Maximum Entropy Classifier Ensembling using Genetic Algorithm for NER in Bengali

Asif Ekbal[1] **and Sriparna Saha**[2]

[1] Department of Computational Linguistics, University of Heidelberg, Germany, Email: asif.ekbal@gmail.com
[2] IWR, University of Heidelberg, Germany, Email: sriparna.saha@gmail.com

May 21, 2010

Outline

Named Entity Recognition I

NER-Named Entity Recognition (NER) involves identification of proper names in texts, and classification into a set of pre-defined categories of interest as:

- Person names (names of people)
- Organization names (companies, government organizations, committees, etc.)
- Location names (cities, countries etc)
- Miscellaneous names (Date, time, number, percentage, monetary expressions, number expressions and measurement expressions)

Approaches for NER I

- Rule-based NER
  1. based on handcrafted set of rules
  2. suffers from adaptability to a new domain and/or languages
- Machine learning based NER: Supervised, Semi-supervised and Unsupervised
  1. adaptable to different domains and languages
  2. maintenance cost is less
  3. difficult to obtain large annotated corpus for resource-constrained languages
- Hybrid NER
  1. combination of both machine learning and rule-based
  2. maintenance of rule-based component still persists
  3. difficult to obtain large annotated corpus for resource-constrained languages

Problems for NER in Indian Languages I

- Lacks capitalization information
- Indian names are more diverse
  1. Lot of person names appear in the dictionary with other specific meanings
  2. For e.g., KabiTA (Person name vs. Common noun with meaning poem)
- High inflectional nature of Indian languages
  1. Richest and most challenging sets of linguistic and statistical features resulting in long and complex wordforms
- Scarcity of Corpus and NE annotated corpus
- Free word order nature of Indian languages
- Resource-constrained environment of Indian languages
  1. POS taggers, morphological analyzers, name lists etc. are not available in the web
- Non-availability of sufficient published works

Motivation and Contribution I

- The language-Bengali
    1. Emerged in AD 1000
    2. Spoken in West Bengal, Tripura, Assam and Jharkhand states of India (Rank 2 in India)
    3. National language of Bangladesh
    4. Rank 5th in the World in terms of native speakers
- NER in Indian languages
    1. More difficult and challenging
    2. Efforts are still in infancy
- NER system for a less computerized language
- Proposal of a generalized approach that could be applicable for many languages
- Use of Genetic Algorithm (GA) for classifier ensemble is noble
- Application of GA for solving any kind of NLP problem is new

Classifier Ensembling I

Classifier Ensembling

- Well-known in the area of machine learning
- Concept of combining classifiers to improve the performance
- Determining the appropriate classifier combination : very crucial problem

Our proposal

- Posed the classifier ensemble selection problem under the single objective optimization framework
- Solution by genetic algorithm(GA)

Single Objective Formulation of Classifier Ensemble Problem I

Suppose, the N number of available classifiers denoted by $C_1, \ldots, C_N$.
Let, $\mathcal{A} = \{C_i : i = 1; N\}$.
Classifier ensemble selection problem :
Find a set of classifiers B

- Optimize a function $F(B)$
- $B \subseteq A$
- F: a classification quality measure of the combined classifiers, $F \in \{\text{recall}, \text{precision}, \text{F-measure}\}$
- Here $F =$ F-measure

Goal of the paper I

- Maximum Entropy : base classifier
- Depending on various feature representations, different versions of ME are made
- Features are language independent
- GA used to find appropriate classifier ensemble
- System evaluated for Bengali, a resource poor language

Genetic Algorithm I

Genetic Algorithms:

- Randomized search and optimization techniques guided by the principles of evolution and genetics
- Evolution produced good individuals, similar principles might work for solving complex problems
- Many problems can not be solved in polynomial amount of time using a deterministic algorithm
- Near optimal solutions requiring less time more desirable than optimal solutions with huge amount of time
- Perform search in complex, large and multimodal landscapes

Genetic Algorithm II

Genetic Algorithms $\Longleftrightarrow$	Nature
A solution (phenotype)	Individual
Representation of a solution (genotype)	Chromosome
Components of the solution	Genes
Set of solutions	Population
Survival of the fittest (selection)	Darwins theory
Search operators	Crossover and mutation
Iterative procedure	Generations

