Impact of Less Skewed Distributions on Efficiency and Effectiveness of Biomedical Relation Extraction

Md. Faisal Mahbub Chowdhury[1,2] Alberto Lavelli[2]
(1) Fondazione Bruno Kessler (FBK-irst), via Sommarive 18, I-38100 Povo, Trento, Italy
(2) University of Trento, via Sommarive 5, I-38123 Povo, Trento, Italy
chowdhury@fbk.eu, lavelli@fbk.eu

Abstract

Like in other NLP tasks, it has been claimed that advances of machine learning (ML) based approaches to relation extraction (RE) are hampered by the imbalanced distribution of positive and negative instances in the annotated training data. Usually, the number of negative instances is much larger than that of the positive ones and such skewness also exists in the test data. In this paper, we aim at addressing the problem of imbalanced distribution by automatically curbing *less informative* negative instances. We propose some criteria for identifying such instances and incorporate them in an existing state-of-the-art RE approach. Empirical results on 5 benchmark biomedical corpora show that our proposed approach improves both recall and F_1 scores. At the same time, there is a large drop in the number of negative instances and in execution runtime as well.

Title and Abstract in Italian

L'Impatto di Distribuzioni Meno Squilibrate sull'Efficienza e l'Efficacia dell'Estrazione di Relazioni Biomediche

Come per altri compiti di Trattamento Automatico del Linguaggio, si è sostenuto che i progressi degli approcci all'estrazione di relazioni basati su apprendimento automatico sono ostacolati dalla distribuzione squilibrata dei casi positivi e negativi nei dati di addestramento annotati. Generalmente, il numero di istanze negative è molto più grande del numero di quelli positivi e tale squilibrio esiste anche nei dati di test. In questo articolo, ci si propone di affrontare il problema della distribuzione squilibrata eliminando automaticamente le istanze negative *meno informative*. Proponiamo alcuni criteri per individuare tali casi e inserirli in un approccio all'estrazione di relazioni con prestazioni allo stato dell'arte. I risultati empirici su 5 corpora biomedici di riferimento mostrano che l'approccio proposto migliora sia la recall sia il punteggio di F_1. Allo stesso tempo, c'è una diminuzione nel numero di istanze negative e anche nel tempo di esecuzione.

Keywords: Relation Extraction, Imbalanced Distribution, Skewed Distribution, Machine Learning, Biomedical Text Mining, Protein-Protein Interaction.

Keywords in Italian: Estrazione di Relazioni, Distribuzione non Equilibrata, Distribuzione Asimmetrica, Apprendimento Automatico, Text Mining Biomedico, Interazioni proteina-proteina.

Proceedings of COLING 2012: Posters, pages 205–216,
COLING 2012, Mumbai, December 2012.

1 Introduction

The imbalance between negative and positive annotated training instances in machine learning (ML) based approaches is a known issue. Previous studies have empirically shown that unbalanced datasets lead to poor performance for the minority class (Weiss and Provost, 2001). Apart from some exceptions, the number of negative instances is usually higher than that of the positive instances. As Gliozzo et al. (2005) argued, in most cases the error rate of a classifier trained on a skewed dataset is typically very low for the majority class and this results in biased estimation (Kotsiantis and Pintelas, 2003) and suboptimal classification performance (Chawla et al., 2004).

Some ML techniques have built-in mechanisms to deal with the skewness in somewhat limited scope[1]. But, according to the empirical results (obtained using SVM) presented in this paper, this might not guarantee to overcome completely the impact of skewness. Some ML algorithms (e.g. kNN) do instance pruning during training while maintaining the generalization accuracy. However, the main drawback of such techniques is the increased time complexity, which is generally quadratic in the data set size, without any guarantee of performance improvement (Gliozzo et al., 2005).

There exist some works in NLP that deal with the problem of skewed data distribution. For example, in the context of named entity recognition (NER), stopword filtering is used to reduce the number of candidate tokens to be considered as target entities (Gliozzo et al., 2005; Giuliano et al., 2006b). In the context of relation extraction (RE), recently Sun et al. (2011) adopted the strategy of discarding any pair of mentions proposed as candidate instance if they were separated by more than two other mentions. They conducted their experiments on news domain texts. However, the increment/decrement of performance due to such filtering was not reported.

