-Score
	step1	3451	7911	2160	62.59%	27.30%	38.02%
	step1-2	3451	5173	2129	61.69%	41.16%	49.37%
	step1-3	3451	5082	2127	61.63%	41.85%	49.85%
	step1-4	3451	5164	2155	62.45%	41.73%	50.03%
	step1-5	3451	4915	2120	61.43%	43.13%	50.68%
	step1-6	3451	4923	2123	51.52%	43.12%	50.70%

4.2 HMM-Based Approaches

The statistical approach for NER is implemented by a concise HMM model (Concise-HMM) which employs a rich set of input features. Its performance is verified with SRC and GENIA 3.02p by comparing two other models, namely, traditional model (Traditional-HMM) and mutual information model (MI-HMM) which was presented in [9] and produced high F-scores in MUC-6 and MUC-7. The comparison is made in the same environment settings.

In this paper, all the models are trained with the same set of useful features including internal, external and global features. Internal features are those surface clues in tokens (e.g. initial character is upper case). There are 17 internal features mined from the training corpus. External features indicate the external information associated with tokens. We treated POS tags as our external features. Global features are the trigger nouns extracted from whole training



corpus by using Chi-square test. Besides, the complete-link clustering algorithm is applied to the mined nouns so as to reduce their dimensions. For window size of three sentences, we have 214 and 142 noun clusters in SRC and GENIA corpus respectively.

Traditional HMM. Given a token sequence $T_1^n = t_1 t_2 \dots t_n$, the goal is to find an optimal state sequence $S_1^n = s_1 s_2 \dots s_n$ that maximizes $\log Pr(S_1^n|T_1^n)$, the logarithm probability of state sequence S_1^n corresponding to the given token sequence T_1^n. By applying Bayes's rule to

$$Pr(S_1^n|T_1^n) = \frac{Pr(S_1^n|T_1^n)}{Pr(T_1^n)} \qquad (1)$$

we have

$$\arg {}^{\max}_{S} \log Pr(S_1^n|T_1^n) = \arg {}^{\max}_{S} \log Pr(S_1^n|T_1^n) + \log Pr(S_1^n)) \qquad (2)$$

where

$$Pr(T_1^n|S_1^n) = \prod_{i=1}^{n} Pr(t_i|s_i) \qquad (3)$$

and

$$Pr(S_1^n) = \prod_{i=1}^{n} Pr(s_i|s_{i-1}) \qquad (4)$$

with the assumption of conditional probability independence and considering preceding state. Therefore equation (2) can be rewritten as:

$$\arg {}^{\max}_{S} \log Pr(S_1^n|T_1^n) = \arg {}^{\max}_{S} \left(\sum_{i=1}^{n} (\log Pr(t_i|s_i) + \log Pr(s_i|s_{i-1})) \right) \qquad (5)$$

MI-HMM. Different from traditional HMM, MI-HMM is aimed to maximize the equation:

$$\arg {}^{\max}_{S} \log Pr(S_1^n|T_1^n) = \arg {}^{\max}_{S} \left(\log Pr(S_1^n) + \log \frac{Pr(S_1^n, T_1^n)}{Pr(S_1^n) \bullet Pr(T_1^n)}) \right) \qquad (6)$$

In order to simplify the computation, the mutual information independence is assumed to be:

$$MI(S_1^n, T_1^n) = \sum_{i=1}^{n} MI(s_i, T_1^n) \qquad (7)$$

or

$$\log \frac{Pr(S_1^n, T_1^n)}{Pr(S_1^n) \bullet Pr(T_1^n)} = \sum_{i=1}^{n} \log \frac{Pr(s_i, T_1^n)}{Pr(s_i) \bullet Pr(T_1^n)} \qquad (8)$$

Applying it to equation (6), we have:

$$\arg {}^{\max}_{S} \log Pr(S_1^n|T_1^n) = \arg {}^{\max}_{S} \left(\log Pr(S_1^n) - \sum_{i=1}^{n} \log Pr(s_i) + \sum_{i=1}^{n} \log Pr(s_i|T_1^n) \right) \qquad (9)$$

Concise HMM. The presented concise HMM is based on the idea of maximizing the fundamental $\log Pr(S_1^n|T_1^n)$. In the equation (9), $\log Pr(S_1^n|T_1^n)$ and $\sum_{i=1}^n \log Pr(s_i)$ are found to carry less meaning because the weak probabilities of states and state transitions are merely 3-by-3 and 3-by-1 matrices respectively. Thus, a concise HMM can be obtained by simplifying the formula (9) to be equation (10):

$$\arg {}^{\max}_{S} \log Pr(S_1^n|T_1^n) = \arg {}^{\max}_{S} \log Pr(S_1^n) - \sum_{i=1}^{n} \log Pr(s_i|T_1^n) \qquad (10)$$

