Integrating Genetic Algorithms with Conditional Random Fields to Enhance Question Informer Prediction

Min-Yuh Day [1, 2], Chun-Hung Lu [1, 2], Chorng-Shyong Ong [2], Shih-Hung Wu [3], and Wen-Lian Hsu [1,*] ,*Fellow, IEEE*

[1] *Institute of Information Science, Academia Sinica, Taiwan*
[2] *Department of Information Management , National Taiwan University, Taiwan*
[3] *Department of CSIE, Chaoyang University of Technology, Taiwan*

{myday,enrico,hsu}@iis.sinica.edu.tw; ongcs@im.ntu.edu.tw; shwu@cyut.edu.tw

Outline

- Introduction
- Research Background
- The Hybrid GA-CRF Model
- Experimental Design
- Experimental Results
- Conclusions

Introduction

- Question informers play an important role in enhancing question classification for factual question answering
- Question Informer
  - choosing a minimal, appropriate contiguous span of a question token, or tokens, as the informer span of a question, which is adequate for question classification.
- An example of Question Informer
  - "What is the biggest city in the United States?"
  - Question informer: "city"
  - "city" is the most important clue in the question for question classification.

Introduction (cont.)

- Previous works have used Conditional Random Fields (CRFs) to identify question informer spans.
- We propose a hybrid approach that integrates GA with CRF to optimize feature subset selection in CRF-based question informer prediction models.

Research Background

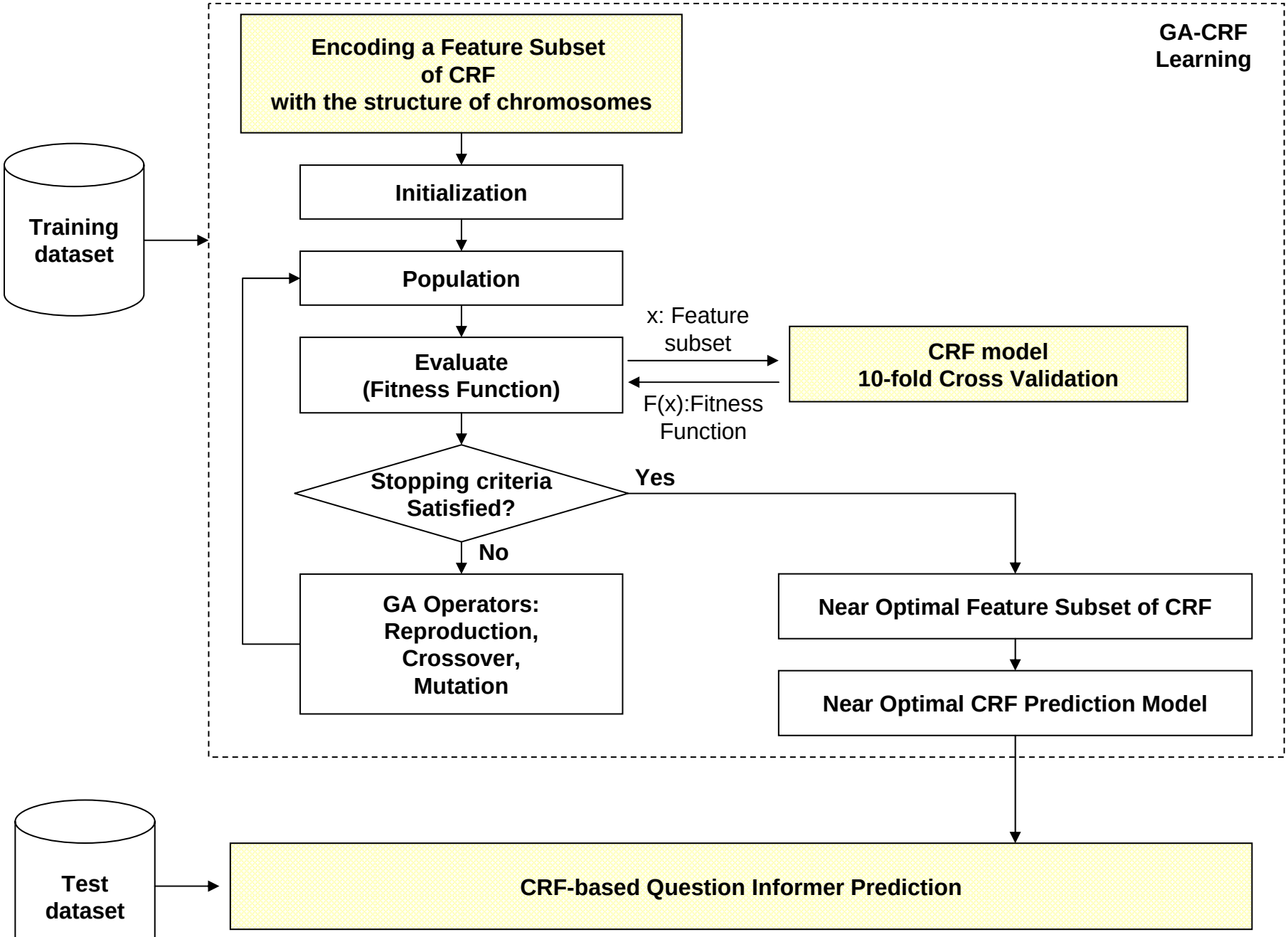
- Conditional Random Fields (CRFs)
  - A framework for building probabilistic models
    - To segment and label sequence data
    - A CRF models *Pr(y|x)* using a Markov random field
  - Advantage over traditional models
    - Hidden Markov Models (HMMs)
    - Maximum Entropy Markov Models (MEMMs)
  - CRF++
    - Open source implementation of CRFs
    - Segmenting and labeling sequenced data
    - Flexible to redefine feature sets in feature templates