In this paper to improve RE system's performance we have tried to reduce skewness in data by automatically identifying and removing what we call *"less informative"* instances[2]. In particular, we aim at explicitly addressing the following questions through empirical investigation:

1. Would reducing skewness in data distribution through negative instance reduction really lead to better RE results?
2. If the answer is 'yes', then can we achieve such goal by randomly discarding negative instances? Or, do we need to define an automatic methodology for singling out less informative instances?
3. To what extent could the data skewness and the efficiency (i.e. runtime) be reduced? Would the reduction of skewness help to train a better ML model/classifier?

The task chosen for our experiments is Protein–Protein Interaction (PPI) extraction from scientific papers, a widely investigated topic in biomedical RE. We adopt a state-of-the-art PPI extraction approach as a baseline system for our experiments and apply various techniques to reduce the number of negative instances being considered. We show that, by discarding less informative instances, it is possible to improve both efficiency and effectiveness of the system.

The remaining of this paper is organized as follows. First, Section 2 includes a brief discussion of the related work. Then in Section 3, we describe the data used in our experiments. Section 4 provides details about the baseline RE method used for the experiments. Following that, in Sections 5 and 6, we discuss our proposed approach. Empirical results are presented in Section 7. Finally, we summarize our work and discuss future directions.

[1] E.g., SVM allows to provide a cost-factor by which training errors on positive instances outweigh errors on the negatives.

[2] These are groups of instances that share some common characteristics and whose exclusion results in better performance.

2 Background

In Section 1, we have briefly discussed the problem of skewness of positive and negative instances in annotated data and mentioned some of the works in NLP on reducing such skewness. Due to space limitation, we cannot discuss other related NLP works not focussed on RE.[3] Previous studies (e.g. Sun et al. (2011)) hypothesized (without providing empirical evidence) that unbalanced distribution of instances is an obstacle for further improving the performance of RE.

Ideally, before reducing skewness of instances, informativeness of both positive and negative instances should be taken on account. In their seminal work regarding selection of features and instances Blum and Langley (1997) pointed out that as learning progresses and the learner's knowledge about certain parts of the training data increases, the remaining data which are similar to the already "well-understood" portion become less useful.

Our goal is to get rid of such instances from the training data prior to training the ML classifier to reduce imbalance in instance distribution and obtain a more accurately learned model/classifier. Ideally, a well trained classifier is expected to successfully identify all negative test instances. But, in practice, sometimes it would mistakenly label some of the negative instances as *(false) positives*. So, we also want to automatically get rid of as many (true) negative instances as possible from the test data (before applying the learned classifier) using the same knowledge used to reduce skewness in training data. Hopefully this would reduce the number of *false positives* produced by the classifier.

Different techniques are employed in open domain IE[4] for filtering irrelevant data to construct datasets. For example, whether the semantic type of the retrieved entity mentions and that of the target mentions are the same[5], or the number of words between the candidate mentions is greater than a certain limit, etc (Banko et al., 2007; Wu and Weld, 2010; Wang et al., 2011). However, such filtering is applied in a setting substantially different from ours.

Regarding the previous work on PPI extraction, several RE approaches have been reported to date. Most of them are based on kernel methods (Bunescu and Mooney, 2006; Giuliano et al., 2006a; Airola et al., 2008; Miwa et al., 2009a,b; Kim et al., 2010; Tikk et al., 2010; Chowdhury and Lavelli, 2012b). Among the state-of-the-art systems, Miwa et al. (2009a) proposed a hybrid kernel by combining graph, tree and bag-of-words kernel. They further boosted system performance by training on multiple PPI corpora and adopting a corpus weighting concept with SVM (Miwa et al., 2009b). Chowdhury and Lavelli (2012b) proposed an approach in which they combined different types of information and their different representations into a hybrid kernel and showed that they can complement each other to obtain state-of-the-art results.

