Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 2583 – 2594

doi:10.3906/elk-1709-185

Enlarging multiword expression dataset by co-training

Senem KUMOVA METİN*
Department of Software Engineering, Faculty of Engineering, İzmir University of Economics, İzmir, Turkey

Received: 19.09.2017 • **Accepted/Published Online:** 30.06.2018 • **Final Version:** 28.09.2018

Abstract: In multiword expressions (MWEs), multiple words unite to build a new unit in language. When MWE identification is accepted as a binary classification task, one of the most important factors in performance is to train the classifier with enough number of labelled samples. Since manual labelling is a time-consuming task, the performances of MWE recognition studies are limited with the size of the training sets. In this study, we propose the comparison-based and common-decision co-training approaches in order to enlarge the MWE dataset. In the experiments, the performances of the proposed approaches were compared to those of the standard co-training [1] and manual labelling where statistical and linguistic features are employed as two different views of the MWE dataset [2]. A number of tests with different settings were performed on a Turkish MWE dataset. Ten different classifiers were utilized in the experiments and the best performing classifier pair was observed to be the *SMO-SMO* pair. The experimental results showed that the common-decision co-training approach is an alternative to hand-labeling of large MWE datasets and both newly proposed approaches outperform the standard co-training [2] when the training set is to be enlarged in MWE classification.

Key words: Multiword expression, classification, training set, co-training

1. Introduction

The notion of learning machine is defined in [1] as follows:

"The machine that experienced the tasks of type T with evaluation metric P is stated to be learned if its performance increases to perform same type of tasks with same evaluation metric".

As given in the definition, a learning machine and/or the task of learning require the experience. There are 2 main categories of learning machines where the required experience is gained in different ways. In first category of machines, supervised machines, a labelled training set is used to supervise the learning process. In the second category of learning machines, unsupervised machines, the machine is expected to learn from unlabeled samples and discover the features of the dataset on its own. Learning machines of the these two categories have many drawbacks in different types of problems. The main drawback of supervised learning is the need for a training set of labelled samples. In cases where the required amount of labelled samples cannot be provided, the learning system commonly fails. On the other hand, in unsupervised learning, the result strongly depends on prior assumptions and appropriate choice of, e.g., distance measure, distribution function, and expected number of classes/clusters [3]. The disadvantages of supervised and unsupervised learning lead researchers to semisupervised learning which is actually the half way between the supervised and unsupervised approaches. The main goal of semisupervised learning is stated to overcome limited amounts of labelled data [4]. Simply, in semisupervised learning, the machine is forced to employ unlabeled samples in training together

*Correspondence: senem.kumova@ieu.edu.tr

with the labelled samples, commonly in order to avoid the overcosting task of constructing a large labelled dataset.

In this study, we introduce two new versions of a well-known semisupervised learning method, co-training, in order to enlarge the datasets employed in an important problem of natural language processing field: multiword expression detection. Multiword expressions are combinations of words where the semantic and/or syntactical role of the composition may be different from the individual words. In natural language processing studies, a variety of different types of multiunits such as idioms (e.g., Turkish –*"garibine gitmek"*, English –*"cut corners"*), multiword (technical) terms (e.g., Turkish –*"yapay zeka"*, English –*"expert systems"*), complex function words (e.g., Turkish –*"bununla birlikte"*, English –*"as well"*), open/hyphenated compounds (e.g., Turkish –*"açık fikirli"*, English –*"ice cream"*, *"long-term"*), named entities (e.g., Turkish –"*Atatürk Caddesi"*, English –*"Alan Turing"*) are accepted to be multiword expressions [5]. The task of MWE detection may be defined simply as the extraction of MWEs in a given text or may be accepted as a binary classification problem where a given word combination is classified as MWE or non-MWE[1]. In order to distinguish MWEs from random word combinations, a group of features/indicators (e.g., occurrence frequency) are employed [6]. In MWE identification studies where supervised methods are employed, both positive and negative MWE samples are required in training. Since manual labelling of great amounts of samples is a hard and time-consuming task, we proposed the use of standard co-training in our previous work [2]. In co-training, two independent views of the same data are provided to a pair of classifiers in order to label unlabeled samples during training [1,2]. It is accepted that employing two different views empowers the classification by providing judgement based on different perspectives.