Research Background (cont.)

- Genetic Algorithms (GAs)
  - A class of heuristic search methods and computational models of adaptation and evolution based on mechanics of natural selection and genetics.
  - Feature selection in machine learning
    - Feature subset optimization

The Hybrid GA-CRF Model

- Encoding a feature subset of CRF with the structure of chromosomes
- Initialization
- Population
- Evaluate (Fitness Function)
- CRF model 10-fold Cross validation
- Stopping criteria satisfied
- Apply GA operators and produce a new generation
- Apply the selected feature subsets to CRF test dataset



Hybrid GA-CRF Approach for Question Informer Prediction

Gene structure of chromosomes for a feature subset

Population	Feature subset selection						
	F_1	F_2	F_3	…	F_{n-2}	F_{n-1}	F_n
Chromosome 1	1	0	1	…	0	1	1
Chromosome 2	0	0	1	…	1	1	0
Chromosome 3	1	1	0	…	0	1	1
. . .				. . .			
Chromosome m-2	0	1	1	…	1	0	1
Chromosome m-1	1	0	0	…	0	1	1
Chromosome m	1	0	1	…	1	1	0

Example of feature subset encoding for GA

Feature	F_1	F_2	F_3	…	F_{n-2}	F_{n-1}	F_n
Chromosome	1	0	1	…	0	1	1

Feature subset = $\{F_1, F_3, \ldots, F_{n-1}, F_n\}$

Experimental Design

- Data set
  - UIUC QC dataset (Li and Roth, 2002)
  - Question informer dataset (Krishnan et al., 2005)
- Training questions: 5500
- Test questions: 500

Features of Question Informer

- Question informer tags for CRF model
  - O-QIF0: outside and before a question informer
  - B-QIF1: the start of question informer
  - O-QIF2: outside and after a question informer
- 21 basic feature candidates
  - Word, POS, heuristic informer, Parser Information, Token Information, Question wh-word, length, position.
- 5 sliding windows
- We Generate 105 (21*5) features (genes) for each chromosome

Features for question informer prediction

ID	Feature name	Description	Feature Template for CRF ++	F-score	Feature Rank
1	Word	Word	U01:%x[0,0]	58.35	1
2	POS	POS	U01:%x[0,1]	48.29	6
3	HQI	Heuristic Informer	U01:%x[0,2]	52.21	4
4	Token	Token	U01:%x[0,3]	58.35	2
5	ParserL0	Parser Level 0	U01:%x[0,4]	58.35	3
6	ParserL1	Parser Level 1	U01:%x[0,5]	50.98	5
7	ParserL2	Parser Level 2	U01:%x[0,6]	48.13	7
8	ParserL3	Parser Level 3	U01:%x[0,7]	37.76	9
9	ParserL4	Parser Level 4	U01:%x[0,8]	38.45	8
10	ParserL5	Parser Level 5	U01:%x[0,9]	21.45	17
11	ParserL6	Parser Level 6	U01:%x[0,10]	22.43	13
12	IsTag	Is Tag	U01:%x[0,11]	21.57	15
13	IsNum	Is Number	U01:%x[0,12]	21.57	16
14	IsPrevTag	Is Previous Tag	U01:%x[0,13]	21.21	18
15	IsNextTag	Is Next Tag	U01:%x[0,14]	28.75	11
16	IsEdge	Is Edge	U01:%x[0,15]	21.58	14
17	IsBegin	Is Begin	U01:%x[0,16]	15.45	20
18	IsEnd	Is End	U01:%x[0,17]	28.26	12
19	Wh-word	Question Wh-word (6W1H1O)	U01:%x[0,18]	30.17	10
20	Length	Question Length	U01:%x[0,19]	20.93	19
21	Position	Token Position	U01:%x[0,20]	13.17	21

