Single and Multi-objective Optimization for Feature Selection in Anaphora Resolution

Sriparna Saha [1] **Asif Ekbal** [1] **Olga Uryupina** [2] **Massimo Poesio** [3,2]
[1] Department of Computer Science and Engineering, IIT Patna, India,
{sriparna,asif}@iitp.ac.in
[2] University of Trento, Center for Mind/Brain Sciences, uryupina@unitn.it
[3] University of Essex, Language and Computation Group, poesio@essex.ac.uk

Abstract

There is no generally accepted metric for measuring the performance of anaphora resolution systems, and the existing metrics—MUC, B^3, CEAF, Blanc, among others—tend to reward significantly different behaviors. Systems optimized according to one metric tend to perform poorly with respect to other ones, making it very difficult to compare anaphora resolution systems, as clearly shown by the results of the SEMEVAL 2010 Multilingual Coreference task. One solution would be to find a single completely satisfactory metric, but it's not clear whether this is possible and at any rate it is not going to happen any time soon. An alternative is to optimize models according to multiple metrics simultaneously. In this paper, we show, first of all, that this is possible to develop such models using Multi-objective Optimization (MOO) techniques based on Genetic Algorithms. Secondly, we show that optimizing according to multiple metrics simultaneously may result in better results with respect to each individual metric than optimizing according to that metric only.

1 Introduction

In anaphora resolution,[1] as in other HLT tasks, optimization to a metric is essential to achieve good performance (Hoste, 2005; Uryupina, 2010). However, many evaluation metrics have been proposed for anaphora resolution, each capturing what seems to be a key intuition about the task: from MUC (Vilain et al., 1995) to B^3 (Bagga and Baldwin, 1998), from the ACE metric (Doddington et al., 2004) to CEAF (Luo, 2005) to BLANC (Recasens and Hovy, 2011). And unlike in other areas of HLT, none has really taken over. This would not matter so much if those metrics were to reward the same systems; but in fact, as dramatically demonstrated by the results of the Coreference Task at SEMEVAL 2010 (Recasens et al., 2010), the opposite is true—almost every system could come on top depending on which metric was chosen.

It seems unlikely that the field will converge on a single metric any time soon. Given that many of the proposed metrics do capture what would seem to be plausible intuitions, it would seem desirable to develop methods to optimize systems according to more than one metric at once—in particular, according to at least one metric of what we might call the 'link-based' cluster of metrics (e.g., MUC) and at least one of what we will call the 'entity-based' cluster (e.g., CEAF).

As it happens, techniques for doing just that have been developed in the area of Genetic Algorithms: so-called **multi-objective optimization** (MOO) (Deb, 2001) techniques. In this paper, we will show how these techniques can be used to optimize anaphora resolution models (we focused for the time being on feature selection) by looking for a solution in the space defined by a multiplicity of metrics (we used MUC and CEAF (in two variants) as the optimization functions). Perhaps the most interesting result of this work is the finding that by working in such a multi-metric space it is possible to find solutions that are better with respect to an individual metric than when trying to optimize for that metric alone—which arguably suggests that indeed both families of metrics capture some fundamental intuition about anaphora, and taking into account both intuitions we avoid local optima.

The structure of the paper is as follows. We first review the literature on using genetic algorithms for both single function and multi function opti-

[1] We use the term 'anaphora resolution' to refer to the task perhaps most commonly referred to as 'coreference resolution,' which many including us find a misnomer. For the purposes of the present paper the two terms could be seen as interchangeable.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 93–101,
Chiang Mai, Thailand, November 8 – 13, 2011.

mization. Next, we discuss the particular method of multi-objective optimization we used in this paper, Non-Dominated Sorting Genetic Algorithm II (Deb et al., 2002). After that we discuss how the method was used, and present our results. We then compare our work with other approaches to optimization for anaphora found in the literature.