3 Data

There are 5 frequently used benchmark PPI corpora: IEPA (Ding et al., 2002), LLL (Nédellec, 2005), AIMed (Bunescu et al., 2005), HPRD50 (Fundel et al., 2007) and BioInfer (Pyysalo et al., 2007). We use the common annotation format of these corpora provided by Pyysalo et al. (2008).

Although all these corpora are annotated for PPI extraction, the differences in performance of the same system on these corpora reported by previous studies are quite dramatic. This is due to the fact that there is no general consensus regarding PPI annotation. Furthermore, there are differences

[3] There exist many related ML studies (e.g. He and Garcia (2009)), apart from the ones discussed in this paper.

[4] Open domain IE has substantial differences with traditional RE some of which are discussed in Wang et al. (2011).

[5] In traditional RE, any pair of mentions to be considered as an instance must satisfy the already known argument types of the target relation. Hence, this technique does not qualify as a criterion for negative instance filtering in traditional RE.

in the entity types, too (i.e. the PPI annotations are not just restricted to proteins). Pyysalo et al. (2008) reported their findings of quantitative and qualitative analyses of the annotations and their differences. In a different study, Chowdhury and Lavelli (2012a) reported statistics of various characteristics of these five corpora and this study pointed out that they are quite distinct datasets.

4 Baseline RE System

As a starting point, we use the state-of-the-art system proposed by Chowdhury and Lavelli (2012b) (where more details about the system can be found). We made few minor changes in data preprocessing of the system. The main change concerns entity blinding. Originally, entity mentions were replaced by placeholders such as *Entity0*, *Entity1*, . . . where the digits represent corresponding entity mention indices inside the given sentence. Such replacement did not include a co-reference mechanism when a particular entity is mentioned multiple times inside a given sentence (using exactly the same string). For the experiments in this paper, if it appears that two (or more) mentions consist of the same string, then we replace them with the same placeholder. In the remaining of the paper, we will refer to this system as the ***primary baseline system***.

5 Anti-positive Governors

The semantic roles of the entity mentions somehow contribute either to relate or not to relate them in a particular relation type (e.g. PPI) in the corresponding context. In other words, the semantic roles of two mentions in the same context could provide an indication whether the relation of interest does *not* hold between them. Interestingly, the word on which a certain entity mention is (syntactically) dependent (along with the dependency type) could often provide a clue of the semantic role of such mention in the corresponding sentence.

Our goal is to automatically identify the words (if there any) that tend to prevent mentions, which are directly dependent on those words, from participating in a certain relation of interest with any other mention in the same sentence. We call such words as ***anti-positive governors*** and assume that they could be exploited to identify negative instances (i.e. negative entity mention pairs) in advance. Below we describe our approach for the automatic identification of such words.

Let $\mathtt{EN}$ be the set of entity mentions such that if ${\mathtt{e}^{i}}_{s} \in \mathtt{EN}$ (where s indicates the corresponding training sentence and i indicates the corresponding entity mention index inside such sentence), then ${\mathtt{e}^{i}}_{s}$ does not have any relation of interest (i.e. PPI) with any other mention inside the same sentence.

Let $\mathtt{EP}$ be the set of entity mentions such that if ${\mathtt{e}^{k}}_{s} \in \mathtt{EP}$ (where s indicates the corresponding training sentence and k indicates the corresponding entity mention index inside such sentence), then ${\mathtt{e}^{k}}_{s}$ has at least one relation of interest with one of the mentions inside the same sentence.

For example, consider the following sentence (taken from the IEPA corpus) where there are three entity mention annotations – *oxytocin*^{1}, *oxytocin*^{2} and *IP3*^{3}.

> *These results indicate that oTP-1 may prevent luteolysis by inhibiting development of endometrial responsiveness to* ***oxytocin***^{1} *and, therefore, reduce* ***oxytocin***^{2}*-induced synthesis of* ***IP3***^{3} *and PGF2 alpha.*

Here, the mention *oxytocin*^{1} does not participate in any PPI relation in this sentence. So, it would be included in $\mathtt{EN}$. The other two mentions would be added to $\mathtt{EP}$, because they are in PPI relation with each other. Note that, the two mentions of the entity *oxytocin* are treated separately.