In this study, sets of statistical and linguistic features are considered as two independent groups of features/views in our task, similar to [2]. In order to improve the previous performance results, we propose two new versions of co-training: comparison-based and common-decision co-training. Comparison-based co-training is the version where the most reliable/confident decision on classification is considered, and common-decision co-training is the version where the common decision of two classifiers is taken into account. The experimental results revealed that semisupervised learning may enlarge MWE dataset strengthening the previous results in [2] and the proposed methods (especially the common-decision approach) improve the performance in the task of MWE detection. The article is structured as follows: In Section 2, a review on co-training is given. In Section 3, the proposed co-training approaches are presented together with standard co-training method. In Sections 4 and 5, the experimental setup and results are given respectively, and the paper is concluded in Section 6.

2. Review on co-training

Co-training, proposed by Blum and Mitchell [1], is a generative semisupervised method. In co-training, a classifier $f : x \rightarrow y$ is to be built by L number of labelled and U number of unlabeled samples where each sample is represented by a feature vector x and a class label y.In co-training, in order to overcome the disadvantage of having a limited number of labelled samples, (L) Blum and Mitchell [1] proposed to split the feature vector into two independent groups of features, $x = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}$, where each group of features represents a different view of the regarding dataset. Each group of features/split/view is used to train one of the classifiers. In other words, co-training algorithm concurrently trains two classifiers, $f^{(1)} : x^{(1)} \rightarrow y$ and $f^{(2)} : x^{(2)} \rightarrow y$, that have

[1]In cases where MWE detection is accepted as a binary classification task, each word combination that is observed in the given text or provided individually is accepted to be an MWE candidate and classified as MWE or non-MWE.

different views/perspectives on the same set. After the first training phase, the first classifier labels/classifies U number of unlabeled samples considering the first group of features $x^{(1)}$. p number of positive and n number of negative samples that are most confidently labelled by the first classifier are added to the labelled dataset. The same procedure is applied for the second classifier employing $x^{(2)}$. In the next round/iteration of training, the classifiers are retrained by the enlarged dataset. The algorithm may iterate till a predefined number of times and till a predefined number of labelled samples are obtained or till there exists no samples in the unlabeled set. The assumptions that guarantee the success of co-training are listed as "Both groups of features (splits/views) must be available for classification" and "Given the label, the feature groups must be conditionally independent for each sample in the dataset" in [1]. For further details on the assumptions and different studies where co-training is employed readers may refer to [7–9].

3. Method: co-training approaches

In this study, accepting linguistic and statistical features as two different views in identification of MWEs, two classifiers were trained by 3 different approaches. In this section, the regarding approaches and features will be presented briefly.

3.1. The approaches

Co-training-based approaches in this study are as follows:

1. Standard co-training [1,2],
2. Comparison-based co-training,
3. Common-decision co-training.

In all the algorithms, the dataset L that includes a limited number of samples is enlarged iteratively by labelling the samples in the unlabeled set U employing two different views of the dataset. The final labelled dataset size is set to $N = U + L$ for standard and comparison-based algorithms. The algorithms differentiate in selection of samples that will be added to the labelled set.

The first approach in our study is standard co-training (given in Figure 1) whose performance results were previously presented in [2]. To summarize, in standard co-training, in each iteration, p number of positive samples (labelled as MWE), and n number of negative samples (labelled as non-MWE) are added to the labelled set. The labelled dataset size is increased by $2p + 2n$ number of new samples in each iteration till there exists no samples in the unlabeled dataset. In our experiments, we set $p = 1$ and $n = 1$.

In the newly proposed comparison-based approach (given in Figure 2), in each iteration, each classifier assigns p number of samples as MWE and n number of samples as non-MWE providing the confidence value of each assignment. Following, MWE labelled samples (MWE candidates) are sorted based on the confidence value in descending order and p number of MWE candidates that hold the highest confidence values are chosen to be added to the labelled dataset. The same procedure is applied to select the n most confidently non-MWE labelled samples and this set is also used to enlarge the labelled set. In this approach, when $p = n = 1$ is set, it is guaranteed that for each class the most confident sample is inserted to the labelled set.

