Signet: Web Information Retrieval with NE Disambiguation based on HMM and CRF

S. Balaji and S. Sasikala, *Member, IACSIT*

***Abstract*—As compared to many other techniques used in natural language processing, hidden markov models (HMMs) are an extremely flexible tool and has been successfully applied to a wide variety of information extraction tasks. This work focus on webpage perceptive through model of Hierarchical Conditional Random Fields (i.e. HCRF) and offer results in free text segmentation and labelling. This paper specially addresses the problem of research community of academic people integration (SIGNET-similar interest group) through perceiving the entities of them.**

***Index Terms*— HMM, HCRF, Named-Entity, SIGNET.**

I. Introduction

Information Extraction, the task of locating textual mentions of specific types of entities and their relationships, aims at representing the information contained in text documents in a structured format that is more amenable to applications in data mining, question answering, or the semantic web[1]. The goal of this paper is to design information extraction model that obtain improved performance by exploiting types of evidence that have not been explored in previous approaches. In this paper, we introduce a novel framework called SIGNET that enables bidirectional integration of page structure understanding and text understanding in an iterative manner for offering the community formation of the research students to explore the knowledge through named-entity extraction.

A. HMM and CRF in NE Recognition

This work on webpage understanding makes the first attempt toward such an integrated solution. It first uses the HCRF model to label the HTML elements and then uses the EM (Expectation-Maximization) model to segment the text fragment within the HTML elements considering their labels assigned by the HCRF model.

Fig. 1 illustrates the state transition diagram of a two-level HHMM. At the top level there are two parent states {A,B}. The parent A has three children, i.e. ch(A) = {1, 2, 3} and B has four, i.e. ch(B) = {1, 2, 3, 4}. Note that we have assumed that the parents share some common children, i.e. ch(A) ∩ ch(B) = {1, 2, 3}. For example, in applications such as Part-of-Speech (POS) tagging, the output variables are a sequence of POS tags that we want to predict from the input sentence. In image scene segmentation the output variables are 2D arrays of scene interpretation of the raw pixels [4]. For training HMM the Baum-Welch Algorithm is applied.

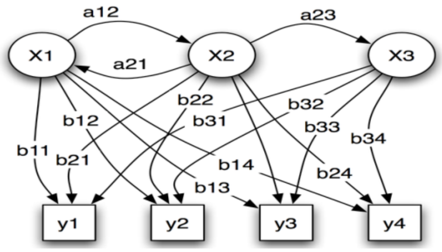


Fig. 1.Probabilistic parameters of a hidden Markov model
x — states, y — possible observations, a — state transition probabilities, b — output probabilities

B. Baum-Welch Algorithm

Problem: Learn the parameters of HMM, transition probabilities A and emission probabilities B.

Input: Set of possible states in HMM and observation sequence O.

Output: Transition probabilities A and emission probabilities B.

It is a special case of Expectation-Maximization algorithm. It can do completely do unsupervised learning of A and B.

The transition probability is:

$$\hat{a}_{ij} = \frac{E(i \rightarrow j)}{\sum_{q\varepsilon Q} E(i \rightarrow q)} \tag{1}$$

The emission probability of symbol vk in state j is calculated as:

$$\hat{b}_j(v_k) = \frac{E(i, v_k)}{\sum_{i=1}^{|v|} E(i, v_i)} \tag{2}$$

II. HHMM And Expectation-Maximisation (EM) Algorithm

Denote by x = (v, h), where # is the subset of visible variables, and h the hidden. The EM attempts to maximise the data log-likelihood log Pr(v|w):

$$\omega = \underset{(w)}{\arg\max} \log PR(v|w) = \arg\max \log \sum_h PR(v|w) \tag{3}$$

where w is the model parameters. In Bayesian Networks w is the set of all local conditional distribution:

$$\{PR(X_i | pa(i))\}_{i=1}^{N} \tag{4}$$

The summation inside the log function couples the two variables # and h. To decouple them, applying the Jensen's inequality to the concave log function, we have:

Manuscript received May 17, 2012; revised June 13, 2012.