Now, let $\mathtt{GV}$ be the set of governor words where for each $\mathtt{w} \in \mathtt{GV}$, *(i)* there is at least one mention ${\mathtt{e}^{i}}_{s} \in \mathtt{EN}$ which is syntactically dependent on $\mathtt{w}$ in the corresponding training sentence s, and *(ii)* there is *no* mention $\mathtt{e}^k{}_s \in \mathtt{EP}$ which is syntactically dependent on `w` in the corresponding training sentence s. We call this set GV as the list of *anti-positive governors*.

6 Detection and Removal of Less Informative Negative Instances

We exploit static (i.e. already known, heuristically motivated) and dynamic (i.e. automatically collected from the data) knowledge for identifying less informative negative instances as described by the following criteria:

- **C1:** If two entity mentions in a sentence refer to the same entity, then it is unlikely that they would have a relation (for our experiments, PPI relation) between themselves.
- **C2:** If each of the two entity mentions (of a candidate pair) have *anti-positive governors* with respect to the type of the relation, then they are not likely to be in a given relation.
- **C3:** If a mention is the abbreviation of another mention (i.e. they refer to the same entity), then they are unlikely to be in a relation.

Criteria C1 and C3 (static knowledge) are quite intuitive. Criterion C2 is motivated by our analyses of some randomly selected sentences from the PPI corpora (and also by what we described at the beginning of Section 5). For criterion C1, we simply check whether two mentions have the same name and there is more than one character between them[6]. As for criterion C2, we construct a list of *anti-positive governors* (dynamic knowledge) from the training data on the fly and use them for detecting pairs that are unlikely to be in relation. For criterion C3, we look for any expression of the form *"Protein1 (Protein2)"* and consider *"Protein2"* as an abbreviation or alias of *"Protein1"*.

7 Results and Discussion

All experiments are conducted (on a computer having Intel(R) Xeon(R) CPU W3520 @ 2.67GHz processor and 4GB RAM) by doing 10-fold cross validation using exactly the same procedures and folds used by Tikk et al. (2010) and Chowdhury and Lavelli (2012b). SVM hyperparameters are tuned separately on 25% data of each dataset during each experiment. The ratio of negative and positive examples is used as value of the SVM parameter known as cost-factor.

7.1 Experiments using the Three Criteria Incrementally

In these experiments, we created another baseline system (henceforth, **2nd baseline system**) by applying the strategy of Sun et al. (2011) for limiting negative instances (see Section 1) in the *primary baseline system*. Also, we created three new different systems (henceforth, **new systems**) by incrementally incorporating the three criteria (see Section 6) into the *primary baseline system*.

The *2nd baseline system* and the three *new systems* use a less skewed distribution and a smaller number of training instances than the *primary baseline system*. They also consider a smaller number of candidate test instances since some of them are automatically discarded by their corresponding criteria for the exclusion of (possible) negative (candidate) instances.

To make sure that results of the *2nd baseline system* and the *new systems* are directly comparable with the *primary baseline system*, we simply consider all the discarded candidate test instances by these four systems as negatives. If the actual label of a discarded test instance is *true*, then we consider it as a *false negative (FN)* during the calculation of precision, recall and F_1 scores.

[6] In biomedical literature sometimes expressions such as *"Protein1-Protein1"* refer to PPI. We wanted to keep mention pairs of such expressions even if the mentions have the same name.

As Table 1 shows, the *2nd baseline system* performs almost similarly as the *primary baseline system*, except that it obtains quite lower F_1 score on BioInfer. On the contrary, all the *new systems* perform better than (or as effectively as) both the *primary baseline* and *2nd baseline* systems.

The improvement of F_1 scores for the *new system v3*, which integrates all the three criteria, with respect to the *primary baseline system* is as follows – LLL: **+1.3**, HPRD50: **+6.3**, IEPA: **+1.8**, AIMed: **+1.1**. The F_1 score for BioInfer remained the same (more on this in Section 7.4). The differences in F_1 scores (except on BioInfer) are statistically significant (verified using *Approximate Randomization Procedure* (Noreen, 1989); number of iterations = 1,000, confidence level = 0.05).