In the common-decision approach (given in Figure 3), both classifiers label p number of MWEs and n number of non-MWEs (in each iteration) similar to standard co-training. The difference in the algorithm is that for each sample to be inserted to the labelled dataset, the same label must be assigned by both classifiers.

Given
Labelled data set L, Unlabelled data set U

Repeat until there is no sample in set U:

Train first classifier, $f^{(1)}$ by labelled data set L, using only $x^{(1)}$ features
Train second classifier, $f^{(2)}$ by labelled data set L, using only $x^{(2)}$ features
Label/classify samples of unlabelled data set U by first classifier
Label/classify samples of unlabelled data set U by second classifier.

Insert n number of negative and p number of positive samples that are most confidently labelled by first classifier to the labelled data set L.

Insert n number of negative and p number of positive samples that are most confidently labelled by second classifier to the labelled data set L.

Remove the samples that are inserted to the labelled set from the unlabelled set U.

Figure 1. Standard co-training algorithm [2].

Given
Labelled data set L, Unlabelled data set U, List T (initially empty)

Repeat until the size of the set L is N:

Train first classifier with linguistic view of set L
Train second classifier with statistical view of set L
Label the samples in U by first classifier
Label the samples in U by second classifier

Append p number of samples that are most confidently labelled as MWE by first classifier to the list T
Append p number of samples that are most confidently labelled as MWE by second classifier to the list T
Sort the samples in list T in descending order of the confidence value.
Insert first (most confident) p samples from the list T to L
Remove these p samples from U and Empty the list T

Append n number of samples that are most confidently labelled as non-MWE by first classifier to the list T
Append n number of samples that are most confidently labelled as non-MWE by second classifier to the list T
Sort the samples in list T in descending order of the confidence value.
Insert first (most confident) n samples from the list T to L
Remove these n samples from U and Empty the list T

Remove the samples that are inserted to the set L from the data set U.

Figure 2. Comparison-based co-training.

For example, a sample that is confidently labelled as MWE by the first classifier is added to the labelled set if it is also labelled as MWE by the second classifier, otherwise not. The same rule holds also for the non-MWE labelled samples. In this algorithm, it may not be possible to insert $2p + 2n$ samples to the labelled set in each iteration. In addition, the algorithm may be stuck in an infinite loop. In order to break such loops, the number of iterations is limited to some predefined number K or the algorithm iterates till there exists no samples in the unlabeled sample set. In our experiments, we set $p = 1$, $n = 1$.

Given
Labelled data set *L*, Unlabelled data set *U*, Maximum number of iterations K

Repeat till there exists no samples in *U* or number of iterations= *K*:

Train first classifier with linguistic view of set *L*
Train second classifier with statistical view of set *L*
Label the samples in *U* by first classifier
Label the samples in *U* by second classifier

Determine *p* number of samples that are most confidently labelled as MWE by first classifier
Label these samples with second classifier
Insert samples that are labelled as MWE by both classifiers to the set *L*

Determine *p* number of samples that are most confidently labelled as MWE by second classifier
Label these samples with first classifier
Insert samples that are labelled as MWE by both classifiers to the set *L*

Determine *n* number of samples that are most confidently labelled as non-MWE by first classifier
Label these samples with second classifier
Insert samples that are labelled as non-MWE by both classifiers to the set *L*

Determine *n* number of samples that are most confidently labelled as non-MWE by second classifier
Label these samples with first classifier
Insert samples that are labelled as non-MWE by both classifiers to the set *L*

Figure 3. Common-decision co-training.

3.2. Features

Co-training experiments require two different views of the same dataset. In this study, statistical and linguistic features are accepted as views of the same data in order to obtain comparable results with those of our previous work [2].

