Multiobjective Optimization and Unsupervised Lexical Acquisition for Named Entity Recognition and Classification

Govind
IIT Patna, India
govind.mc12@iitp.ac.in

Asif Ekbal
IIT Patna, India
asif@iitp.ac.in

Chris Biemann
TU Darmstadt, Germany
biem@cs.tu-darmstadt.de

Abstract

In this paper, we investigate the utility of unsupervised lexical acquisition techniques to improve the quality of Named Entity Recognition and Classification (NERC) for the resource poor languages. As it is not *a priori* clear which unsupervised lexical acquisition techniques are useful for a particular task or language, careful feature selection is necessary. We treat feature selection as a multiobjective optimization (MOO) problem, and develop a suitable framework that fits well with the unsupervised lexical acquisition. Our experiments show performance improvements for two unsupervised features across three languages.

1 Introduction

Named Entity Recognition and Classification (NERC) (Nadeau and Sekine, 2007) is a subtask of information extraction that has great importance in many Natural Language Processing (NLP) application areas. The objective of NERC is to find and assign tokens in unstructured text to pre-defined classes such as the names of organizations, persons, locations, miscellaneous (e.g. date-times, quantities, monetary expression etc.); and other-than-NE.

There have been a good number of research works in NERC area but these are mostly limited to the resource-rich languages such as English, the majority of the European languages and a few Asian languages like Japanese, Chinese and Korean. Research in NLP relating to the resource-scarce languages like the Indian ones is still evolving and poses some interesting problems. Some of the problems outlined previously in (Ekbal and Saha, 2011b) with reference to a specific NERC task include the absence of capitalization information, appearance of named entities (NEs) in the dictionary with other word classes, and the non-availability of various NLP resources and processing technology for non-Latin resource-poor languages.

In present work, we propose some novel methods based on the concepts of unsupervised lexical acquisition and multiobjective optimization (MOO) (Deb, 2001) for solving the problems of NERC for several languages. While we evaluate the proposed method with only three languages, the technique is generic and language-independent, and thus should adapt well to other languages or domains.

1.1 Multiobjective Optimization

The multiobjective optimization problem (MOOP) can be stated as follows: find the vectors x of decision variables that simultaneously optimize the M objective values $f_1(x), f_2(x), ..., f_M(x)$, while satisfying the constraints, if any. An important concept of MOO is that of domination. In the context of a maximization problem, a solution x_i is said to dominate x_j if $\forall k \in 1, 2, \ldots, M, f_k(x_i) \geq f_k(x_j)$ and $\exists k \in 1, 2, \ldots, M$, such that $f_k(x_i) > f_k(x_j)$.In general, a MOO algorithm usually admits a set of solutions that are not dominated by any solution encountered by it.

Genetic algorithms (GAs) are known to be more effective than classical methods such as weighted metrics, goal programming (Deb, 2001), for solving multiobjective problems primarily because of their population-based nature. Evolutionary approaches have also been used to solve few NLP problems including NERC (Ekbal and Saha, 2011a; Sofianopoulos and Tambouratzis, 2010).

1.2 Unsupervised Lexical Acquisition

One of the major problems in applying machine learning algorithms for solving information extraction problems is the availability of large annotated corpora. We explore possibilities aris-

ing from the use of unsupervised part-of-speech (PoS) induction (Biemann, 2009) and lexical expansion (Miller et al., 2012) with distributional thesauri (Riedl and Biemann, 2013). Unsupervised PoS induction is a technique that induces lexical-syntactic categories through the statistical analysis of large, raw text corpora. As shown in (Biemann et al., 2007a), using these induced categories as features results in improved accuracies for a variety of NLP tasks, including NERC.

Lexical expansion (Miller et al., 2012) is also an unsupervised technique that needs a large corpus for the induction, and is based on the computation of a distributional thesaurus (DT), see (Riedl and Biemann, 2013; Lin, 1998). While (Miller et al., 2012) used a DT for expanding lexical representations and showed performance gains in knowledge-based word sense disambiguation (WSD), the expansion technique can also be used in other text processing applications including NERC: especially for rare words and unseen instances, lexical expansion can provide a useful back-off technique as it performs a generalization of the training and test data.