Data format for CRF model

Question: “What is the oldest city in the United States?”

	Features f_{ij} for x_i																				
j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
i	x_i	POS	HQI	Token	ParL0	ParL1	ParL2	ParL3	ParL4	ParL5	ParL6	IsTag	IsNum	IsPrevTag	IsNextTag	IsEdge	IsBegin	IsEnd	Wh-word	L	P	y_i
0	What	WP	city	What	What	WP_1	WHNP_1	Null_1	Null_1	Null_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag1	IsNextTag0	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	1	O-QIF0
1	is	VBZ	city	is	is	VBZ_1	Null_1	Null_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag1	IsNextTag0	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	2	O-QIF0
2	the	DT	city	the	the	DT_1	NP_1	Null_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag1	IsNextTag0	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	3	O-QIF0
3	oldest	JJS	city	oldest	oldest	JJS_1	NP_1	Null_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag1	IsNextTag0	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	4	O-QIF0
4	city	NN	city	city	city	NN_1	NP_1	Null_1	VP_1	SQ_1	SBARQ_1	IsTag1	IsNum1	IsPrevTag0	IsNextTag0	IsEdge1	IsBegin1	IsEnd1	Wh_what	10	5	B-QIF1
5	in	IN	city	in	in	IN_1	Null_2	PP_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag0	IsNextTag1	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	6	O-QIF2
6	the	DT	city	the	the	DT_2	NP_2	PP_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag0	IsNextTag1	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	7	O-QIF2
7	United	NNP	city	United	United	NNP_1	NP_2	PP_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag0	IsNextTag1	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	8	O-QIF2
8	States	NNS	city	States	States	NNS_1	NP_2	PP_1	VP_1	SQ_1	SBARQ_1	IsTag0	IsNum0	IsPrevTag0	IsNextTag1	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	9	O-QIF2
9	?	.	city	?	?	._1	Null_3	Null_2	Null_2	Null_2	SBARQ_1	IsTag0	IsNum0	IsPrevTag0	IsNextTag1	IsEdge0	IsBegin0	IsEnd0	Wh_what	10	10	O-QIF2

	Features f_{ij} for x_i		
j	0	1	
i	x_i	POS	y_i
0	What	WP	O-QIF0
1	is	VBZ	O-QIF0
2	the	DT	O-QIF0
3	oldest	JJS	O-QIF0
4	city	NN	B-QIF1
5	in	IN	O-QIF2
6	the	DT	O-QIF2
7	United	NNP	O-QIF2
8	States	NNPS	O-QIF2
9	?	.	O-QIF2

Sliding Windows

i -2
i -1
i +0
i +1
i +2

	Features f_{ij} for x_i		
j	0	1	
i	x_i	POS	y_i
-2	the	DT	O-QIF0
-1	oldest	JJS	O-QIF0
0	city	NN	B-QIF1
+1	in	IN	O-QIF2
+2	the	DT	O-QIF2

Features f_{ij} for x_i = Uid:*%x[i, j]*

the	=> $f_{-2,0}$	=> U00:%x[-2,0]	=> F1
oldest	=> $f_{-1,0}$	=> U01:%x[-1,0]	=> F2
city	=> $f_{0,0}$	=> U02:%x[0,0]	=> F3
in	=> $f_{+1,0}$	=> U03:%x[+1,0]	=> F4
the	=> $f_{+2,0}$	=> U04:%x[+2,0]	=> F5
DT	=> $f_{-2,1}$	=> U05%x[-2,1]	=> F6
JJS	=> $f_{-1,1}$	=> U06:%x[-1,1]	=> F7
NN	=> $f_{0,1}$	=> U07:%x[0,1]	=> F8
IN	=> $f_{+1,1}$	=> U08:%x[+1,1]	=> F9
DT	=> $f_{+2,1}$	=> U09:%x[+2,1]	=> F10