- Parameters of the search space encoded in the form of strings (called *chromosomes*)
- A collection of such *chromosomes* called a *population*
- Initial step: A random population representing different points in the search space

Genetic Algorithm III

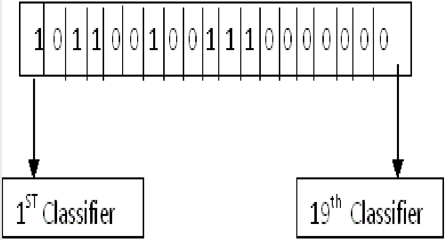
- *objective* or *fitness* function: associated with each string
  - represents the degree of *goodness* of the string
- Selection
  - Based on the principle of survival of the fittest, a few of the strings selected
- Biologically inspired operators like *crossover* and *mutation* applied on these strings to yield a new generation of strings
- Process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition satisfied

Basic Steps of Genetic Algorithm I

1. $t = 0$
2. initialize population $P(t)$ /* $Popsize = |P|$ */
3. for $i = 1$ to $Popsize$
   compute fitness $P(t)$
4. $t = t + 1$
5. if termination criterion achieved go to step 10
6. select (P)
7. crossover (P)
8. mutate (P)
9. go to step 3
10. output best chromosome and stop

End

String Representation I



Total number of available classifiers: M
Length of the chromosome : M
0 in the i^{th} position of chromosome$\rightarrow$ i^{th}classifier does not participate in ensemble
1 in the i^{th} position of a chromosome $\rightarrow$ i^{th} classifier participates in the classifier ensemble

Fitness Computation I

1. N: number of classifiers present in the ensemble represented in a particular chromosome (Total N number of 1's in that chromosome)
2. Overall average F-measure values of the 3-fold cross validation on the training data for these N classifiers be F_i, $i = 1 \ldots N$
3. Training data divided into 3 parts
4. Each classifier trained using 2/3 of the training data and tested with the remaining 1/3 part
5. Output class label for each word in the 1/3 training data determined using the weighted voting of these N classifiers' outputs
6. The weight of the o/p label provided by the i^{th} classifier = F_i.
7. The overall F-measure value of this ensemble classifier for the 1/3 training data calculated.

Fitness Computation II

8. Average F-measure value of the ensemble classifier used as the fitness value of that particular chromosome

Objective: Maximize F-measure using the search capability of GA

Selection I

- New generation created from a proportion of the existing population
- Individual solutions selected through a fitness-based process
  - Fitter solutions more likely to be selected
- Roulette wheel selection: Resemblance to a Roulette wheel in a casino
  - Fitness function associates a probability of selection with each individual chromosome
  - f_i : the fitness of individual i in the population, its probability of being selected :

$$p_i = \frac{f_i}{\Sigma_{j=1}^{N} f_j},$$

  where N: the number of individuals in the population

Crossover I

- Normal single point crossover
- Suppose, there are 8 classifiers. The two chromosomes look like:
  $P_1 = 0\ 1\ 1\ 0\ 0\ 0\ 1\ 1$
  $P_2 = 1\ 1\ 1\ 0\ 0\ 0\ 1\ 0$
- Consider the crossover point : 4. After single point crossover the new chromosomes will look like:
  $O_1 = 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0$
  $O_2 = 1\ 1\ 1\ 0\ 0\ 0\ 1\ 1.$
- Crossover probability selected adaptively

Mutation I

- Mutation operator applied to each entry of the chromosome
  - Entry randomly replaced by either 0 or 1
- Fitness computation, selection, crossover, and mutation executed for a maximum number of generations
- The best string seen upto the last generation provides the solution
- Elitism implemented at each generation by preserving the best string seen upto that generation in a location outside the population
- On termination, this location contains the best classifier ensemble

Feature Set Used I

1. Context Word: Preceding and succeeding words
2. Word Suffix:
   1. Not necessarily linguistic suffixes
   2. Fixed length character strings stripped from the endings of words
   3. Variable length suffix -binary valued feature
3. Word Prefix
   1. Fixed length character strings stripped from the beginning of the words
4. First Word (binary valued feature): Check whether the current token is the first word in the sentence
5. Length (binary valued): Check whether the length of the current word less than three or not (shorter words rarely NEs)
6. Position (binary valued): Position of the word in the sentence