2 Technical Background

Unlike supervised techniques, unsupervised PoS tagging (Christodoulopoulos et al., 2010) techniques require no pre-existing manually tagged corpus to build a tagging model and hence highly suitable for the resource poor languages.

There have been various approaches to unsupervised PoS induction. One such approach, reported in (Brown et al., 1992) is based on the class based n-gram models. In (Clark, 2003) distributional and morphological information is used for PoS induction. We use the unsupervised PoS tagging system of (Biemann, 2009) because of its availability as an open source software. We use web-based corpus of 34 million tokens for Bengali (Ekbal and Bandyopadhyay, 2008), and the datasets reported in (Biemann et al., 2007b) for Hindi and German. These datasets were used for unsupervised lexical acquisition.

A Distributional Thesaurus (DT) is an automatically computed resource that relates words according to their similarity. A DT contains, for every sufficiently frequent word, the most similar words as computed over the similarity of contexts these words appear in, which implements the distributional hypothesis (Harris, 1951). We use the scalable, open source implementation of (Riedl and Biemann, 2013), based on the MapReduce paradigm.

Feature selection is the vital task which involves selecting a subset of relevant features for building robust classifier by eliminating the redundant and irrelevant features. It therefore, reduces the time complexity of the learning algorithm and improves performance. Overall results as reported in (Biemann, 2009) suggest that unsupervised PoS tagging provides an additional word-level feature, which can be computed for any language and domain, and has been proven to be useful in domain adaptation and in situations where we have scarcity of labelled training data. In our work, we employ unsupervised PoS tags as one of the important language independent features which can benefit NERC task for various Indian languages and German.

We also investigate the use of features based on distributional similarity. We incorporate three most similar words to a particular token as three features in training and test datasets. As an example, Figure 1 shows the three most similar words for tokens in a Hindi language sentence.

Tokens from a Hindi sentence	Similar words from Distributional Thesaurus		
राम(rAma)	कृष्ण(kRRiShNa)	नारायण(nArAyaNa)	प्रसाद(parsAda)
लंका(lankA)	कोलकाता(kolakAtA)	तो(to)	घर(ghara)
में(men)	मे(me)	मंे	मेंे
सुरक्षित(surakSita)	खतरनाक(khataranAka)	उचित(uchita)	बेहतर(behatara)
स्थान(sthAna)	मिनट(minata)	ओवर(ovara)	नंबर(nambara)
पर(para)	से(se)	को(ko)	तक(taka)
बन्दिनी(bandinI)	ND	ND	ND
सीता(sitA)	संतोष(santoSha)	रमेश(ramesha)	मुकेश(mukesha)
को(ko)	उन्हें(unhen)	उसे(use)	द्वारा(dwArA)
खोज(khoja)	तलाश(talAsha)	ढूंढ(DhUnDha)	निकाल(nikAla)
सके(sake)	सकते(sakate)	सकती(sakatI)	सकता(sakatA)

Figure 1: Lexical expansion of tokens in Hindi language with ITRANS transliteration to English. Here, ND denotes the ”not defined”.

3 Named Entity Features

Following features constitute the available feature set for building the various models based on a first order Conditional Random Field (CRF) (Lafferty et al., 2001) classifier. Most of the following features do not require any language and domain specific resources or rules for their computation.
Context words: These denote the local contexts surrounding the current token.
Word suffix and prefix: Fixed length character sequences stripped from the leftmost and right

most positions of words.

First word: A binary valued feature which takes the value 1 when the current word is the first token of the sentence and 0 for the other case.

Length of the word: This feature takes the value 1 when the number of characters in a token is greater than a predetermined threshold value (here, set to 5).

Infrequent word: A binary valued features which checks whether frequency of current word in the training set exceeds a threshold value (here, set to 5).

Last word of sentence: This binary valued feature checks whether the word is the last word of a sentence or not and turn on/off accordingly.

Capitalization: This binary valued feature checks whether the word starts with a capital letter or not and takes values accordingly. This feature is used only for German.

Part-of-speech (POS) information: PoS tags of the current and/or the surrounding token(s).

Chunk information: Chunk of the current and/or surrounding tokens. This is used only for German.

Digit features: These features are defined based upon the presence and/or the number of digits and/or symbols in a token.