2 Background: Optimizing for Anaphora Resolution

A great number of statistical approaches to anaphora resolution have been proposed in the past ten years. These approaches differ with respect to their underlying models (e.g., mention pair model (Soon et al., 2001) vs. tournament model (Iida et al., 2003; Yang et al., 2005), vs. entity-model (Luo et al., 2004)), machine learners (e.g., decision trees vs. maximum entropy vs. SVMs vs. TiMBL) and their parameters, and with respect to feature sets used. There have been, however, only few attempts at explicit optimization of these aspects, and in those few cases, optimization tends to be done by hand.

An early step in this direction was the work by Ng and Cardie (2002), who developed a rich feature set including 53 features, but reported no significant improvement over their baseline when all these features were used with the MUC6 and MUC7 corpora. They then proceeded to manually select a subset of features that did yield better results for the MUC-6/7 datasets. A much larger scale and very systematic effort of manual feature selection over the same dataset was carried out by Uryupina (2007), who evaluated over 600 features.

The first systematic attempt at automatic optimization of anaphora resolution we are aware of was carried out by Hoste (2005), who investigated the possibility of using genetic algorithms for automatic optimization of both feature selection and of learning parameters, also considering two different machine learners, TiMBL and Ripper. Her results suggest that such techniques yield improvements on the MUC-6/7 datasets. Recasens and Hovy (2009) carried out an investigation of feature selection for Spanish using the ANCORA corpus.

These approaches focused on a single metric only; the one proposal simultaneously to consider multiple metrics, Zhao and Ng (2010) still optimized for each metric individually.

The effect of optimization on anaphora resolution was dramatically demonstrated by Uryupina's contribution to SEMEVAL 2010 Multilingual Coreference Task (Uryupina, 2010). Uryupina directly optimizes two parameters of her system: the choice of a model (mention-pair vs. ILP with various constraints) and the definition of mention types for training separate classifiers. The optimization is done on the development data in a brute-force fashion, in order to maximize the performance according to a pre-defined metric (MUC, CEAF or BLANC). The results on the SEMEVAL-10 dataset clearly show that existing metrics of coreference rely on different intuitions and therefore a system, optimized for a particular metric, might show inferior results for the other ones. For example, the reported BLANC difference between the runs optimized for BLANC and CEAF is around 10 percentage points.

This highlights the importance of the multi-objective optimization (MOO) for coreference, that suggests a family of systems, showing reliable performance according to all the desired metrics. A form of MOO was applied to coreference by Munson et al. (2005). Their general conclusion was negative, stating that "ensemble selection seems too unreliable for use in NLP", but they did see some improvements for coreference.

3 Optimization with Genetic Algorithms

In this section, we review optimization techniques using genetic algorithms (GAs) (Goldberg, 1989). We first discuss single objective optimization, that can optimize according to a single objective function, and then multi-objective optimization (MOO), that can optimize more than one objective function, in particular, a popular MOO technique named Non-dominated Sorting Genetic Algorithm (NSGA)-II (Deb et al., 2002).

3.1 Genetic Algorithms

Genetic algorithms (GAs) (Goldberg, 1989) are randomized search and optimization techniques guided by the principles of evolution and natural genetics. In GAs the parameters of the search space are encoded in the form of strings (called *chromosomes*). A collection of such strings is called a *population*. Initially, a random population is created, which represents different points in the search space. An *objective* or *fitness* function is associated with each string that represents the degree of *goodness* of the string. Based on the principle of survival of the fittest, a few of the strings are selected and each is assigned a number of copies that go into the mating pool. Biologically inspired op-

erators like *crossover* and *mutation* are applied on these strings to yield a new generation of strings. The processes of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied.

3.2 Multi-objective Optimization

Multi-objective optimization (MOO) can be formally stated as follows (Deb, 2001). 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.

An important concept in MOO is that of **domination**. In the context of a maximization problem, a solution $\overline{x_i}$ is said to dominate $\overline{x_j}$ if $\forall k \in 1, 2, \ldots, M,\ f_k(\overline{x_i}) \geq f_k(\overline{x_j})$ and $\exists k \in 1, 2, \ldots, M,$ such that $f_k(\overline{x_i}) > f_k(\overline{x_j})$.