	Features f_{ij} for x_i		
j	0	1	
i	x_i	POS	y_i
-2	the [-2,0]	DT [-2,1]	O-QIF0
-1	oldest[-1, 0]	JJS [-1,1]	O-QIF0
0	city [0, 0]	NN [0,1]	B-QIF1
+1	in [+1, 0]	IN [+1,1]	O-QIF2
+2	the [+2, 0]	DT [+2,1]	O-QIF2

Feature generation and feature template for CRF++

Feature	Features	Feature Template	Feature ID
the	$f_{-2,0}$	*U*00:%x[-2,0]	F1
oldest	$f_{-1,0}$	U01:%x[-1,0]	F2
city	$f_{0,0}$	U02:%x[0,0]	F3
in	$f_{+1,0}$	U03:%x[+1,0]	F4
the	$f_{+2,0}$	U04:%x[+2,0]	F5
DT	$f_{-2,1}$	U05%x[-2,1]	F6
JJS	$f_{-1,1}$	U06:%x[-1,1]	F7
NN	$f_{0,1}$	U07:%x[0,1]	F8
IN	$f_{+1,1}$	U08:%x[+1,1]	F9
DT	$f_{+2,1}$	U09:%x[+2,1]	F10

Encoding a feature subset with the structure of chromosomes for GA

Features	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}
Chromosome	1	0	1	1	0	0	1	1	0	1

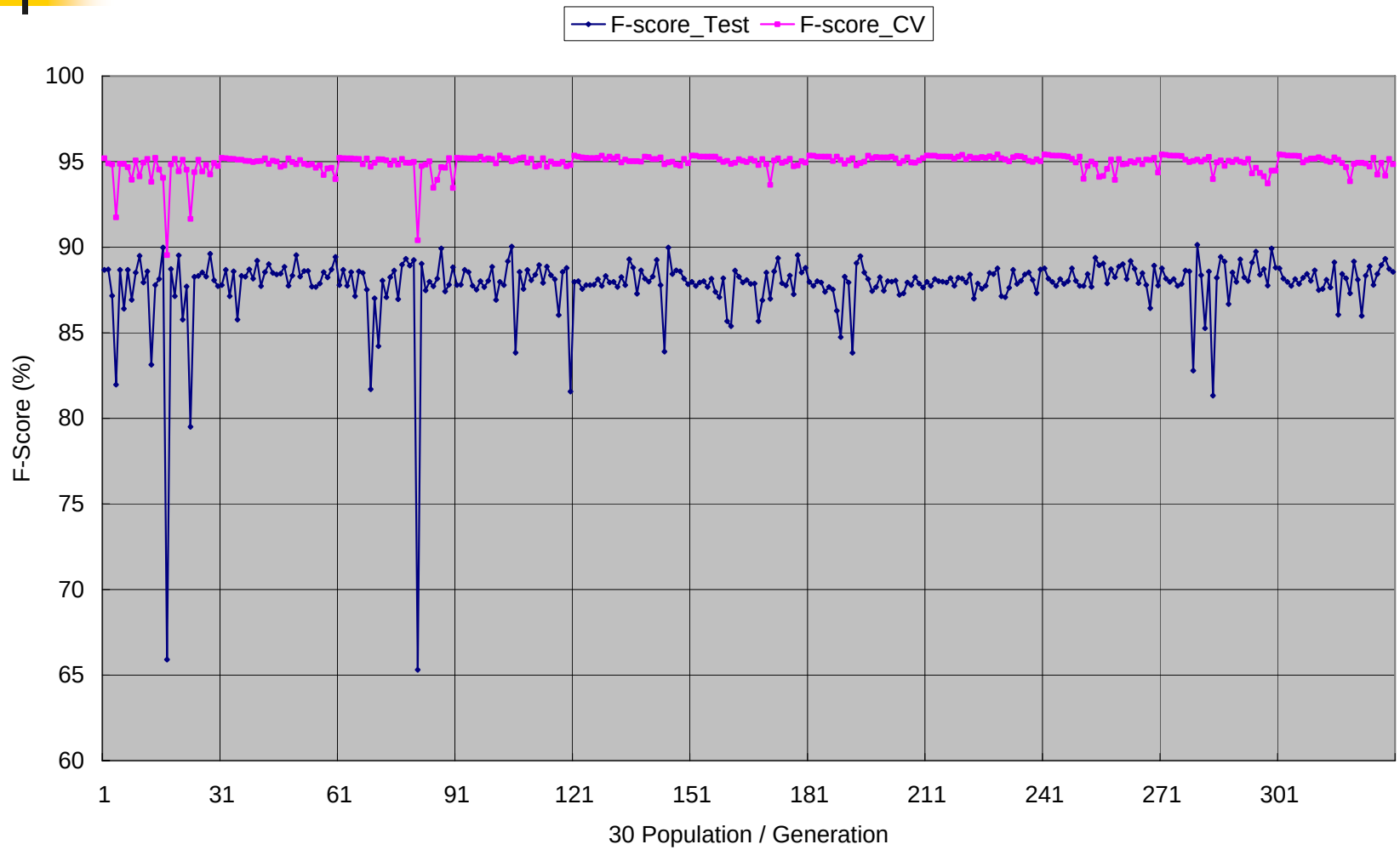
Feature subset = {F_1, F_3, F_4, F_7, F_8 , F_{10}}