The improvement of F_1 scores for the *new system v3* with respect to the *2nd baseline system* is as follows – LLL: **+1.2**, HPRD50: **+5.5**, IEPA: **+1.8**, AIMed: **+1.6**, BioInfer: **+2.2**. These differences are statistically significant, too.

A noticeable observation is that the *new system v3* obtains better recall than the *primary baseline system* on each of the corpora except LLL. For LLL, the recall remains the same but precision increases by *1.9* points. Similarly, the *new system v3* obtains better recall than the *2nd baseline system* on each of the corpora except AIMed. Although the recall decreases on AIMed, the F_1 scores improves by *1.6* points due to a significant improvement in precision.

Table 1 also reports the *AUC* scores computed following the same way as Airola et al. (2008) did. It is hard to draw any conclusion from these *AUC* scores. It should be noted that the practical value of *AUC* has been called into question by some recent ML studies (Lobo et al., 2008; Hand, 2009).

LLL	HPRD50	IEPA	AIMed	BioInfer
AUC / P / R / F_1	**AUC / P / R / F_1**	**AUC / P / R / F_1**	**AUC / P / R / F_1**	**AUC / P / R / F_1**
Primary baseline system: Using all the instances				
88.1 / 69.6 / 96.3 / 80.8	77.2 / 55.7 / 81.0 / 66.0	86.0 / 76.1 / 75.8 / 75.9	88.1 / 63.3 / 58.0 / 60.5	92.9 / 78.0 / 74.7 / 76.3
2nd baseline system: Adding the approach proposed by Sun et al. (2011) in the *primary baseline system*				
88.0 / 72.5 / 91.5 / 80.9	77.8 / 57.5 / 79.8 / 66.8	85.9 / 76.1 / 75.8 / 75.9	87.2 / 55.2 / 65.6 / 60.0	93.2 / 79.0 / 69.8 / 74.1
New system v1: Adding criterion **C1** in the *primary baseline system*				
88.4 / 70.0 / 98.2 / **81.7**	79.0 / 60.3 / 82.8 / **69.8**	85.2 / 78.2 / 76.1 / **77.2**	87.4 / 64.0 / 58.7 / **61.2**	93.1 / 77.5 / 75.2 / **76.3**
New system v2: Adding criterion **C2** in the *new system v1*				
88.4 / 71.5 / 96.3 / **82.1**	80.8 / 65.0 / 81.0 / **72.1**	85.3 / 78.0 / 77.3 / **77.7**	87.3 / 63.6 / 58.9 / **61.2**	93.1 / 77.5 / 75.2 / **76.3**
New system v3: Adding criterion **C3** in the *new system v2*				
88.4 / 71.5 / 96.3 / **82.1**	80.1 / 64.9 / 81.6 / **72.3**	85.3 / 78.0 / 77.3 / **77.7**	86.7 / 63.3 / 59.9 / **61.6**	93.1 / 77.2 / 75.4 / **76.3**

Table 1: Results on the five corpora for the *primary baseline, 2nd baseline* and *new systems*. Note that discarded positive and negative test instances (for the *2nd baseline* and *new systems*) are automatically considered as *false negatives* and *true negatives* during the calculation of the scores.

As we can see, the improvement of F_1 scores (mentioned above) varies from corpus to corpus. One of the major differences among these corpora concerns their size. But since they also have other differences, e.g. different annotation guidelines regarding entity mentions and relations, these variations of F_1 scores cannot be exclusively attributed to the disparity of corpus size. To test the influence of data size on performance, we carried out a set of experiments on a single corpus using different proportions of training data. These experiments are done on AIMed, the largest corpus among the 5 corpora considered, having 1,955 sentences collected from 225 PubMed abstracts. The learning curve for the *primary baseline system* and the *new system v3* (not reported because of lack of space) using 25, 40, 50, 60, 75 and 100% data shows that our approach for reducing skewness obtains slightly better results when the size of the corpus gets bigger.

7.2 Random Removal of Negative Instances

We wanted to investigate what happens if one decides to reduce the skewed distribution by randomly removing instances of the majority class (i.e. negative instances). This would help to better understand the effectiveness of our idea of singling out less informative instances.