Feature Set Used II

7. Infrequent (binary valued): Infrequent words in the training corpus most probably NEs
8. Digit features: Binary-valued
   1. Presence and/or the exact number of digits in a token
   2. CntDgt : Token contains digits
   3. FourDgt: Token consists of four digits
   4. TwoDgt: Token consists of two digits
   5. CnsDgt: Token consists of digits only
9. Combination of digits and punctuation symbols
   1. CntDgtCma: Token consists of digits and comma
   2. CntDgtPrd: Token consists of digits and periods
10. Combination of digits and symbols
    1. CntDgtSlsh: Token consists of digit and slash
    2. CntDgtHph: Token consists of digits and hyphen

Feature Set Used III

3. CntDgtPrctg: Token consists of digits and percentages

11. Combination of digit and special symbols
    1. CntDgtSpl: Token consists of digit and special symbol such as $, # etc.
12. Part of Speech (POS) Information: POS tag(s) of the current and/or the surrounding word(s)
    1. SVM-based POS tagger
    2. Accuracy=90.2%

Data Sets I

- Web-based Bengali news Corpus (Ekbal and Bandyopadhyay, 2008)
  1. 34 million wordforms
  2. news data collection of 5 years
- NE annotated corpus
  1. Manually annotated 250K wordforms
  2. IJCNLP-08 Shared Task on NER for South and South East Asian Languages (available at http://ltrc.iiit.ac.in/ner-ssea-08)
- NE Tagset
  1. Person name
  2. Location name
  3. Organization name
  4. Miscellaneous name (date, time, number, percentages, monetary expressions and measurement expressions)

Data Sets II

- IJCNLP-08 NERSSEAL Shared Task Tagset: Fine-grained 12 NE tags (available at http://ltrc.iiit.ac.in/ner-ssea-08)
- Tagset Mapping (12 NE tags$\rightarrow$ 4 NE tags)
    1. NEP $\rightarrow$ Person name
    2. NEL $\rightarrow$ Location name
    3. NEO $\rightarrow$ Organization name
    4. NEN [number], NEM [Measurement] and NETI [time]$\rightarrow$ Miscellaneous name
    5. NETO [title-object], NETE [term expression], NED [designations], NEA [abbreviations], NEB [brand names], NETP [title persons]$\rightarrow$ O

Experimental Results I

- Parameters for GA:
  1. population size=100
  2. number of generations=50
- MaxEnt experiment: OpenNLP Java based ME package (http://maxent.sourceforge.net/)
- Baselines:
  - *Baseline 1*: Majority voting of all classifiers
  - *Baseline 2*
    - Weighted voting of all classifiers
    - Weight: average F-measure value of the 3-fold cross validation on the training data

Training set size: 313K wordforms Test set size: 37K wordforms

Table: Statistics of training and test sets

Set	PER	LOC	ORG	MISC
Training	6,717	5,591	3,070	8,058
Test	648	670	374	1,008

Results I

Table: Feature types and parameters used for training different ME based classifiers for Bengali. X: Denotes the presence of the corresponding feature

Classifier	CW	FW	PRE-SIZE	SUF-SIZE	WL	IW	PW	DI	POS	recall	precision	F-measure
M_1	X	X						X	X	35.59	62.74	45.42
M_2	X	X	3					X	X	63.12	78.61	70.02
M_3	X	X	3	3				X	X	68.81	81.34	74.55
M_4	X	X	3	3	X			X	X	68.65	81.57	74.55
M_5	X	X	3	3	X	X		X	X	69.35	81.37	74.88
M_6	X	X	3	3	X	X	X	X	X	69.15	81.53	74.83
M_7	X	X	4					X	X	65.45	79.43	71.76
M_8	X	X	4	3				X	X	68.42	81.58	74.42
M_9	X	X	3	4				X	X	69.39	81.66	75.03
M_{10}	X	X	4	4				X	X	68.65	81.13	74.37
M_{11}	X	X	4	3	X			X	X	67.81	81.53	74.04
M_{12}	X	X	3	4	X			X	X	69.39	82.02	75.18
M_{13}	X	X	4	4	X			X	X	68.01	81.00	73.94
M_{14}	X	X	4	3	X	X		X	X	68.69	81.46	74.53
M_{15}	X	X	3	4	X	X		X	X	**69.76**	**81.75**	**75.28**
M_{16}	X	X	4	4	X	X		X	X	68.87	80.89	74.40
M_{17}	X	X	4	3	X	X	X	X	X	68.58	81.64	74.54
M_{18}	X	X	3	4	X	X	X	X	X	69.67	81.85	75.27
M_{19}	X	X	4	4	X	X	X	X	X	68.51	81.01	74.24