Unsupos: Unsupervised PoS tag as obtained from the system developed in (Biemann, 2009) is used as a feature.

Unsupervised DT features: Three most similar word from the DT for each token in training and test dataset.

4 Feature Selection using MOO

In this section we formulate feature selection as an optimization problem that involves choosing an relevant feature subset for NERC. Mutiobjective optimization (MOO) can be effective for solving the problem of feature selection. Here we develop a feature selection method based on a popular MOO based technique, namely non-dominated sorting genetic algorithm (NSGA-II) (Deb, 2001). In order to implement our MOO-based feature selection we make use of NSGA–II (Deb et al., 2002). As a supervised learner we used Conditional Random Field (CRF) (Lafferty et al., 2001), and carried out experiments using its CRF++[1] implementation.

[1]CRF++:Yet another CRF toolkit http://crfpp.googlecode.com/svn/trunk/doc/index.html

4.1 Formulation of feature selection problem

Let us denote the N number of available features by $f_1, f_2, \ldots, f_N$ and suppose that the set of all features be denoted by $F = f_i : i = 1, 2 \ldots N$. Then the problem of feature selection can be stated as follows: Find a set of features G that will optimize a function $O(F)$ such that: $G \subseteq F$. Here, O is a measure of classification efficiency for the classifier trained using the features set G. The feature selection problem can be formulated under the MOO framework as: Find a set of features G such that maximize $[O_1(G), O_2(G)]$, where $O_1, O_2 \in$ recall, precision, F-measure, $-$(feature count). Here, we choose $O_1 =$ F-measure and $O_2 = -$(feature count)

4.2 Problem encoding

Let the total number of features is N and size of the population is P. The length of the chromosome is determined from the number of available features and hence its size is N. If the i^{th} position of chromosome is 0, then it represents that i^{th} feature does not participate in feature template set for construction of CRF-based classifier and opposite in case of 1.All the P number of chromosomes of initial population are initialized with a random sequence of 0 and 1.

4.3 Fitness Computation

For the fitness computation, the following steps are executed.

- There are $|G|$ number of features present in a particular chromosome (i.e., total $|G|$ number of 1's are there in the chromosome).
- Build a CRF classifier with only these $|G|$ features. We perform 3-fold cross validation and compute the F-measure value.
- Our objective is to maximize F-measure and minimize the feature count. NSGA-II (Deb, 2001) is used for optimization process using these two objective functions.

4.4 Selecting a single solution

The MOO based feature selection technique produces a set of solutions on the Pareto front. All these are best in their own and incomparable on the basis of aforementioned two objectives collectively. But in order to report the final results we build a CRF classifier with that particular feature combination that yields the highest F_1 measure

Language	Set	#tokens
Bengali	Training	328,064
	Test	34,200
Hindi	Training	462,120
	Test	60,810
German	Training	220,187
	Test	54,711

Table 1: Statistics of annotated training and test datasets

value among all the solutions of the final population.

5 Datasets and Experimental Setup

We use the web-based Bengali news corpus for our NERC experiments (Ekbal and Bandyopadhyay, 2008) in Bengali. A part of this corpus was manually annotated with four MUC NE categories, 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 (Ekbal and Bandyopadhyay, 2008). In addition we also use the NER on South and South East Asian Languages (NERSSEAL)[2] Shared Task datasets of Bengali after mapping the fine-grained tagset to our coarse-grained form. For German we use the datasets obtained from datasets from the CoNLL 2003 challenge (Tjong Kim Sang and De Meulder, 2003). Statistics of training and test datasets are reported in Table 5.

The feature selection algorithm is run three times with different set of available features. Specifically we design three experiments, one with only basic lexical features, the second with lexical features along with unsupos tag, and the third experiment with three features from DT in addition to unsupos tag and lexical features. In order to properly denote the boundaries of a NE, we follow the IOB2 encoding scheme of the CoNLL-2003 shared task[3].

6 Evaluation of NERC for the Indian Languages

In this section we present the results along with the analysis for NERC on two Indian languages, namely Hindi and Bengali. For each of the languages, we extracted the features as defined in Section 3 including the token itself. We also incorporate features from the immediate contextual tokens (i.e. preceding token and following token). So the available number of features becomes equal to 27*3=81, and our goal is to find the best feature subset from this available features set which optimizes our objective functions.