For each corpus, the number of randomly discarded negative instances from the training data was kept equal to that discarded by the *new system v3*. To put it differently, the ratio of positive and negative training instances of this *3rd baseline system* (which uses random sampling) is equal to that of the *new system v3*.

As shown in Table 2, in 3 out of the 5 corpora there was a slight increase of F_1 scores for the *3rd baseline system* with respect to those for the *primary baseline system*. There was no change on BioInfer but deteriorating F_1 score on LLL. Overall, the results of the *new system v3* are better than that obtained by exploiting random sampling.

7.3 Impact on Efficiency Improvement and Skewness Reduction

Table 3 shows how much the runtime and the distribution of positive and negative instances were reduced in the *new systems* with respect to those of the *primary baseline system*. All these systems are much faster than their original *primary baseline system*. The reduction of runtime for the final version (*new system v3*) ranges from 15% to 33% depending on the corpora.

As for the reduction of skewness in the instance distribution, the number of negative instances decreased quite sharply (>= 20%) for all the corpora except BioInfer. While positive instances also decreased, the percentage of such reduction is negligible with respect to that of the negative instances.

It is evident from the numbers in Table 3 that for LLL and IEPA the greater number of negative instances were discarded in *new systems v1* and *v2*. For HPRD50 and AIMed, a considerable decrease in negative instances can be observed in each of the new systems.

The decrease of negative instances for the *2nd baseline system* (see Table 3) is negligible, while the decrease of positive instances is worrying. This suggests that merely considering the number of entity mentions in between the target mentions (as in Sun et al. (2011)) is not an effective strategy.

7.4 Peculiarities of the BioInfer Corpus

Since the F_1 score on the BioInfer corpus did not change (Table 1), we wanted to understand why it is so. The first peculiarity that we observed in the BioInfer corpus is that 2.19% of its PPIs are between entity mentions having the same name. The only other corpus which has such annotations is AIMed, but only 0.20%. So, although the criterion *C1* discarded 6.69% negative instances in BioInfer (Table 3), that was perhaps not enough to counter the loss of information due to the discarded PPIs.

The second peculiarity is that the usage of *anti-positive governors* (criterion *C2*) actually discarded positive instances in BioInfer and failed to filter any negative instance. To check why it is so, we extracted the list of *anti-positive governors* from the whole BioInfer corpus (total 1,100 sentences) and found there are only 10 such words. By comparison, the number of *anti-positive governors* in AIMed (total 1,955 sentences) and IEPA (total 486 sentences) are 300 and 161 respectively. Further investigation revealed that there are startling differences for the concentration of PPIs/sentence between BioInfer and the other corpora. For BioInfer, it is 2.30 PPIs/sentence. If we compare this with AIMed and IEPA then the respective numbers are 0.51 PPIs/sentence and 0.70 PPIs/sentence. As a result, it is quite difficult to spot a word which is not governing any mention that participate in PPI, but only governing those mentions that are not in any PPI in the corresponding sentence.

LLL			**HPRD50**			**IEPA**			**AIMed**			**BioInfer**		
P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
3rd baseline system (using random selection)														
66.0	98.2	78.9	57.8	77.3	66.1	75.5	77.3	76.4	60.3	61.7	61.0	76.5	76.2	76.3
New system v3: Implementing criteria **C1, C2,** and **C3** in the *primary baseline system*														
71.5	96.3	**82.1**	64.9	81.6	**72.3**	78.0	77.3	**77.7**	63.3	59.9	**61.6**	77.2	75.4	**76.3**

Table 2: Comparison between the results of the *3rd baseline system* (that randomly discards negative training instances) and the *new system v3*.

