Feature Selection Using Multiobjective Optimization for Named Entity Recognition

Asif Ekbal*, Sriparna Saha†, Christoph S. Garbe‡
*Department of Computational Linguistics, Heidelberg University, Heidelberg, Germany
Email: ,ekbal@cl.uni-heidelberg.de, asif.ekbal@gmail.com
†Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany
Email: [2]sriparna.saha@iwr.uni-heidelberg.de, sriparna.saha@gmail.com, [3] Christoph.Garbe@iwr.uni-heidelberg.de

* First two authors are the joint first authors.

***Abstract*—Appropriate feature selection is a very crucial issue in any machine learning framework, specially in Maximum Entropy (ME). In this paper, the selection of appropriate features for constructing a ME based Named Entity Recognition (NER) system is posed as a multiobjective optimization (MOO) problem. Two classification quality measures, namely recall and precision are simultaneously optimized using the search capability of a popular evolutionary MOO technique, NSGA-II. The proposed technique is evaluated to determine suitable feature combinations for NER in two languages, namely Bengali and English that have significantly different characteristics. Evaluation results yield the recall, precision and F-measure values of 70.76%, 81.88% and 75.91%, respectively for Bengali, and 78.38%, 81.27% and 79.80%, respectively for English. Comparison with an existing ME based NER system shows that our proposed feature selection technique is more efficient than the heuristic based feature selection.**

***Keywords*-Multiobjective Optimization; Feature Selection; Maximum Entropy; Named Entity Recognition.**

I. Introduction

Named Entity Recognition (NER) is an important pipelined module in almost all Natural Language Processing (NLP) application areas such as Information Retrieval, Information Extraction, Machine Translation, Question Answering and Automatic Summarization etc. The objective of Named Entity Recognition (NER) is to identify and classify every word/term in a document into some predefined categories like person name, location name, organization name, miscellaneous name (date, time, percentage and monetary expressions etc.) and "none-of-the-above".

The performance of any classification technique depends on the features of training and test data sets. Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique, commonly used in machine learning, of selecting a subset of relevant features for building robust learning models. In a machine learning approach, feature selection is an optimization problem that involves choosing an appropriate feature subset. In Maximum Entropy (ME) based models, relevant feature selection is a very crucial problem and also a key issue to improve the classifier's performance. Maximum entropy, however, does not provide a method for automatic feature selection. Usually, heuristics are used to find the most relevant set of features in ME model. In this paper, the task of feature selection is posed as a multiobjective optimization (MOO) [1] problem. Generally feature selection problems are solved using some single objective optimization techniques like genetic algorithm (GA) [2]. But, these single objective optimization techniques can only optimize a single quality measure, for e.g., recall, precision or F-measure at a time. Sometimes, a single measure can not capture the quality of a good classifier reliably. A good classifier should have it's recall, precision and F-measure values optimized simultaneously rather than only the high value of any parameter. In this paper, we simultaneously optimize both recall and precision in order to determine the best feature combination for NER in ME framework using a MOO technique [1].

We use a set of language independent features that can be easily obtained for many languages with little effort. Thereafter, a MOO technique based on a popular multiobjective evolutionary algorithm (MOEA), non-dominated sorting GA-II (NSGA-II)[3], is proposed to search for the appropriate subset of features for NER. Initially, the proposed approach is evaluated for a resource poor language like Bengali. In terms of native speakers, Bengali is the *fifth* popular language in the world, *second* in India and the *national* language in Bangladesh. Evaluation results show the effectiveness of the proposed feature selection technique with the overall recall, precision and F-measure values of 70.76%, 81.88% and 75.91%, respectively for Bengali. Thereafter, the system is evaluated for the standard CoNLL-2003 shared task [4] English datasets where it yields the overall recall, precision and F-measure values of 78.38%, 81.27%, and 79.80%, respectively.

II. Named Entity Features

Below, we describe the set of features that we use for our NER tasks. These features are mostly language independent in nature, and can be easily obtained for almost all the languages.

1. **Context words**: These are the preceding and succeeding words of the current word. This is based on the observation

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.477

that surrounding words carry effective information for identification of NEs.

2. **Word suffix and prefix**: Fixed length (say, n) word suffixes and prefixes are very effective to identify NEs and work well for the highly inflective Indian languages. This is also an useful feature for English. Actually, these are the fixed length character strings stripped either from the rightmost or from the leftmost positions of the words. This feature is included with the observation that NEs share some common suffixes and/or prefixes.