In MWE studies, statistical features provide MWE identification based on occurrence frequencies and are commonly categorized into two groups: associative measures (e.g., joint probability, point-wise mutual information) where the strength of ties between the composing words is expected to be higher for MWEs and term-hood measures (e.g., C and NC values [10]) where weak ties between the given word combination and the surrounding words is accepted to indicate the MWEs. On the other hand, linguistic features enable the identification of MWEs by examination of the properties that are extracted from written texts such as part of speech tags and inflectional suffixes of words [11]. In order to be used in our experiments, statistical and linguistic features, given in Tables 1 and 2 respectively, were chosen by the feature selection process presented in [11]. Further information/detail on the regarding features and the whole process of feature selection are given in [2,11,12].

4. Experimental setup

4.1. Dataset

In this study, we conducted tests on a Turkish MWE dataset of 8176 bigrams[2] (3946 MWEs and 4230 non-MWEs) where the samples were obtained from a merged corpus of 6 Turkish corpora (e.g., Bilkent [13], Leipzig

[2]A contiguous (uninterrupted) sequence of two words.

Table 1. Statistical view [2].

Feature	Formula
Bigram-backward variety	$v_b\left(w_1w_2\right)/f\left(w_1w_2\right)$
Bigram-forward variety	$v_f\left(w_1w_2\right)/f\left(w_1w_2\right)$
Bigram-word forward variety	$v_f\left(w_1w_2\right)/v_f\left(w_2\right)$
Fager	$\frac{f(w_1w_2)}{\sqrt{(f(w_1w_2)+f(w_1\ \bar{w}_2\ ).\ (f(w_1w_2)+f(\ \bar{w}_1\ w_2)}} - 1/2max?(f\left(w_1\ \bar{w}_2\ \right)f\left(\ \bar{w}_1\ w_2\right))$
First Kulcznsky	$f(w_1w_2)/\left(f\left(w_1\,\bar{w}_2\,\right)+\ f\left(\ \bar{w}_1\ w_2\right)\right)$
Jaccard	$\left(f\left(w_1w_2\right)\right)/\left(f\left(w_1w_2\right)+\ f\left(w_1\ \bar{w}_2\ \right)+\ f\left(\ \bar{w}_1\ w_2\right)\right)$
Joint probability	$P(w_1w_2)$
Mutual dependency	$\log\left(P(w_1w_2)^2\Big/\left(P(w_1)P(w_2)\right)\right)$
Normalized expectation	$2f(w_1w_2)/\left(f\left(w_1\right)+f(w_2)\right)$
Neighborhood Unpredictability (NUP) [29]	$FNUP\left(w_1w_2\right)=1-\left(v_f\left(w_1w_2\right)-1\right)/\left(v_f\left(w_2\right)-1\right)BNUP\left(w_1w_2\right)=1-\left(v_b\left(w_1w_2\right)-1\right)/\left(v_b\left(w_1\right)-1\right)$ $NUP\left(w_1w_2\right)=\sqrt{FNUP(w_1w_2)^2+BNUP(w_1w_2)^2}$
Point-wise mutual information	$\log\left(P(w_1w_2)/\left(P\left(w_1\right)P(w_2)\right)\right)$
Piatersky–Shapiro	$P\left(w_1w_2\right)-P\left(w_1\right)P\left(w_2\right)$
R cost	$\log\left(1+f\left(w_1w_2\right)/\left(f\left(w_1w_2\right)+f\left(w_1\,\bar{w}_2\,\right)\right)\right)+\log\left(1+f\left(w_1w_2\right)/\left(f\left(w_1w_2\right)+f\left(\ \bar{w}_1\ w_2\right)\right)\right)$
S cost	$\log\left(1+\frac{min?(f(w_1\ \bar{w}_2\ ),\ f(\ \bar{w}_1w_2\ )}{f(w_1w_2)+1}\right)$
U cost	$\log\left(1+\frac{\min(f(w_1\ \bar{w}_2\ ),\ f(\ \bar{w}_1w_2\ ))+f(w_1w_2)}{\max(f(w_1\ \bar{w}_2\ ),\ f(\ \bar{w}_1w_2\ ))+f(w_1w_2)}\right)$
Second Kulcznsky	$\frac{1}{2}\left(\frac{f(w_1w_2)}{(f(w_1w_2)+f(w_1\ \bar{w}_2\ ))}+\ \frac{f(w_1w_2)}{(f(w_1w_2)+f(\ \bar{w}_1\ w_2))}\right)$
Second Sokal–Sneath	$\frac{f(w_1w_2)}{f(w_1w_2)+2(f(w_1\ \bar{w}_2\ )+f(\ \bar{w}_1\ w_2))}$
Word forward variety	$v_f\left(w_2\right)/f\left(w_2\right)$

Table 2. Linguistic view (For further details, see [2,12,16,30]).