S.Balaji is with Sengunthar Engineering College, Tiruchengode, Tamilnadu, India (e-mail: hereiambalaji@yahoomail.com).

S.Sasikala was with KSR College of arts and science, Tiruchengode, Tamilnadu, India. She is now with the Department of Computer Science (e-mail: sasi_sss123@rediff.com).

$$\ell(w) = \log \sum\nolimits_h PR(v,h\Big|w) \geq \sum\nolimits_h Q(h) \log \frac{PR(v,h|w)}{Q(h)}$$

$$Q\left[\log PR(v,h|w)\right] + H\left[Q\right] \quad (5)$$

for any proper distribution Q(h). A nice property of the lower-bound is that the gap between L(w) and its lower-bound is closed by setting Q(h) = Pr(h|w;w). Since log Pr(v, h|w) is typically decomposable into the sum of simpler components, the lower bound nicely decouples variables. Let Q = RQ[log Pr(v, h|w)], since H[Q] does not depend on w, maximizing the lower-bound with respect to w is equivalent to maximizing Q. This suggests an iterative procedure which loops through two steps until convergence:

E-step: compute Q(h) = Pr(h|v;wt) and M-step: optimize the parameter (6)

Essentially, the M-step increases the lower-bound, and the E-step closes the gap between the true log-likelihood and the lower-bound. The overall effect is that the log-likelihood monotonically increases until it reaches a local maximum. In computer vision, CRFs are often used for image segmentation, object recognition and as a general approach to combine features from different sources. Here in this paper a model has been proposed for named entity extraction based on Semi-CRF with HMM and this model derives the PERSON_NAME and AREA_OF _EXPERTISE this seems to be useful for the extraction and reuse the same.

$$(H,S)^* = \arg\max_{(S,H)} P(H,S\big|X) \quad (7)$$

III. Overall Workflow of Signet

A. SIGNET Phase #1

Preprocessing: Segment, POS tagging and general NER is primarily conducted Generating Institution NE Candidates: First,AREA or SPECIALIZATION and NAME are triggered by domain word list and some word features respectively. Here categorize the triggering word features into six classes: alphabet string, alphanumeric string, digits, alphabet string and other symbols. Then AREA is triggered by NAME candidate as well as some clue words indicating type information to some extent. In this step the model structure (topology) of HHMM is dynamically constructed, and some conjunction words or punctuations and specified maximum length of AREA OF EXPERTISE NE are used to control it.

B. SIGNET Phase #2

Disambiguating Candidates: In this module, boundary and classification ambiguities between candidates are resolved simultaneously. And Viterbi algorithm is applied for most-likely state sequences based on the HHMM topology. Both detection and disambiguation are formulated as a ranking problem, using a ranking function that takes as arguments a proper name and a named entity. The value computed by the ranking function is a measure of the similarity between the document context of the proper name and the text of the article corresponding to the argument entity[2]. Standard similarity functions (e.g., TF-IDF cosine similarity are based on the assumption that both the context and the article have many relevant words in common. Hence the raw text from the academic websites (.edu and .ac.in extensions) has been considered to be the input and based on the data the named entity PERSON and EXPERTISE_AREA are extracted and depending on the domain the data is clustered. So it is possible to answer relational queries such as:

1) Who are all the specialized persons in "Data Mining" in indian univeristies?
2) "Sasikala Subramani", assistant professor in KSR College is specialized in which domain?

Thus the system has been organized and tested as mentioned.