3. **First word**: This is a binary valued feature that checks whether the current token is the first word of the sentence or not. We consider this feature with the observation that the first word of the sentence is most likely a NE.

4. **Length of the word**: This binary valued feature checks whether the length of the token is less than a predetermined threshold value and based on the observation that very short words are most probably not the NEs.

5. **Infrequent word** : A cut off frequency has been chosen in order to consider the infrequent words in the training corpus with the observation that very frequent words are rarely NEs. Then, a binary valued feature is defined that fires if the current word appears in this list.

6. **Position of the word**: This checks the position of the word in the sentence. Sometimes, position of a word in a sentence acts as a good indicator for NE identification. In the present work, this binary valued feature is used only for Bengali with the observation that verbs generally appear in the last position of the sentence.

7. **Capitalization**: This binary valued feature is used to check whether the word starts with a capital letter or not. It is found to be an useful feature for English. In the present work, this feature is only used for English as Bengali does not have any capitalization cues.

8. **Part of Speech (POS) information**: POS information of the current and/or the surrounding word(s) are effective for NE identification. In the present work, we consider the POS information only for English and this information was already provided with the training data.

9. **Chunk information**: This is useful for NE identification. In the present work, this is used only for English due to the non-availability of any chunker in Bengali.

10. **Digit features**: Several digit features are defined for the Indian languages depending upon the presence and/or the number of digits and/or symbols in a token. These features are digitComma (token contains digit and comma), digitPercentage (token contains digit and percentage), digitPeriod (token contains digit and period), digitSlash (token contains digit and slash), digitHyphen (token contains digit and hyphen) and digitFour (token consists of four digits only). These features are helpful to identify miscellaneous NEs. In the present work, these features are only used for Bengali.

III. Problem Formulation

In general, feature selection problem is formulated under the single objective optimization framework. It is stated as follows: Given a set of features S and a classification quality measure P, determine the feature subset F^* such that: $P(F^*) = \max_{F \in S} P(F)$

In general the search space for this type of problems is 2^d, where d is the total number of possible features. Thus, exhaustive search strategies can not be applied in this case. Some heuristics based techniques like GA [2] can be used to search for the appropriate feature combination.

A. Multiobjective Formulation of Feature Selection Problem

The MOO can be formally stated as follows [1]. Find the vectors $\overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T$ of decision variables that simultaneously optimize the *M* objective values $\{f_1(\overline{x}), f_2(\overline{x}), \ldots, f_M(\overline{x})\}$, while satisfying the constraints, if any.

The feature selection problem under multiobjective framework is stated as follows:
$maximize\{P_1(F), P_2(F)\}$,
where $P_1, P_2 \in \{\text{recall}, \text{precision}, \text{F-measure}\}$ and $F \subset S$. We choose $P_1 =$ recall and $P_2 =$ precision.

Selection of Objectives: Performance of MOO largely depends on the choice of the objective functions which should be as contradictory as possible. In this work, we choose recall and precision as two objective functions. From the definitions, it is clear that while recall tries to increase the number of tagged entries as much as possible, precision tries to increase the number of correctly tagged entries. Thus, these two capture two different classification qualities.

IV. Proposed Approach

A multiobjective GA, along the lines of NSGA-II [3], is now used for solving the feature selection problem. Note, that although the proposed approach has some similarity in steps with NSGA-II, any other existing multiobjective GAs could have been used as the underlying MOO technique.

A. Chromosome Representation and Population Initialization

If the total number of features is F, then the length of the chromosome is F. As an example, the encoding of a particular chromosome is represented in Figure 1. Here, $F = 12$ (i.e., total 12 different features are available). The chromosome represents the use of 7 features for constructing a classifier (first, third, fourth, seventh, tenth, eleventh and twelfth features). The entries of each chromosome are randomly initialized to either 0 or 1. Here, if the i^{th} position of a chromosome is 0 then it represents that i^{th} feature does not participate in constructing the classifier. Else, if it is 1 then the i^{th} feature participates in constructing the classifier. If the population size is P then all the P number of

chromosomes of this population are initialized in the above way.