Feature	Formula/Definition
Partial variety in surface forms (PVSF)	$PVSF_m$: Manhattan distance between the actual surface-form histogram of the MWE candidate and the possible/expected uniform histogram. $PVSF_n$: The ratio of $PVSF_m$ to total occurrence frequency of the candidate.
Orthographical variety (OV)	OV_h: The proportion of the occurrence frequency of MWE candidate that is formed with and without a hyphen. OV_a: The proportion of the occurrence frequency of MWE candidate that is formed with and without an apostrophe.
Frozen form (FF)	FF = 1 (if MWE candidate has a single surface form) FF=0 (if MWE candidate has multiple surface forms)
Ratio of uppercase letters (RUP)	The ratio of occurrence frequency of MWE candidate where capital letters are used to the total frequency of the candidate.
Suffix sequence (SS)	The number of matches in the last n (3 to 10) characters of the candidate with the idiomatic suffix sequences in Turkish [2].
Named entity words (NEW)	The number of matches with the list of named entity word categories that are commonly used in Turkish named entities [2]. NEW may be in range [0,5] theoretically.

[14], METU [15]), similar to our previous work [2]. The regarding dataset includes bigram samples belonging to one or more of the following groups:

1. Top 200 samples in the sorted bigram lists based on occurrence frequency, point-wise mutual information, t-test, and/or χ^2 tests,

2. Frequently occurring samples that have matching part of speech tags with one of 11 predefined linguistic patterns (e.g., adverb+noun, noun+verb, adjective+noun),

3. Frequently occurring samples that mimic idioms or domain-specific terms/multiword (technical) terms by holding the same first word with one of the idioms/terms in the dictionary of idioms/technical terms,

where the term "frequently occurring" refers to the samples that are observed at least 3 times in the regarding corpus. Further details on the regarding MWE dataset can be found in [2,12,16].

In our experiments, the MWE dataset was employed to obtain the subsets given in Table 3. In Table 3, where 12 different settings are presented, the training set size represents the set size to be reached after co-training and the regarding labelled set sizes for each test-training size are given in column L.

Table 3. Dataset sizes.

Test/Training size	L (Labelled dataset size)				
100/300	50	100	200	-	-
250/750	50	100	200	500	-
300/1000	50	100	200	500	750

4.2. Evaluation

The classification performance is assessed by F1 and error rate (ER) measures. F1 is:

$$F1 = \frac{2TP}{2TP+FN+FP} \tag{1}$$

where TP is the number of true positives (candidates that are both expected and predicted to belong to the same class), FN is the number of false negatives, FP is the number of false positives. Error rate is given as:

$$ER = \frac{FP + FN}{N} \tag{2}$$

where N is the dataset size in classification. In our experiments, full-MWE comparison was performed where partial matches were ignored. Standard deviation and significance levels for test results were also provided whenever required.