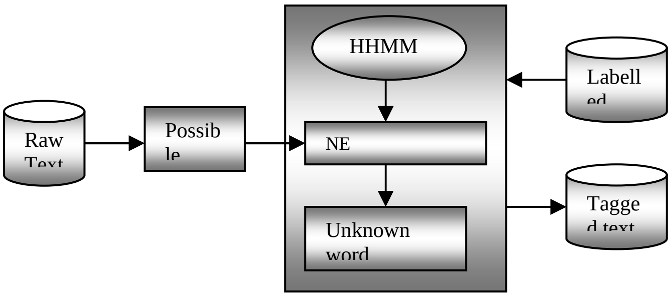


Fig. 2. Flow of SIGNET

IV. Experiments

This application is on building a social network among persons and area of expertise by extracting their relationship from crawled webpages. Therefore, in this experiment, the two most important named entities considered are PERSON_NAME and AREA_OF _EXPERTISE. The webpages used in the experiments are crawled from academic sites and institution websites and personal blogs. The open source utility, Deixto[3] has been used for creating the vision tree with HTML tags and to consider the weblog data.In order to better show the effectiveness of the proposed framework, it is only selected some webpages containing multiple mentions of the same entity. These pages include biography and personal homepages. It is randomly sampled 25 pages for training and 100 pages for testing. In this experiment, we only compared the results from the traditional named entity recognition algorithm with Semi CRF and HHMM.

V. Conclusion

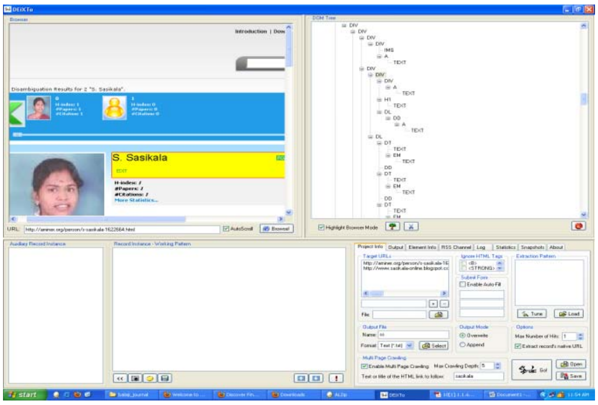
The extraction results of different algorithms are reported in Table 1. It is seen that the proposed SIGNET framework improved both the precision and the recall of the entity extraction task. Especially, it increased the recall significantly. For example, for PERSON, the recall of the BHS was only 65.8%, but the recall of the SIGNET was 71.1%, which was increased. In total, the SIGNET framework increased the recall of the Named Entities. This paper presented a hierarchical HMM (hidden Markov model) and CRF based approach of academic named entity recognition.

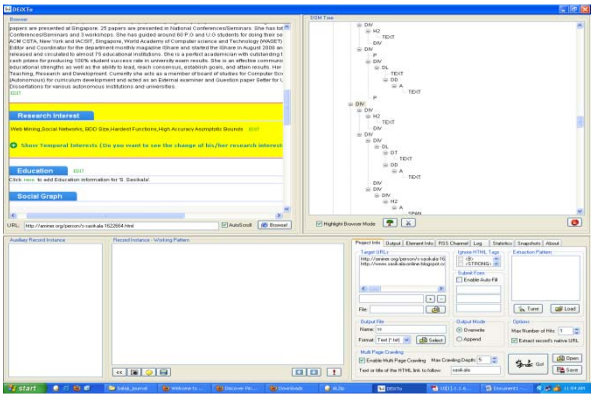
TABLE I: PRECISION AND RECALL COMPARISON

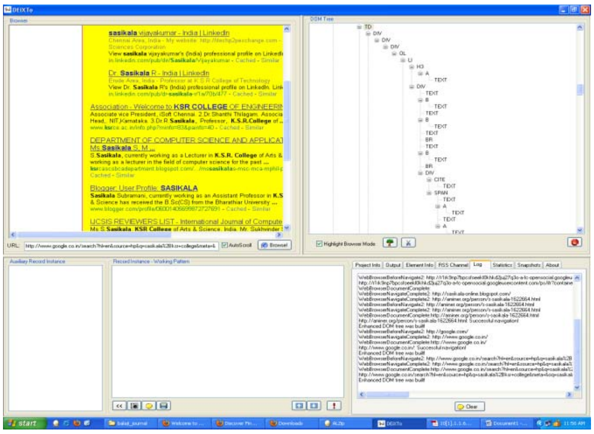
odel	Recall(%)	Precision(%)	F- Score (%)
BHS	65.8	62.8	64.8
NHS	69.1	51.4	58.9
SIGNET	71.1	70.2	70.6