Features f_{ij}* for x_i = Uid:*%x[i, j]

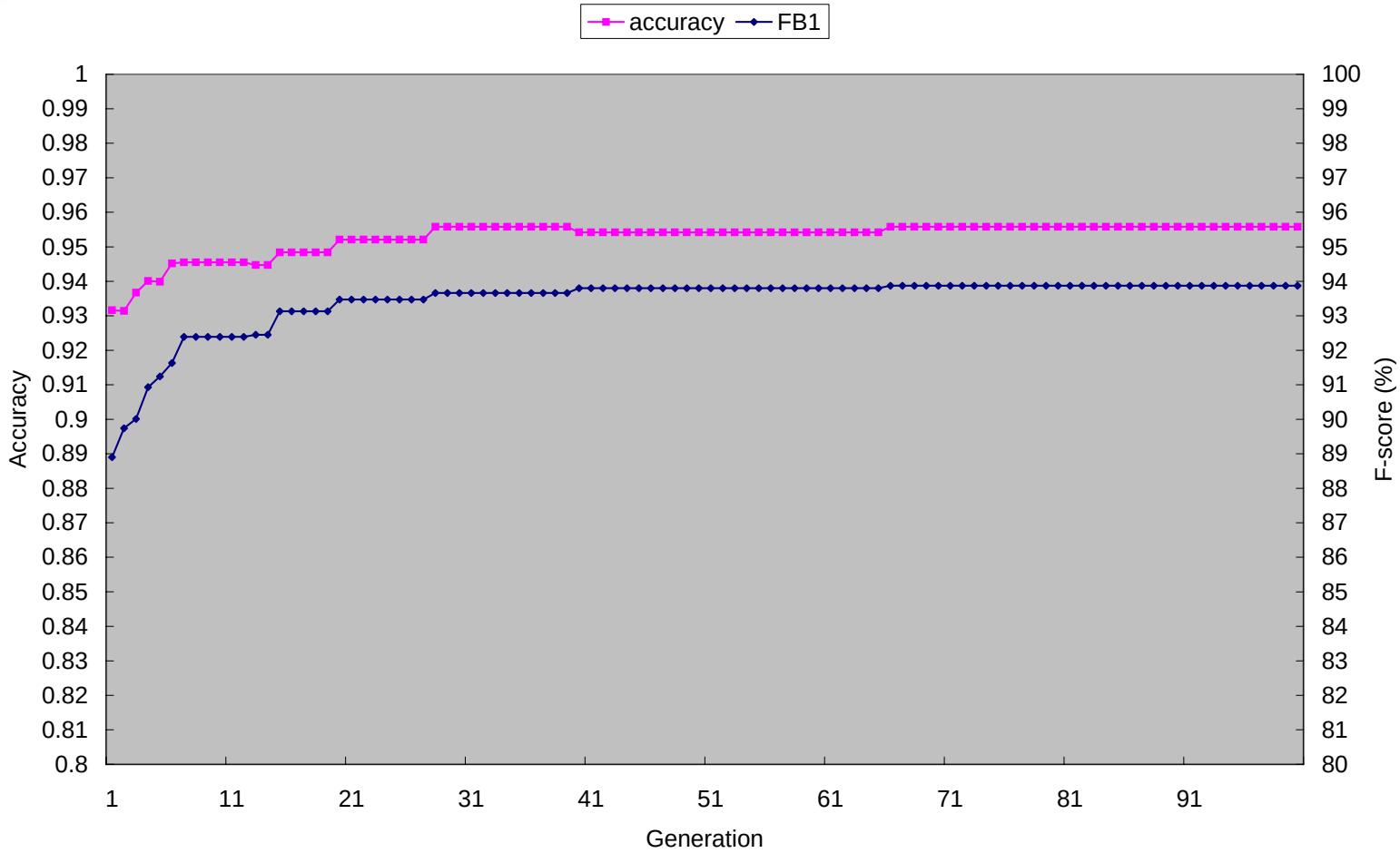
Feature	Features	Feature Template	Feature ID
the	*f-2,0*	*U*00:%x[-2,0]	F1
oldest	*f-1,0*	U01:%x[-1,0]	F2
city	*f0,0*	U02:%x[0,0]	F3
in	*f+1,0*	U03:%x[+1,0]	F4
the	*f+2,0*	U04:%x[+2,0]	F5
DT	*f-2,1*	U05%x[-2,1]	F6
JJS	*f-1,1*	U06:%x[-1,1]	F7
NN	*f0,1*	U07:%x[0,1]	F8
IN	*f+1,1*	U08:%x[+1,1]	F9
DT	*f+2,1*	U09:%x[+2,1]	F10