Among a set of solutions P, the nondominated set of solutions P' are those that are not dominated by any member of the set P. The nondominated set of the entire search space *S* is called the **globally Pareto-optimal set**. In general, a MOO algorithm usually admits a set of solutions not dominated by any solution encountered by it.

3.3 Nondominated Sorting Genetic Algorithm-II (NSGA-II)

Genetic algorithms (GAs) are known to be more effective than classical methods such as weighted metrics, goal programming (Deb, 2001), for solving MOO primarily because of their population-based nature. A particularly popular GA of this type is NSGA-II (Deb et al., 2002).

In NSGA-II, initially a random parent population P_0 is created and the population is sorted based on the *partial order* defined by the non-domination relation. This results in a sequence of nondominated **fronts**. Each solution is assigned a fitness value which is equal to its non-domination level in the partial order. A child population Q_0 of size N is then created from the parent population P_0 by using binary tournament selection, recombination, and mutation operators. In general, in the t^{th} iteration, a combined population $R_t = P_t + Q_t$ is formed. The size of R_t is $2N$, as the size of both P_t and Q_t is N. All the solutions of R_t are sorted according to non-domination. If the total number of solutions belonging to the best nondominated set F_1 is smaller than N, then F_1 is totally included in $P_{(t+1)}$. The remaining members of the population $P_{(t+1)}$ are chosen from the subsequent nondominated fronts in the order of their ranking. To choose exactly N solutions, the solutions of the last included front are sorted using the crowded comparison operator (Deb et al., 2002) and the best among them (i.e., those with lower crowding distance) are selected to fill in the available slots in $P_{(t+1)}$. The new population $P_{(t+1)}$ is then used for selection, crossover and mutation to create a population $Q_{(t+1)}$ of size N.

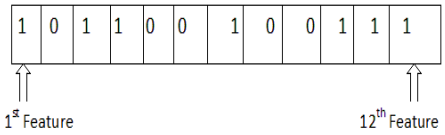

Figure 1: Chromosome representation for GA based feature selection

4 Two Algorithms for Feature Selection in Anaphora Resolution

Below we discuss how single and multi-objective optimization techniques can be used feature selection in the anaphora resolution task.

4.1 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.

4.2 Fitness Computation

For 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 the coreference resolution system (i.e., BART) with only these N features.
3. This coreference system is evaluated on the development data. The recall, precision and F-measure values of three metrics are calculated.

In case of single objective optimization (SOO), the objective function corresponding to a particular chromosome is the F-measure value of a single metric. This objective function is optimized using the search capability of GA. For MOO, the objective functions corresponding to a particular chromosome are F_{MUC} (for the MUC metric), F_{ϕ_3} (for CEAF using the ϕ_3 entity alignment function (Luo, 2005)) and F_{ϕ_4} (for CEAF using the ϕ_4 entity alignment function). These three objective functions are simultaneously optimized using the search capability of NSGA-II.

4.3 Genetic Operators

In case of SOO, a single point crossover operation is used with a user defined crossover probability, μ_c. A mutation operator is applied to each entry of the chromosome with a mutation probability, μ_m, where the entry is randomly replaced by either 0 or 1. In this approach, the processes of fitness computation, selection, crossover, and mutation are executed for a maximum number of generations. The best string seen up to the last generation provides the solution to the above feature selection problem. Elitism has been implemented at each generation by preserving the best string seen upto that generation in a location outside the population. Thus on termination, this location contains the best feature combination.

We use crowded binary tournament selection as in NSGA-II, followed by conventional crossover and mutation for the MOO based feature selection. The most characteristic part of NSGA-II is its elitism operation, where the non-dominated solutions (Deb, 2001) 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.

5 Methods

5.1 The BART System

For our experiments, we use BART (Versley et al., 2008), a modular toolkit for anaphora resolution that supports state-of-the-art statistical approaches to the task and enables efficient feature engineering. BART implements different models of anaphora resolution (mention-pair and entity-mention; best-first vs. ranking), has interfaces to different machine learners (MaxEnt, SVM, decision trees) and provides a large set of linguistically motivated features, along with the possibility to design new ones. It is thus ideally suited for experimenting with optimization and feature selection.