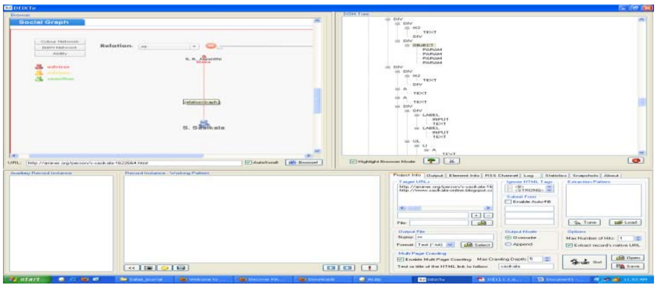
APPENDIX

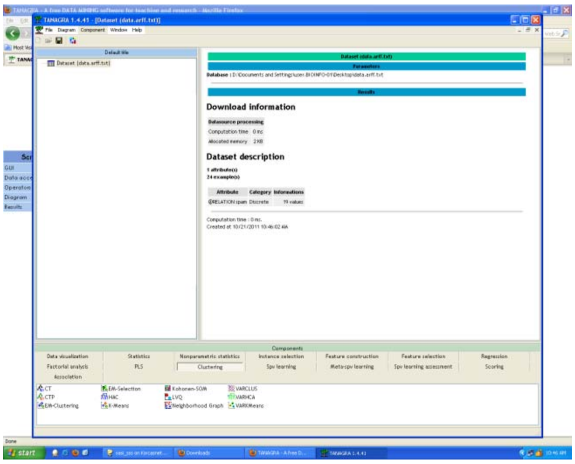
Screenshots of the Simulated Result with Tool Deixto and Tanagra:











REFERENCES

[1] J. Cowie and W. Lehnert, “Information extraction,” Commun.ACM, vol. 39, no. 1, pp. 80–91, 1996.

[2] S. K. Jayanthi andS.Sasikala, “Hyperlink Structure Attribute Analysis for Detecting Link Spamdexing “ International Conference on Advances in Computer Science – AET-ACS 2010, Trivandrum, Kerela, Dec 2010.

[3] Deixto. [Online]. Available: http://www.deixto.com/index.php

[4] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. T. Yu, “Fully automatic wrapper generation for search engines,” in Proceeding of WWW, 2005.

Balaji Subramani has Received the M.Tech (IT) in K.S.Rangasamy College of Technology has received the B.E (CSE) from the Anna University in 2009. He is currently working as an assistant professor in Sengunthar Engineering College. He has published a paper in the journal (IJEST), 3 papers in International Seminars, 5 papers in national seminars and has a total of 8 publications. And participated in various symposiums and workshops held at different places. His area of interest includes Word sense disambiguation, web mining, Information retrieval and social network analysis.

Sasikala currently working as an Assistant Professor in K.S.R. College of Arts & Science has received the B.Sc(CS) from the Bharathiar University, M.Sc(CS) from the Periyar University, M.C.A. from Periyar University , M.Phil from Periyar University, PGDPM & IR from Alagappa university in 2001, 2003, 2006, 2008 and 2009 respectively. And she is currently pursuing her Ph.D in computer science at Bharathiar University. Her area of Doctoral research is Web mining. She secured University First Rank in M.Sc(CS) Programme under Periyar University and received Gold Medal from Tamilnadu State Governor Dr.RamMohanRao in 2004. She has published 5 papers in International Journals, 10 papers in International Conferences/Seminars (1 paper in IEEE Xplore, 1 paper in Springerlink, 1 paper in ACEEE Search digital library, 6 papers with ISBN Numbers). 27 papers are presented in National Conferences/Seminars and acts as a reviewer for 8 journals and 2 conferences and her papers are cited at various publications including IEEE Xplore, International Journals, Wikispaces and Conference Proceedings. She has totally 42 publications and participated in 5 National Conferences/Seminars and 3 workshops and acted as a reviewer for 8 International journals and 2 conferences.