Since the concise HMM does not take its state transition into account, we put previous state in the model to ensure correct state induction. Because the presented HMM approach concerned many features mentioned above, it is possible to train a high-accuracy probability model. To overcome spareseness problem, we use a back-off strategy which aims at the token sequence T_1^n in $Pr(S_1^n|T_1^n)$ or in $Pr(s_i|T_1^n)$ where T_1^n represents not only a token sequence but also the full set of sequence's features. There are two back-off levels. First level is based on different combinations of tokens and their features, and T_1^n will be assigned in the descending order:
$< s_{-1}, t_{-1}, t_0, f_0 >$, $< s_{-1}, t_0, f_0 >$, $< s_{-1}, t_{-1}, f_0 >$, $< s_{-1}, f_0 >$
where f_i represents the feature set including internal, external and global features. t_i is a token, s_i expresses a HMM state, and i is the i^{th} one relative to current token. Second level is based on different combinations of features, and f_i in first level is assigned in the descending order:
$< f_i^I, f_i^E, f_i^G >$, $< f_i^I, f_i^E >$, $< f_i^I >$
where f_i^I , f_i^E and f_i^G represent internal, external and global features, respectively.

4.3 Method Comparisons

Method comparisons for the three HMM-based models were made on both SRC corpus and GENIA corpus in the same environment settings. We used the same back-off model for concise and mutual information HMM, but not for traditional HMM. Table 9 shows that concise HMM with rule-based features (i.e. concise-ruled) yielded the best result. Traditional HMM obtains good high precision, but low recall since we chose a severe probability model to get the best F-score. It is also noticed that the performance of MI-HMM turned out to be the worst because the back-off model was used to optimize concise HMM. On the other hand, Table 10 shows all kinds of features turned out to be positive effect ($f^E > f^I > f^G$) for concise HMM. Such result is similar to that concluded from [10]. Table 11 lists the comparisons of the presented approaches to other well-known approaches on the public evaluation GENIA 3.x corpus. It is noticed that the presented rule-based approach with its simple general rules outperformed the other two complicated rule-based systems. On the other hand, the performance of the presented concise HMM-based models is comparable to the best model presented in [4]. However, we do not need any dictionary or rules in our model.

Future work includes the manual annotation correction of SRC for fine classification, exploitation of dictionaries for better recognition performance and the improvement of the resolution for coordination variants by using the semantic type information of biomedical thesaurus like UMLS. In addition, novel mining techniques to resolve other types of term variants should be explored for full NER automation.

Acknowledgements

This research is partially supported by MediaTek Research Center, National Chiao Tung University, Taiwan and partially supported by National Science Council under the contract NSC 93-2213-E-009-074.

References

1. Fukuda, K. and Tsunoda, T. and Tamura, A. and Takagi, T.: Towards Information Extraction: identifying Protein Names from Biological Papers. The 3rd Pacific Symposium on Biocomputing. (1998) 707-718.
2. Hou, W. J. and Chen, H. H.: Enhancing Performance of Protein Name Recognizers using Collocation. ACL 2003 Workshop on Natural Language Processing in Biomedicine, (2003) 25-32.
3. Lee, K.J. and Hwang, Y.S. and Rim, H.C.: Two-Phase Biomedical NE Recognition based on SVMs. ACL 2003 Workshop on Natural Language Processing in Biomedicine, (2003) 33-40.
4. Lin, Y. and Tsai, T. and Chiou, W. and Wu K. and Sung, T.-Y. and Hsu, W-L.: A Maximum Entropy Approach to Biomedical Named Entity Recognition. 4th Workshop on Data Mining in Bioinformatics (2004).
5. Olsson, F., Eriksson, G., Franzen, K., Asker, L., and Liden, P.: Notions of Correctness when Evaluating Protein Name Taggers. 19th International Conference on Computational Linguistics. (2002) 765-771.
6. Settles, B.: Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. Int'l Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA), Geneva, Switzerland (2004).
7. Takeuchi, K. and Collier, N.: Bio-Medical Entity Extraction using Support Vector Machines. ACL 2003 Workshop on Natural Language Processing in Biomedicine, (2003) 57-64.
8. Tsuruoka, Y. and Tsujii, J.: Boosting Precision and Recall of Dictionary-based Protein Name Recognition. ACL 2003 Workshop on Natural Language Processing in Biomedicine (2003) 41-48.
9. Zhou, G.D. and Su, J.: Named Entity Recognition using an HMM-based Chunk Tagger. 40th Annual Meeting of the Association for Computational Linguistics (2002).
10. Zhou, G., Zhang, J., Su, J., Shen, D. and Tan, C. L.: Recognizing Names in Biomedical Texts: A Machine Learning Approach. Bioinformatics, Vol. 20, (2004)1178-1190.