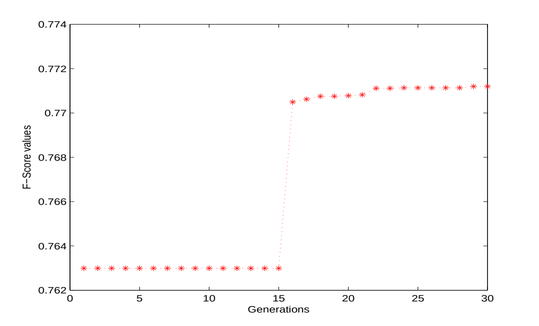
Results continued.. I

Table: Overall results for Bengali

Model	R	P	F
Best classifier	69.76	81.75	75.28
Baseline 1	69.83	82.90	75.81
Baseline 2	70.25	82.97	76.08
GA	71.14	84.07	77.11

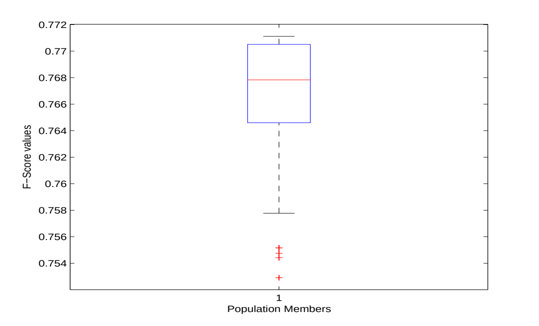
- Classifiers selected: M_2, M_3, M_4, M_5, M_7, M_9, M_{10}, M_{11}, M_{12}, M_{14}, M_{16}, M_{18} and M_{19}.

Variation of the best F-measure values over generations I



- Observation: convergence within 21 generations for this particular data set

Boxplot of the F-measure values of the solutions on the final population I



Conclusion and Future Works I

- Proposed the use of GA to develop a classifier ensemble for NER
- Base classifier: ME framework
- Language independence
- Evaluation with a resource poor language: Bengali
    1. Recall= 71.14%, Precision=84.07%, F-measure=77.11%
    2. Performed better than two conventional *baseline* ensembles

Future Works I

- Incorporation of some more language independent (dynamic NE information etc.) as well as the language specific features to generate more classifiers
- Development of vote based classifier ensembles using some other well-known classifiers like CRF and SVM
- Use of Multiobjective optimization

References I

- A. Ekbal and S. Bandyopadhyay. 2007. Lexical Pattern Learning from Corpus Data for Named Entity Recognition. In Proceedings of the 5th International Conference on Natural Language Processing (ICON), pages 123 128, India.
- A. Ekbal and S. Bandyopadhyay. 2008a. Bengali Named Entity Recognition using Support Vector Machine. In Proceedings of Workshop on NER for South and South East Asian Languages, 3rd International Joint Conference on Natural Languge Processing (IJCNLP), pages 5158, India.
- A. Ekbal and S. Bandyopadhyay. 2008b. Web-based Bengali News Corpus for Lexicon Development and POS Tagging. POLIBITS, ISSN 1870-9044, 37:2029.

References II

- A. Ekbal and S. Bandyopadhyay. 2008c. A Webbased Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal, 42(2):173182.
- A. Ekbal and S. Bandyopadhyay. 2009a. A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi. Linguistic Issues in Language Technology (LiLT), 2(1):144.
- A. Ekbal and S. Bandyopadhyay. 2009b. Voted NER System using Appropriate Unlabeled Data. Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), ACL-IJCNLP 2009, pages 202210.
- A. Ekbal, S.K. Naskar, and S. Bandyopadhyay. 2007. Named Entity Recognition and Transliteration in Bengali. Named Entities: Recognition, Classification and Use, Special Issue of Lingvisticae Investigationes Journal, 30(1):95114.

References III

- D. E. Goldberg. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, New York.

Thank You I

Thank You For Listening