In this study, we specifically focus on feature selection.[2] The complete list of features currently implemented in BART is listed in Table 1; all were considered in the present experiments. We used a simple mention-pair model without ranking as in (Soon et al., 2001). In the mention-pair model, anaphora resolution is recast as a binary classification problem. Each classification instance consists of two mentions, i.e. an anaphor M_j and its potential antecedent M_i ($i < j$). Instances are modeled as feature vectors (cf. Table 1) and are handed over to a binary classifier that decides, whether the anaphor and its candidate antecedent are mentions of the same entity or not. All the feature values are computed automatically.

We train a maximum entropy classifier and follow the approach of (Soon et al., 2001) to partition mentions into coreference sets given the classifier's decisions.

5.2 The Data Sets

We evaluated our approach on the ACE-02 dataset, which is divided in three subsets: bnews, npaper, and nwire. We provide results for both gold (hand-annotated) versions of the datasets (gbnews, gnpaper, gnwire) and system mentions extracted with CARAFE[3] (cbnews, cnpaper, cnwire).

Table 2 compares the performance level obtained using all the features in Table 1 with that of a loose re-implementation of the system proposed by Soon et al. (2001), commonly used as baseline and relying only on very shallow information. Our reimplementation of the Soon et al. model uses only a subset of features: those marked with an asterisk in Table 1. We also provide in Table 2 typical state-of-the-art figures on the ACE-02 dataset, as presented in an overview by Poon and Domin-

[2] The choice of the best model and the best machine learner, along with its parameters, is the main direction of our future work.

[3] `http://sourceforge.net/projects/carafe`

Table 1: Features used by BART: each feature describes a pair of mentions $\{M_i, M_j\},\ i < j$, where M_i is a candidate antecedent and M_j is a candidate anaphor

Mention types and subtypes	
MentionType*	relevant types of M_i and M_j, as identified in Soon et al.
MentionType_Ante_Salient	M_i is demonstrative; M_i is an NE
MentionType_Ante_Extra	M_i is a pronoun
MentionType_Ana	M_j is a definite, demonstrative or indefinite NP, or pronoun of a specific type
MentionType2	relevant types of M_i and M_j, as identified in Soon et al.
MentionType_Salience	combination of $MentionType$ and $MentionType\_Ana$
FirstSecondPerson	M_i is a pronoun of the 1st/second person, same for M_j
PronounLeftRight	4 possible values for $< M_i$ is a pronoun$> * < M_j$ is a pronoun $>$
PronounWordForm	lemma for M_i if it's a pronoun; same for M_j
SemClassValue	semantic class of M_i, and M_j and the pair
BothLocation	both M_i and M_j are locations or geo-political
Agreement	
GenderAgree*	M_i and M_j agree in gender
NumberAgree*	M_i and M_j agree in number
AnimacyAgree*	M_i and M_j agree in animacy
Aliasing	
Alias*	heuristical NE-matching
BetterNames	heuristical matching for personal names
Syntax	
Appositive*	M_i and M_j are in an apposition
Appositive2	M_i and M_j are adjacent
Coordination	M_i is a coordination; same for M_j
HeadPartOfSpeech	POS of M_i's head; same for M_j and the pair
SynPos	depth of M_i's node in the parse tree
Attributes	M_i and M_j have incompatible premodifiers
Relations	M_i and M_j have incompatible postmodifiers
Matching	
StringMatch*	M_i and M_j have the same surface form after stripping off the determiners
NonPro_StringMatch	both M_i and M_j are non-pronominal and $Stringmatch(M_i, M_j) == 1$
Pro_StringMatch	both M_i and M_j are pronominal and $Stringmatch(M_i, M_j) == 1$
NE_StringMatch	both M_i and M_j are NE and $Stringmatch(M_i, M_j) == 1$
HeadMatch	M_i and M_j have the same head
MinSame	M_i and M_j have the same minimal span
LeftRightMatch	M_j is a prefix or suffix substring of M_i or vice versa
StringMatchExtra	extra string-macthing for bare plurals
StringKernel	approximate matching
Salience	
First_Mention	M_i is the first mention in its sentence
CorefChain	Size of the coreference chain suggested for M_i so far (with a threshold)
NonProSalience	for non-pronominal M_i, number of preceding mentions with the same head lemma
Web	
Wiki	M_i and M_j have the same wikipedia entry
Yago	M_i and M_j are linked in Yago via `means` or `typeof` relation
WebPatterns	specific contexts for co-reference extracted from the web
Proximity	
DistanceMarkable	distance in mentions between M_i and M_j
DistanceSentenceInt*	distance in sentences between M_i and M_j
DistanceSentence	log-distance in sentences between M_i and M_j
DistanceSentence2	log-distance in sentences between M_i and M_j, different formula
DistDiscrete	distance in sentences between M_i and M_j discretized into {0,1,>=2}
Miscellaneous	
Speech	M_i is in quoted speech; same for M_j and the pair