		LLL	**HPRD50**	**IEPA**	**AIMed**	**BioInfer**
2nd baseline system	*Reduction of runtime*	5.77%	4.84%	1.04%	6.05%	5.83%
	Reduction of positive instances	4.88%	2.45%	0.00%	4.20%	11.37%
	Reduction of negative instances	1.81%	0.37%	0.00%	1.67%	3.21%
New system v1	*Reduction of runtime*	7.69%	19.35%	19.69%	28.13%	10.80%
(criterion C1)	*Reduction of positive instances*	0.00%	0.00%	0.00%	0.20%	2.19%
	Reduction of negative instances	6.63%	12.59%	19.71%	12.83%	6.69%
New system v2	*Reduction of runtime*	17.31%	20.97%	23.32%	31.61%	13.23%
(criteria C1, C2)	*Reduction of positive instances*	1.83%	0.61%	0.60%	0.40%	2.19%
	Reduction of negative instances	19.88%	21.48%	24.07%	15.34%	6.69%
New system v3	*Reduction of runtime*	15.38%	20.97%	23.32%	33.24%	15.93%
(criteria C1, C2, C3)	*Reduction of positive instances*	1.83%	0.61%	0.60%	0.60%	2.46%
	Reduction of negative instances	19.88%	26.30%	24.07%	20.18%	9.22%

Table 3: Percentage of the decrease in runtime and number of instances for the *2nd baseline system* and for each of the *new systems* (shown in Table 1) with respect to the *primary baseline system.*

7.5 Effect of Excluding Negative Instances during Learning

An obvious question would be whether the exclusion of less informative negative instances provides any gain in learning, i.e. whether less skewed data provide a better trained model. To answer this, we performed two different sets of experiments. At first, we applied the *primary baseline system* and *New system v3* on the filtered (using the three criteria) test data. These results are reported in Table 4 which shows the recall of *New system v3* is always considerably higher than that of the *primary baseline system* for any of the corpora. As for the F_1 scores, *New system v3* obtains slightly better scores on all corpora apart from LLL (the smallest PPI corpus considered), even for BioInfer.

In a second set of experiments, we applied these two systems on the unfiltered test data. It is not possible to include details on these results in this paper for space limitation. Nevertheless, we found similar trend of better recall and slightly better F_1 scores for *New system v3* in these results. However, F_1 scores for both the systems degrades with respect to those of the previous set of experiments.

Some of the F_1 score differences in the above experiments are statistically significant, while others are not. So, the answer to the question posed above is inconclusive.

LLL			**HPRD50**			**IEPA**			**AIMed**			**BioInfer**		
P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Primary baseline system: Training using all the instances and testing only on the instances not filtered by the three criteria														
73.8	96.3	83.6	63.6	80.9	71.2	78.3	75.7	77.0	64.1	58.3	61.0	78.8	75.8	77.2
New system v3														
71.5	98.1	82.7	64.9	82.1	**72.5**	78.0	77.8	**77.9**	63.3	60.3	**61.8**	77.2	77.8	**77.5**

Table 4: Results obtained discarding the less informative test instances for the *primary baseline system* too.

	LLL	**HPRD50**	**IEPA**	**AIMed**	**BioInfer**
	P / R / F_1	**P / R / F_1**	**P / R / F_1**	**P / R / F_1**	**P / R / F_1**
Miwa et al. (2009a)	77.6 / 86.0 / 80.1	68.5 / 76.1 / 70.9	67.5 / 78.6 / 71.7	55.0 / 68.8 / 60.8	65.7 / 71.1 / 68.1
Miwa et al. (2009b)	– / – / 80.5	– / – / 69.7	– / – / 74.4	– / – / **64.2**	– / – / 67.6
Chowdhury and Lavelli (2012b)	70.4 / 95.7 / 81.1	72.9 / 59.5 / 65.5	81.1 / 69.3 / 74.7	64.2 / 58.2 / 61.1	80.0 / 71.4 / 75.5
Our proposed approach	71.5 / 96.3 / **82.1**	64.9 / 81.6 / **72.3**	78.0 / 77.3 / **77.7**	63.3 / 59.9 / 61.6	77.2 / 75.4 / **76.3**

Table 5: Comparison of our results with other state-of-the-art approaches.

8 Conclusion

In this paper, we have addressed the well known issue of skewed distribution, which has been hypothesized as one of the stumbling blocks for the advancement of ML based RE approaches (Sun et al., 2011). To the best of our knowledge, there are no existing studies showing that the reduction of skewed distribution could lead to better RE results. Since negative instances play important role for accurate ML training, only the less informative negative instances should be discarded.