4.3. Determining classifier pair(s)

In co-training, a pair of classifiers is required to enlarge the dataset iteratively as mentioned before. In order to determine the methods to be used in co-training classifiers, performances of 10 different classification methods (Naïve Bayes [17]; voted perceptron [18,19]; sequential minimal optimization (SMO) [20]; random forest [21]; one rule [22]; logistic regression (Logistic) [23]; J48 [24]; k-nearest neighbor; Bayes network (BayesNet) [25], and AdaBoostM1 [18] where J48 algorithm is boosted) were examined individually prior to the co-training process. We considered F1 and error rate (ER) measures to evaluate the classification performances of the candidate methods. The average evaluation results were obtained utilizing labelled training sets (L) with sizes N = 50, 100, 250, 500, 800, 1000 and a testing set of approximately 25% of the training set. The experiments were repeated for 5 runs and 5-fold cross validation was performed in each run. The experimental results showed that when the training set size is in the range [50,100], the performance of almost all the methods varies too much. In other words, when ER and F1 curves are drawn where the horizontal axis presents the training set size and the vertical axis gives the performance values, there exist many fluctuations in curves till $N = 250$ that results with high average standard deviations, 0.015 for ER and 0.023 for F1 measure. On the other hand, when the training set size is in the range [250,1000], it was observed that the classification performances tend to stabilize such that average standard deviations for the range [250,1000] are decreased to 0.011 for ER and 0.013 for F1 measures. As a result, we decided to determine the best performing methods based on their performances in the range [250,1000]. Table 4 gives the best performing 3 methods in sorted order both for ER and F1 measures. For example, when statistical features are utilized, SMO is the method that generates the highest average F1 value ($No = 1$, Average $F1 = 0.712$) and the best performing method in terms of ER is observed to be Logistic (Average ER = 0.306) method by generating the lowest average ER. The SD column in Table 4 gives the standard deviation values and the p(%) column shows the significance level of the performance difference (in percentages) between the best performing method and the method in the regarding row. In our experiments, p-values were calculated based on the area under the curves of performance measures. It was observed that for almost all the competing methods, p-values(%) were less than 5%, meaning that the best performing methods not only generate better average performance values but also outperform the others significantly. There exists a single exceptional/opposing case in the p(%) column where p(%) value of IBk10 row in ER measure is 20%.

Since the average ER value of IBk10 is relatively much higher than the best performing Logistic method, we decided to accept Logistic as the best performing method in the rest of our experiments.

Based on the results given in Table 4, two classifier pairs J48-Logistic and SMO-SMO are decided to be used in further experiments. In co-training, linguistic features are to be used by the first and the statistical features are to be used by the second classifier of the regarding classifier pair.

Table 4. The performances of candidate classification methods.

Measure	Statistical features					Linguistic features				
	No	Method	Average	SD	p(%)	No	Method	Average	SD	p(%)
F1	1.	SMO	0.712	0.015	-	1.	SMO	0.658	0.004	-
	2.	Logistic	0.706	0.020	0.130	2.	Logistic	0.639	0.018	0.060
	3.	J48	0.692	0.010	2.010	3.	J48	0.624	0.010	4.890
ER	1.	Logistic	0.306	0.020	-	1.	J48	0.398	0.011	-
	2.	SMO	0.309	0.012	1.080	2.	BayesNet	0.400	0.007	0.570
	3.	IBk10	0.324	0.019	20.460	*3.*	SMO	0.402	0.005	1.910

5. Results

Table 5 presents the classification performances of the two predetermined (best) classifier pairs in terms of average (Average), minimum (Min), and maximum (Max) F1 values of 12 settings given in Table 3. In order to compare the performances of the newly proposed methods, the rows Classification without co-training (L), Classification without co-training $(U+L)$, and Standard are given presenting the previously reported results in [2]. Simply, Classification without co-training (L) and $(U+L)$ rows hold the performance values of classification without co-training when L and $U+L$ number of samples are employed. The row Standard shows the performance values when standard co-training is utilized. The shaded values in Table 5 present the highest co-training results in the regarding column. For example, the highest Max F1 value (0.71) of the statistical classifier is observed when common-decision algorithm is used with SMO-SMO classifier pair. The results in Table 5 may be interpreted as follows:

1. The classification results of the co-trained SMO-SMO and J48-Logistic pairs show that in almost all F1 types (Min, Max, and Average) and approaches, the SMO-SMO pair outperforms the J48-Logistic pair.

2. When the highest co-training results are compared to Classification without co-training $(U+L)$ results that are expected to be the highest scores (upper baseline), three important remarks are observed. Firstly, Max values are equal. Secondly, Average values are 3.0% to 4.3% lower than the values obtained in Classification without co-training $(U+L)$. Thirdly, Min values are remarkably lower than the Min values of the Classification without co-training $(U+L)$.

3. The statistical classifier returns higher performance values compared to the Classification without co-training (L) results for all approaches, except common-decision co-training with J48-Logistic pair. In all configurations, linguistic classifier has lower or equal performance with the lower baseline. Based on the results, it may be stated that co-trained statistical classifier is more successful than the linguistic classifier in MWE classification.