Table 2: Baseline performance on the ACE-02 dataset

	gold mentions								
	gbnews			gnpaper			gnwire		
	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}
following Soon et al. (2001)	71.6	67.2	69.6	67.8	62.6	67.5	66.7	67.9	69.7
All features (Table 1)	75.8	70.6	74.4	72.5	64.7	67.0	71.2	70.3	72.2
state-of-the-art	65-69	-	-	70-72	-	-	54-67	-	-
	system mentions								
	cbnews			cnpaper			cnwire		
	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}
following Soon et al. (2001)	61.3	56.7	55.9	63.3	57.6	54.0	60.8	58.2	57.0
All features (Table 1)	62.3	57.9	57.5	65.5	55.9	52.7	60.6	56.8	55.6

Table 3: Feature vectors identified via single-objective optimization.

DataSet	Metric opt.	Features Selected	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}
gbnews	MUC	00100110110111100111111000111001001111111010	76.8	71.5	74.5
	ϕ_3,ϕ_4	10011000111010110000101101010011011011000001	76.7	71.8†	74.9†
gnpaper	MUC	10000001001111110101011101110000101010100111	74.6	67.1	70.1†
	ϕ_3	10101001100100110100100000010100010001101100	72.2	67.6	69.1
	ϕ_4	11101001100100110100111000100101110010001100	71.4	65.2	70.3
gnwire	MUC	10111011011111110010101010011010011011001011	74.0†	70.3†	73.1†
	ϕ_3	11011011100001000011110110101111011110001101	71.4	72.3	73.6
	ϕ_4	11101001100100110100111000100101110010001100	71.7	72.1	74.4
cbnews	MUC,ϕ_3	11111001100101000011011100101101101111001100	64.6	59.7	58.4
	ϕ_4	11111001100001000011110100101111101110001101	63.6	59.6	58.8
cnpaper	MUC,ϕ_3	01000100100101011001000010111100101100001000	66.5	59.7†	54.7†
	ϕ_4	10100101101011100011111110010100100010010011	66.2	59.1	55.6†
cnwire	MUC	00101111101110101001100000010101001011001001	63.8	60.0	58.1
	ϕ_3,ϕ_4	00011000101110100010000010011000100110000100	63.4	61.2	58.4

gos (2008). The results clearly show that although even larger sets of features have been proposed (Uryupina, 2007; Bengtson and Roth, 2008), the set of features already included in BART is sufficient to achieve results well above the state of the art on the dataset we used.

The results in Table 2 also confirm the intuition that, contrary to what is suggested by some of the early papers (Soon et al., 2001; Ng and Cardie, 2002) working on smaller datasets, linguistic factors do play a crucial role in anaphora resolution and therefore rich feature sets may lead to performance improvements once larger datasets are considered (a similar result was also obtained by Bengtson and Roth (2008)). Such improvements, however, come at high costs, as both using larger datasets and larger sets of features learning a model becomes slower and requires much more memory.

This suggests that automatic feature selection may be essential not just to improve performance but also to be able to train a model—i.e., that an efficient coreference resolution system should combine rich linguistic feature sets with automatic feature selection mechanisms.