B. Fitness Computation

For the fitness computation, the following procedure is executed.
(1). Suppose, there are N number of features present in a particular chromosome (i.e., there are total N number of 1's in that chromosome).
(2). Construct a classifier with only these N features.
(3). Here, initially the training data is divided into 3 parts. The above classifier is trained using 2/3 parts of the training data with the features encoded in that particular chromosome and tested with the remaining 1/3 part.
(4). Now, the overall recall, precision and F-measure values of this classifier for the 1/3 training data are calculated.
(5). Steps 2-4 are repeated 3 times to perform 3-fold cross validation. The average recall and precision values of 3-fold cross validation of the classifier are used as the two objective functions in the proposed MOO technique. Thus, the objective functions corresponding to a particular chromosome are $f_1 = \text{recall}_{avg}$ and $f_2 = \text{precision}_{avg}$. The objective is to optimize these two objective functions using the search capability of NSGA-II.

C. Other Operators

We use crowded binary tournament selection as in NSGA-II, followed by conventional crossover and mutation. The most characteristic part of NSGA-II is its elitism operation, where the non-dominated solutions [1] among the parent and child populations are propagated to the next generation. The near-Pareto-optimal strings of the last generation provide the different solutions to the feature selection problem.

D. Selection of a Solution from the Final Pareto Optimal Front

In MOO, the algorithms produce a large number of non-dominated solutions [1] on the final Pareto optimal front. Each of these solutions provides a set of features. All the solutions are equally important from the algorithmic point of view. But, sometimes the user may require only a single solution. Consequently, in this paper a method of selecting a single solution from the set of solutions is now developed.

For every solution on the final Pareto optimal front, (i). ME based classifier is trained using the features present in that particular solution, and (ii). overall average F-measure value is computed from the 3-fold cross validation on the training data. Finally, we select the solution with maximum F-measure value. Final results on the test data are reported using the classifier corresponding to this best solution. There can be many other different approaches of selecting a solution from the final Pareto optimal front.

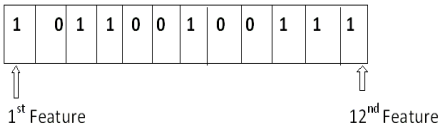

Figure 1. Chromosome representation for MOO based feature selection

V. Experimental Results and Discussions

We use the manually annotated data for Bengali. For English, we use the standard CoNLL-2003 shared task data [4]. We set the following parameter values for NSGA-II: population size=100, number of generations=50, probability of mutation=0.2 and probability of crossover=0.9.

A. Datasets for NER

Like any other Indian languages, Bengali is also resource-constrained in nature. The corpus, NE annotated corpus, POS taggers, morphological analyzers, gazetteers etc. are not readily available. In this work, we use a Bengali news corpus [5], developed from the archive of a leading Bengali newspaper available in the web. A portion of this corpus containing approximately 250K wordforms is manually annotated with a coarse-grained NE tagset of four tags namely, PER *Person name*, LOC *Location name*, ORG *Organization name* and MISC *Miscellaneous name*. The *Miscellaneous name* includes date, time, number, percentages, monetary expressions and measurement expressions. This annotation was carried out by one of the authors and verified by an expert. We also use the IJCNLP-08 NER Shared Task on South and South East Asian Languages (NERSSEAL)[1] data of around 100K wordforms that were originally tagged with a fine-grained tagset of twelve tags. An appropriate mapping is defined to make this data compatible with our manually annotated data. Out of this total 350K wordforms, approximately 37K wordforms are used as the test set in order to report the evaluation results. In order to properly denote the boundaries of NEs, four basic NE tags are further divided into the format I-TYPE (TYPE→PER/LOC/ORG/MISC) which means that the word is inside a NE of type TYPE. Only if two NEs of the same type immediately follow each other, the first word of the second NE will have tag B-TYPE to show that it starts a new NE. For example, the name *sachIna ramesha tenDUlkara* [Sachin Ramesh Tendulkar] is tagged as *sachIna*[Sachin]/I-PER *ramesha*[Ramesh]/I-PER *tenDUlkara*[Tendulkar]/I-PER. But, the pair of names *sachIna tenDUlkara rAhul drAbhiD* [Sachin Tendulkar][Rahul Dravid] are to be tagged as *sachIna*[Sachin]/I-PER *tenDUlkara*[Tendulkar]/I-PER *rAhul*[Rahul]/B-PER *drAbhiD*[Dravid]/I-PER, if they appear sequentially in the text. For English NER, we use the CoNLL-2003 shared task

[1] http://ltrc.iiit.ac.in/ner-ssea-08

Table I
STATISTICS OF THE DATASETS

Language	# words in training	#NEs in training	#words in test	#NEs in test	Unknown NEs (in %)
Bengali	312,947	37,009	37,053	4,413	35.1
English	204,566	49,029	46,666	8,112	35.56

[4]data . The training and test sets statistics are presented in Table I.