Table 5. Testing results- F1 values.

Approach	Classifier pair	Statistical classifier				Linguistic classifier			
		Min	Average.	Max	p(%)	Min	Average	Max	p(%)
Classification without co-training (L)[2]	J48-Logistic	0.50	0.56	0.63	-	0.59	0.61	0.63	-
	SMO-SMO	0.50	0.57	0.63	-	0.60	0.65	0.67	-
Standard[2]	J48-Logistic	0.57	0.62	0.67	0.00	0.53	0.59	0.63	0.00
	SMO-SMO	0.55	0.62	0.69	0.04	0.58	0.64	0.67	0.03
Comparison-based	J48-Logistic	0.59	0.63	0.68	0.00	0.53	0.58	0.63	0.00
	SMO-SMO	0.55	0.62	0.69	0.10	0.58	0.62	0.67	0.00
Common-decision	J48-Logistic	0.54	0.57	0.62	0.00	0.56	0.60	0.63	0.01
	SMO-SMO	0.60	0.66	0.71	-	0.61	0.65	0.67	0.89
Classification without co-training ($U + L$) [2]	J48-Logistic	0.65	0.66	0.68	-	0.62	0.62	0.63	-
	SMO-SMO	0.68	0.69	0.71	-	0.66	0.67	0.67	-

4. Common-decision co-training that employs the SMO-SMO pair commonly performs better compared to other approaches. It was observed that %51.73 of the correctly classified samples are added by the statistical and %48.27 by the linguistic classifier in common-decision co-training. In other words, two classifiers equally enlarge the training set by adding correctly classified samples to the machine. On the other hand, it was observed that by this approach 52.73% of the expected training set size is reached. For instance, when $U + L = 300$ is the expected training set size, the set is enlarged to only $\sim$158 samples.

Examining the average F1 values in Table 5, it may be stated that common-decision that employs SMO-SMO pair with statistical features outperforms the other classification configurations. In Table 5, p(%) values show the significance levels of the performance difference between the common-decision algorithm that employs SMO-SMO pair with statistical features and the competing configurations, in percentages. The p levels that are all lower than 5% reveal that the common-decision approach that employs SMO-SMO pair with statistical features has significantly different F1 results.

6. Conclusion

In this study, we presented two new co-training approaches (common-decision and comparison-based co-training) that address the question of how unlabeled data may be used to enlarge the labelled MWE training dataset. We conducted various tests on a number of methods in classifier pairs that employ statistical and linguistic features as two different views to MWE data. The results showed that the SMO-SMO classifier pair with common-decision approach has a consistent performance advantage over classification scores where a small number of labelled samples exist. Though the best performance of the proposed methods did not exceed the performance of the manually labelled training set, it may be stated that the proposed approaches may be employed in MWE recognition whenever hand-labelling of large MWE datasets is not available. As a further work, we plan running experiments on other datasets (e.g., IMST-IWT [26–28]) that include labels for other linguistic features (e.g., part of speech tags) and considering MWEs of more than two words.

Acknowledgment

This work was carried out under the grant of The Scientific and Technological Research Council of Turkey (Project No. 115E469, Identification of Multiword Expressions in Turkish Texts). Further information/statistics on the MWE dataset is available on the project web page (http://app.ieu-nlpteam.com:8000).

References

[1] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: 11th Annual Conference on Computational learning Theory; 24–26 July 1998; Madison, Wisconsin, USA. USA: ACM. pp. 92-100.

[2] Kumova Metin S. Standard co-training in multiword expressiın detection. In: International Conference on Intelligent Human Computer Interaction; 10–11 December 2017; Evry, France. Springer: Evry. pp. 178-188.

[3] Klose A, Kruse R. Semi-supervised learning in knowledge discovery. Fuzzy Sets Syst 2005; 149: 209-233.

[4] Chapelle O, Schölkopf B, Zien A. Semi-supervised learning. Interdiscip Sci 2006; 2: 151-5.

[5] Constant M, Eryiğit G, Monti J, van der Plas L, Ramisch C, Rosner M et al. Multiword expression processing: a survey. Comput Linguist 2017; 43: 837-892.