5.3 Genetic Algorithm Parameter Setting

We set the following parameter values for both single (i.e., GA) and MOO (i.e., NSGA-II): population size=20, number of generations=30, probability of mutation $\mu_m = 0.2$ and probability of crossover $\mu_c = 0.9$. Both approaches are executed on devel-

opment data to determine the optimal feature vector(s). Final results are reported on the test data. It is to be noted that GA is a stochastic approach and outputs different results for trials with different seeds and initial populations. Initial seeds and population are chosen randomly. Thus for each data set we executed the proposed single and multi objective based approaches 3 times. Finally, we report the maximum values of these 3 runs.

6 Results

6.1 Single Objective Optimization

Single objective GA based feature selection was executed on the six data sets to determine the appropriate set of features. For each data set three sets of experiments were carried out by optimizing the F-measure values of the three different evaluation metrics. The binary-valued feature vectors identified by the single objective GA based feature selection technique for the six data sets and the corresponding F-measure values are shown in Table 3. The order of the features in the vector corresponds to their order in Table 1; the values of 0's and 1's represent the absence and presence of the corresponding features. Significant improvements over the classifier based on all the features are indicated with † (sign test, $p < 0.05$).

These results show that for all the datasets, the proposed single objective GA-based feature selection technique performs better than the baseline approach of using all features. Moreover, the results show that the technique based on SOO (i.e., conventional GA-based method) with different objective functions provides different evaluation figures. Thus, it is meaningful to optimize each objective function separately.

It is also evident from Table 3 that the optimal feature set obtained by optimizing a single objective function may not be optimal with respect to another objective function. Thus, it is not possible to come up with common patterns in the set of optimal features. For example, in case of *gbnews*, the F-measure value of the first metric, i.e. of MUC corresponding to the optimal feature vector optimizing second metric, i.e. ϕ_3 is 76.7. This is obviously less than the evaluation figure obtained by optimizing the first metric.

6.2 Multi-objective Optimization

Thereafter we apply our proposed MOO based feature selection technique on the six data sets. The MOO approach provides a set of non-dominated solutions on the final Pareto optimal front. All the solutions are equally important from the algorithmic point of view. In Table 4, we show the final solutions obtained by the MOO based approach for all the data sets. Significant improvements over the classifier based on all the features are indicated with † (sign test, $p < 0.05$).

The results in Table 4 indicate that the MOO based technique achieves higher performance than the single objective GA based approach. For the *gbnews* data set, MOO achieves 0.6, 0.3 and 0.8 F-measure points increments for three metrics over the single objective GA based technique. For the *gnpaper* data set, there are increments of 2.5 F-measure points on second metric and 1.0 F-measure point on third metric over the corresponding single objective GA based technique. Similarly, for all other datasets the MOO based approach attains superior performance over the SOO-based approach.

7 Comparison with Related Work

As discussed in Section 2 most work on optimization in anaphora resolution relies on manual optimization; the one significant exception is the work of Hoste (2005).

There are two major differences between the approach of Hoste (2005) and that followed in our study. First, the scope of (Hoste, 2005) is restricted to *single-objective* optimization. As we saw above, this might provide unstable solutions, that are too tailored to a particular scoring metric. Second, the feature set of Hoste (2005) is relatively small and therefore does not provide an efficient test-bed for a feature selection approach. Not surprising, parameter optimization shows a more consistent effect on the overall performance than feature selection in (Hoste, 2005)'s experiments.

8 Discussion and Conclusions

In this paper we showed that it may not be necessary to choose one among the existing metrics for anaphora resolution—in fact, that developing systems attempting to optimize according to a combination of them may lead to better results.

In subsequent work, we plan to expand the optimization technique to consider also learning parameters optimization, classifier selection, and learning model selection.

Table 4: Feature vectors identified by the MOO based approach.