To meet this challenge, we proposed three criteria for identifying less informative instances. We applied them on a state-of-the-art RE system and evaluated our approach on 5 different benchmark PPI corpora. Empirical outcome shows that our proposed approach performs better than 3 different *baseline systems* (which were created from an existing state-of-the-art RE approach) on 4 out of 5 corpora. Although the F_1 score remains the same on the 5th corpus, i.e. BioInfer, recall improves. In fact, the proposed approach boosts recall in 4 corpora (in LLL recall remains the same but precision increases) which is a desirable characteristic for biomedical RE. However, it is inconclusive whether less skewed distribution leads to a better trained model. Nonetheless, our approach significantly reduces the number of negatives instances and runtime. Comparison with previous studies shows that our approach provides state-of-the-art results for PPI extraction (see Table 5).

As for future work, we plan to investigate whether the proposed approach can also improve performance of RE from other genres of text such as news domain, since none of the criteria proposed for discarding less informative instances is domain specific.

Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments.

References

Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F., and Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI 2007)*, pages 2670–2676, Hyderabad, India.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271.

Bunescu, R., Ge, R., Kate, R., Marcotte, E., Mooney, R., Ramani, A., and Wong, Y. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine (Special Issue on Summarization and Information Extraction from Medical Documents)*, 33(2):139–155.

Bunescu, R. and Mooney, R. (2006). Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems (NIPS 2006)*, pages 171–178.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6.

Chowdhury, M. and Lavelli, A. (2012a). An Evaluation of the Effect of Automatic Preprocessing on Syntactic Parsing for Biomedical Relation Extraction. In *Proceedings of the 8th conference on International Language Resources and Evaluation (LREC 2012)*, pages 544–551.

Chowdhury, M. F. M. and Lavelli, A. (2012b). Combining tree structures, flat features and patterns for biomedical relation extraction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 420–429, Avignon, France. Association for Computational Linguistics.

Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. (2002). Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing*, pages 326–337.

Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx–relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Giuliano, C., Lavelli, A., and Romano, L. (2006a). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 401–408.

Giuliano, C., Lavelli, A., and Romano, L. (2006b). Simple Information Extraction (SIE): A portable and effective IE system. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, pages 9–16.

Gliozzo, A. M., Giuliano, C., and Rinaldi, R. (2005). Instance filtering for entity recognition. *SIGKDD Explor. Newsl.*, 7(1):11–18.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Kim, S., Yoon, J., Yang, J., and Park, S. (2010). Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11(1).

Kotsiantis, S. B. and Pintelas, P. E. (2003). Mixture of Expert Agents for Handling Imbalanced Data Sets. *Annals of Mathematics, Computing and Teleinformatics*, 1(1):46–55.

Lobo, J., Jimenez-Valverde, A., and Real, R. (2008). Auc: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17:145–151.

Miwa, M., Sætre, R., Miyao, Y., and Tsujii, J. (2009a). Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78.

Miwa, M., Sætre, R., Miyao, Y., and Tsujii, J. (2009b). A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of EMNLP 2009*, pages 121–130, Singapore.

Nédellec, C. (2005). Learning language in logic - genic interaction extraction challenge. In *Proceedings of the ICML 2005 workshop: Learning Language in Logic (LLL05)*, pages 31–37.

Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses : An Introduction*. Wiley-Interscience.

Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., and Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Jarvinen, J., and Salakoski, T. (2007). Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50.

Sun, A., Grishman, R., and Sekine, S. (2011). Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, pages 521–529, Portland, Oregon, USA. Association for Computational Linguistics.

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A Comprehensive Benchmark of Kernel Methods to Extract Protein-Protein Interactions from Literature. *PLoS Computational Biology*, 6(7).

Wang, W., Besançon, R., Ferret, O., and Grau, B. (2011). Filtering and clustering relations for unsupervised information extraction in open domain. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM 2011)*, pages 1405–1414, New York, NY, USA. ACM.

Weiss, G. and Provost, F. (2001). The effect of class distribution on classifier learning: An empirical study. Technical report, Rutgers University.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.