[6] Tsvetkov Y, Wintner S. Identification of multiword expressions by combining multiple linguistic information sources. Comput Linguist 2014; 40: 449-468.

[7] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In: Proc 9th Int Conf Inf Knowl Manag 2002; 06 - 11 November 2000; McLean, Virginia, USA. USA:ACM. pp. 86-93.

[8] Mihalcea R. Co-training and self-training for word sense disambiguation. In: 8th Conference on Computational Natural Language Learning; 2004; Boston, MA, USA. pp. 182-183.

[9] Sarkar A. Applying Co-training methods to statistical parsing. In: 2nd *ACL*; 1–7 June 2001; Pittsburgh, PA, USA. pp. 175-182.

[10] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. Int J Digit Libr 2000; 3: 115-130.

[11] Kumova Metin S. Feature selection in multiword expression recognition. Expert Syst Appl 2018; 92: 106-123.

[12] Kumova Metin S, Taze M, Aka Uymaz H, Okur E. Multiword expression detection in Turkish using linguistic features. In: 25th Signal Processing and Communications Applications Conference; 15–18 May 2017, Antalya, Turkey. IEEE. pp. 1-4.

[13] Tur G, Hakkani-Tur D, Oflazer K. A statistical information extraction system for Turkish. Nat Lang Eng 2003; 9: 181-210.

[14] Quasthoff U, Richter M, Biemann C. Corpus portal for search in monolingual corpora. In: 5th Int. Conf. on Lang. Resources and Evaluation; May 2006; Genoa, Italy. pp. 1799-1802 .

[15] Say B, Zeyrek D, Oflazer K, Umut Ö. Development of a corpus and a treebank for present-day written Turkish. In: 11th Conference of Turkish Linguistics; January 2002. pp. 83-192.

[16] Kumova Metin S. Feature selection in multiword expression recognition. Expert Syst Appl 2018; 92: 106-123.

[17] John GHG, Langley P. Estimating continuous distributions in Bayesian classifiers. In: 11th Conf Uncertain Artif Intell; 1995; Quebec, Canada. pp. 338-345.

[18] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Mach Learn Conf (ICML '96); 03–06 July 1996; Bari, Italy. USA: Morgan Kaufmann. pp. 148-156.

[19] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in brain. Psychol Rev 1958; 65: 386-408.

[20] Platt JC. Sequential Minimal Optimization: A fast algorithm for training support vector machines. In: Technical Report MST-TR-98-14; 1998. Microsoft Research.

[21] Breiman L. Random forests. Mach Learn 2001; 45: 5–32.

[22] Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn 1993; 11: 63-90.

[23] Tabachnick BG, Fidell LS. Using Multivariate Statistics. 6th Ed. Boston, MA, USA: Pearson/Allyn & Bacon, 2007.

[24] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufman, 1993.

[25] Pearl J, Russell S. Bayesian networks. In: Arbib MA, editor. The Handbook of Brain Theory and Neural Networks. USA: MIT Press, 2003. pp. 1-11.

[26] Sulubacak U, Eryiğit G. Implementing Universal Dependency, Morphology and Multiword Expression Annotation Standards for Turkish Language Processing. Turk J Elec Eng & Comp Sci 2018; 26: 1662-1672.

[27] Sulubacak U, Gökırmak M, Tyers F, Çöltekin Ç, Nivre J, Eryiğit G. Universal dependencies for Turkish. In: 26th Int Conf on Computational Linguistics; 11–17 December 2016; Osaka, Japan. pp. 3444-3454.

[28] Pamay T, Sulubacak U, Torunaoğlu-Selamet D, Eryiğit G. The annotation process of the ITU web treebank. In: Proceedings of the 9th Linguistic Annotation Workshop; 5 June 2015; Denver, CO, USA. USA: ACL, pp. 95-101.

[29] Kumova Metin S. Neighbour unpredictability measure in multiword expression extraction. Comput Syst Sci Eng 2016; 31: 209-221.

[30] Tsvetkov Y, Wintner S. Identification of multiword expressions by combining multiple linguistic information sources. Comput Linguist 2013; 40: 449-468.