B. Results and Discussions

We consider the following set of features: (i). Context of size five (previous two and next two words) or three (previous one and next one) words, (ii). Prefixes of length upto three or four characters (3 or 4 features), (iii). Suffixes of length upto three or four characters (3 or 4 features), (iv). First word of the sentence, (v). Length of the word, (vi). Infrequent word, (vii). Position of the word in the sentence, and (viii). several digit features including digitComma, digitPercentage, digitDot, digitSlash, digitHyphen, digitFour and digitTwo.

The best solution of the proposed MOO based feature selection technique finally selects the following features for Bengali:
(i). Context of size three (previous one and next one word), (ii). Prefixes of length upto three characters, (iii). Suffixes of length upto four characters, (iv). Position of the word.

The recall, precision and F-measure values of the best individual classifier trained using the feature set identified by the best solution of the proposed MOO based technique are 70.76%, 81.88%, and 75.91%, respectively. Here, we choose the desired solution from the final Pareto optimal front to be the one having the highest F-measure. We also compare the performance of our proposed system with a ME based Bengali NER system[6], where they reported 10-fold cross validation recall, precision and F-measure values of 91.01%, 83.69% and 87.2%, respectively. However, they used a different experimental set up, more complex features (POS information etc.) and gazetteers. Features were selected manually. For the fair comparison, we evaluate this system [6] with our datasets and considering the subset of features (selected from their best feature set) that are inclusive in our system. The system yields the recall, precision and F-measure values of 69.12%, 79.69% and 74.03%, respectively. Hence, this is clearly an improvement of approximately 1.88% F-measure in our proposed system. This shows that appropriate feature selection using MOO works better compared to the heuristics based manual feature selection in a ME framework.

Thereafter, the proposed algorithm is evaluated with the standard CoNLL-2003 shared task English datasets [4]. It selects the following features for English data set:
(i). Context of size three (previous one and next one word), (ii). Chunk information of the current word, (iii). Capitalization information, (iv). Infrequent word, (v). Suffixes of length upto four characters, (vi). Prefixes of length upto four characters and (vii). digitHyphen.

Now, the ME classifier is trained using the feature set identified by the best solution of the proposed MOO based technique. Evaluation results show the overall recall, precision and F-measure values of 78.38%, 81.27%, and 79.80%, respectively. Like Bengali, the best solution is obtained from the final Pareto optimal front of the solution with highest F-measure value of the 3-fold cross validation on training data.

VI. CONCLUSION

In this work, we have posed the problem of appropriate feature selection for ME based NER as a MOO problem. We solved this crucial problem with a technique, based on a popular multiobjective evolutionary algorithm, NSGA-II. The proposed system is evaluated for Bengali, a resource poor language, as well as for English. Results are encouraging. Due to the language independent nature, our proposed algorithm is applicable for any language. Comparison with an existing ME based NER system shows that the proposed feature selection technique is more efficient than the heuristic based feature selection.

In future, we would like to add some more language independent features as well as language dependent features. In this work, we have only considered ME as the underlying classification technique. Conditional Random Fields and Support Vector Machines are some other good classifiers. Future works include selecting best feature combinations for these classifiers.

REFERENCES

[1] K. Deb, *Multi-objective Optimization Using Evolutionary Algorithms*. England: John Wiley and Sons, Ltd, 2001.

[2] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley, 1989.

[3] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 181–197, 2002.

[4] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the Conll-2003 Shared Task: Language Independent Named Entity Recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147.

[5] A. Ekbal and S. Bandyopadhyay, "A Web-based Bengali News Corpus for Named Entity Recognition," *Language Resources and Evaluation Journal*, vol. 42, no. 2, pp. 173–182, 2008.

[6] ——, "Maximum Entropy Approach for Named Entity Recognition in Bengali," in *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP)*, 2007, pp. 1–6.