There are 105 feature subsets in total
(21 basic features * 5 sliding windows)

Experimental Results



Experimental results of CRF-based question informer prediction using GA



Population: 40, Crossover: 80%, Mutation:10%, Generation:100

Optimal feature subset for the CRF model selected by GA

GA-CRF Model

Near Optimal Chromosome:

0 0 0 0 0 1 0 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0
1 1 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0
0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 0 0
1 1 1 0 0 1

Near Optimal Feature Subsets for CRF model:

U001:%x[-2,1] U002:%x[0,1] U003:%x[1,1] U004:%x[-1,2]
U005:%x[0,2] U006:%x[1,2] U007:%x[2,2] U008:%x[-2,3]
U009:%x[-2,5] U010:%x[-1,5] U011:%x[-2,6] U012:%x[1,6]
U013:%x[2,6] U014:%x[2,7] U015:%x[0,8] U016:%x[1,8]
U017:%x[-2,9] U018:%x[1,9] U019:%x[1,10] U020:%x[-2,11]
U021:%x[-2,12] U022:%x[0,12] U023:%x[0,13] U024:%x[2,13]
U025:%x[-2,14] U026:%x[-1,14] U027:%x[2,14] U028:%x[0,16]
U029:%x[1,16] U030:%x[2,16] U031:%x[-2,17] U032:%x[-1,17]
U033:%x[0,17] U034:%x[1,17] U035:%x[-2,19] U036:%x[-1,19]
U037:%x[2,19] U038:%x[-2,20] U039:%x[-1,20] U040:%x[2,20]

Experimental Result of the proposed hybrid GA-CRF model for question informer prediction

Question Informer Prediction	Accuracy	Recall	Precision	F-score
Traditional CRF Model (All features) (105 features)	93.16%	94.33%	84.07%	88.90
GA-CRF Model (Near optimal feature subset) (40 features)	95.58%	95.79%	92.04%	93.87

Conclusions

- We have proposed a hybrid approach that integrates Genetic Algorithm (GA) with Conditional Random Field (CRF) to optimize feature subset selection in a CRF-based model for question informer prediction.
- The experimental results show that the proposed hybrid GA-CRF model of question informer prediction improves the accuracy of the traditional CRF model.
- By using GA to optimize the selection of the feature subset in CRF-based question informer prediction, we can improve the F-score from 88.9% to 93.87%, and reduce the number of features from 105 to 40.

Q & A

Integrating Genetic Algorithms with Conditional Random Fields to Enhance Question Informer Prediction

Min-Yuh Day [a, b], Chun-Hung Lu [a, b], Chorng-Shyong Ong [b], Shih-Hung Wu [c], and Wen-Lian Hsu [a,*] ,*Fellow, IEEE*

[a] *Institute of Information Science, Academia Sinica, Taiwan*

[b] *Department of Information Management , National Taiwan University, Taiwan*

[c] *Department of CSIE, Chaoyang University of Technology, Taiwan*

{myday,enrico,hsu}@iis.sinica.edu.tw; ongcs@im.ntu.edu.tw; shwu@cyut.edu.tw