DataSet	Features	F_{MUC}	F_{ϕ_3}	F_{ϕ_4}
gbnews	00011110110111110011101101011111101110010101	77.20	71.50	75.70
	00110100110111110010101101011111100110010101	77.20	72.00	75.50†
	00111110111111110011101101011111100110010101	77.00	72.10	75.10
	00111010110111110011101101011111100110000100	77.30	71.50	74.40
	00111100100111110010101101011111101100010101	77.40	71.30	74.70†
gnpaper	01001011101010110111111000010010101011000010	73.90	70.10†	71.10†
	10000001001111110101011101110000101010100111	74.60	67.10	70.10
	01001010101010110111111000110110101011000010	73.80	70.10	71.30
	11011111100011110011110011110110100111000010	74.30	67.90	70.00
	11001010101011100111111000110010101011000010	74.10	69.30	70.70
	10011110101011110011110000110111101011000010	74.40	67.20	69.60
	11001110101011100111111010111110100011000010	74.40	67.50	69.10
	10001110101011100111111000110111101011100010	74.50†	66.90	69.40
	01001110101010110111111010011100100011000010	74.20	68.80	70.90
gnwire	10101100111011100110101001011011100110000100	74.90†	72.30†	73.80
	10101100101011100110101011101010100110000100	73.80	73.10†	74.70
	10101100101011100100101001011011100010000100	74.80†	73.40†	74.00†
	10001100111011100110101011101010100110000100	74.30†	72.80†	74.60†
	10001100101011100110101001011011100010000100	74.80†	73.30†	74.10
cbnews	01011010011111001111100110011110001110001011	64.80†	60.30	59.10†
	00111010111111001111100100011010000110011011	65.10	60.60	58.90
cnpaper	10011010110111110001111000110110001111001000	67.40	60.00†	55.00
	11011000110011110000110000111110001011111010	66.40	58.20†	56.10†
	00011111110010010001011110110111000011001001	66.20	59.60	55.20†
	10011011010011110001110010110110000011011000	66.60	58.30†	55.90†
	11011000110011110010110000111110101011111010	66.70	59.40†	55.70†
cnwire	11110000111011010111011000111111100110000100	63.90	60.90	58.50
	11011100111011010111011000101111011110000100	64.30	61.40	58.10
	01011100101011110000101000100110001111100010	63.70	60.70	59.20
	01011110101010110001101000100111001111100010	63.00	61.00	58.70
	01011111101011110001101111100110000111100010	64.50	60.20	58.40
	11011100101011110000100000100110001111100010	63.80	60.30	58.90
	01001101101011110000101000100110001111100010	63.90	60.60	58.80

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proc. of the LREC workshop on Linguistic Coreference*, pages 563–566, Granada.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proc. of EMNLP*.

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.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassell, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proc. of LREC*.

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

Veronique Hoste. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Antwerp University.

Ryu Iida, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proc. EACL Workshop on the Computational Treatment of Anaphora*.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proc. of ACL 2004*, pages 136–143.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. NAACL / EMNLP*, Vancouver.

Art Munson, Claire Cardie, and Rich Caruana. 2005. Optimizing to arbitrary NLP metrics using ensemble selection. In *Proceedings of HLT/EMNLP*, pages 539–546.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for C omputational Linguistics*, pages 104–111.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Marta Recasens and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In S. Lalitha Devi, A. Branco, and R. Mitkov, editors, *Anaphora Processing and Applications (DAARC 2009*, number 5847 in LNAI, pages 29–42, Berlin / Heidelberg. Springer-Verlag.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*.

Marta Recasens, Lluis Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proc. SEMEVAL 2010*, Uppsala.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistic*, 27(4):521–544.

Olga Uryupina. 2007. *Knowledge Acquisition for Coreference Resolution*. Ph.D. thesis, University of the Saarland.

Olga Uryupina. 2010. Corry: a system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Eval uation (SemEval'10)*.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: a modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 9–12.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the Sixth Message Understanding Conference*, pages 45–52.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2005. A twin-candidate model of coreference resolution with non-anaphor identification capability. In *Proc. of IJCNLP*, pages 719–730.

Shanheng Zhao and Hwee Tou Ng. 2010. Maximum metric score training for coreference resolution. In *Proceeding of COLING-2010*.