In all the experiments, we set the following parameter values for NSGA-II algorithm: population size = 32, number of generations = 50, probability of crossover = 0.8 and probability of mutation = 0.0125. The values of these parameters were determined using a held-out dataset (created by taking a portion from the training dataset).

Table 2 depicts the detailed evaluation results for NERC task on Hindi dataset. Results show that without using any lexical acquisition feature, we obtain the best results with a set of 41 features represented in the final population of MOO based feature selection algorithm. These results are considered as baseline for our further experiments on NERC.

In the next experiment we incorporate unsupervised PoS tag in the available set of features and apply the algorithm. It is observed that including unsuporvised PoS, recall increases but at the cost of precision. However this causes a small improvement in F_1 measure. This improvement is attributed because of the incorporation of unsupervised PoS tags for training the classifier. Thus, unsupos features generalize over the vocabulary, and subsume part of the lower-level features. We observe that the presence of the unsupervised PoS tag reduces the optimized feature set from 41 down to 25 features while at the same time improving in F_1.

Features	Tag	Precision	Recall	F_1	FC
Syntactic features only(Baseline)	LOC	82.71%	47.97%	60.72	
	MISC	83.37%	74.22%	78.53	
	ORG	52.63%	29.85%	38.10	
	PER	70.72%	29.15%	41.29	
	Overall	80.15%	52.19%	63.22	41
Syntactic + Unsupos features	LOC	82.20%	49.24%	61.59	
	MISC	83.00%	76.78%	79.77	
	ORG	62.50%	29.85%	40.40	
	PER	67.42%	32.14%	43.53	
	Overall	79.22%	54.45%	64.54	25
Syntactic + Unsupos + DT features	LOC	72.88%	63.39%	67.81	
	MISC	80.08%	82.76%	81.40	
	ORG	55.13%	56.95%	56.03	
	PER	63.87%	43.96%	52.08	
	Overall	73.26%	66.44%	69.68	32

Table 2: NERC performance for Hindi data–set, No. of generations=50, Size of population=32, FC= Feature Count

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

[3]http://www.cnts.ua.ac.be/conll2003/ner/

Next, we explore DT features by adding them to the pool of features. Algorithm for feature selection is again run with these additional features, and the results are reported in Table 2. With these DT features, recall goes up rapidly, but at the cost of precision. Again, we see a drop in precision, yet a relative recall increase of 22% causes the F-measure to increase by 5 percentage points.

The feature selection algorithm determines 32 features to be most relevant for the task. This feature combination includes several lexical expansion features that include the first two expansions of the preceding token and all the three expansions of the current token. It seems that the CRF profits rather from the expansion of contexts than from the expansions themselves. These DT in combination with unsupos features improve a total of 6 points F-measure over the baseline.

Thereafter we experiment with the Bengali datasets and its results are shown in Table 3. It shows how the performance can be improved with the use of unsupervised PoS tag and DT features. Although there is not much difference in the scores between the results obtained in the first two experiments, there is substantial reduction in the feature count. Again, recall is increased at cost of precision, as unsupervised features add coverage, but also noise at subsuming lower-level features. The performance obtained using unsupervised features are quite encouraging and comparable to the existing works (for both Hindi and Bengali). This also open a new direction for performing similar kinds of works in the resource-poor languages.

Feature set	F1–measure	FC
No unsupervised PoS Tag and DT features	72.44	30
With unsupervised PoS Tag	72.72	14
With unsupervised PoS Tag and DT features	73.50	21

Table 3: NERC performance for Bengali data–set, No. of Generations=50, Size of population=52

7 Experiments for NERC on German

In this section we report on our experiments for NERC in German language. For each token we extract twelve features including lexical features, unsupervised PoS tag and three most similar words from DT. We compute the values of these features at the preceding and succeeding tokens. We use the default parameter values of CRF and set of the parameters of NSGA-II as mentioned in the previous section.

Table 4 depicts the performance for NERC task on German dataset for the baseline model, which is constructed without using any unsupervised lexical acquisition features and for models which are constructed after incorporation of lexical acquisition features. For the baseline model, feature selection algorithm selects the solution representing 20 features for training CRF classifier. We obtain precision, recall and F_1 measure of 80.43%, 64.11% and 71.35%, respectively.

Features	Tag	Precision	Recall	F_1	FC
Syntactic features only (Baseline)	LOC	77.36%	67.94%	72.34	
	MISC	80.52%	30.10%	43.82	
	ORG	73.47%	59.76%	65.91	
	PER	86.83%	68.68%	76.70	
	Overall	80.43%	64.11%	71.35	20
Syntactic + DT features	LOC	81.40%	69.93%	75.23	
	MISC	79.22%	29.61%	43.11	
	ORG	74.50%	57.02%	64.60	
	PER	88.31%	72.40%	79.56	
	Overall	82.89%	65.72%	73.31	19
Syntactic + DT + Unsupervised PoS features	LOC	84.87%	72.60%	78.26	
	MISC	79.75%	30.58%	44.21	
	ORG	74.64%	61.99%	67.73	
	PER	93.07%	82.15%	87.27	
	Overall	86.21%	71.52%	78.18	21

Table 4: NERC performance for German data–set, No. of Generations=50, Size of population=52

In the next experiment on German dataset with DT features incorporated, we obtain improvements in both precision and recall, which causes substantial improvement in F_1. Lexical expansion reduces the chances of unseen instances during testing, which results in higher F_1 measure with one less number of features. The third experiment includes three DT features as well as the unsupervised PoS tag in the available set of features for feature selection. It is evident that we obtain significant improvements for both recall and precision, which in turn causing higher F_1 measure. Over the baseline we obtain an improvement of 6.83 in F_1 measure with the 21 most relevant features. The best solution includes all the four unsupervised lexical acquisition features.

8 Conclusion

In this present work, we proposed a unsupervised lexical acquisition and MOO-based technique for building NERC systems. It has been consistently observed that incorporation of unsupervised lexical acquisition features and using

MOO-based feature selection result in significant improvement in NERC performance for a variety of languages. The performance of our models compares favourably with other works in the literature (Tjong Kim Sang and De Meulder, 2003). Also, we present a framework that can easily be transferred to the other languages and applications.

In future we would like to include more language independent features. Rather than selecting a single best-fitting feature set from best population produced by MOO algorithm, we would like to combine an ensemble of several classification systems based on different feature sets and/or different classification techniques.

References

Chris Biemann, Claudio Giuliano, and Alfio Gliozzo. 2007a. Unsupervised part-of-speech tagging supporting supervised methods. In *Proceedings of RANLP*, volume 7.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007b. The Leipzig Corpora Collection - Monolingual corpora of standard size. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.

Chris Biemann. 2009. Unsupervised part-of-speech tagging in the large. *Research on Language and Computation*, 7(2-4):101–135.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of EMNLP*, pages 575–584.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *In proceedings of European chapter of the Association for Computational Linguistics (EACL-03)*, pages 59–66.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):181–197.

Kalyanmoy Deb. 2001. *Multi-objective Optimization Using Evolutionary Algorithms*. John Wiley and Sons, Ltd, England.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. A web-based Bengali news corpus for named entity recognition. *Language Resources and Evaluation*, 42(2):173–182.

Asif Ekbal and Sriparna Saha. 2011a. Multiobjective Optimization for Classifier Ensemble and Feature Selection: An Application to Named Entity Recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 8.

Asif Ekbal and Sriparna Saha. 2011b. Weighted vote-based classifier ensemble for named entity recognition: A genetic algorithm-based approach. *ACM Trans. Asian Lang. Inf. Process.*, 10(2):9.

Zellig S. Harris. 1951. *Methods in Structural Linguistics*. University of Chicago Press, Chicago.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 768–774, Stroudsburg, USA. ACM Press.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. *Proceedings of COLING–12, Mumbai, India*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Martin Riedl and Chris Biemann. 2013. Scaling to large3 data: An efficient and effective method to compute distributional thesauri. In *EMNLP*, pages 884–890.

Sokratis Sofianopoulos and George Tambouratzis. 2010. Multi-objective optimisation of real-valued parameters of a hybrid mt system using genetic algorithms. *Pattern Recognition Letters*, 31(12):1672–1682.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.