Classifying English Verbs According to the Inflection Forms

Hakkı Bağcı hakkib@cs.bilkent.edu.tr
Department of Computer Engineering
Bilkent University
Ankara, Turkey

Abstract

This paper presents results on classification of English verbs according to their inflections using some inductive machine learning algorithms such as ID3 and C4.5. It discusses the ways of getting higher accuracy on classification task and tries to find the most appropriate representation of the input data. The comparison of different data representations is made and it is shown that with a convenient data representation, it is possible to classify the English verbs according to inflection forms with a high accuracy.

1. Introduction

Classifying English verbs according to the inflection forms is important for learning the grammatical structure of English language and it also constitutes a first step of developing natural language parsers. The methods explained in this paper may also be applied to other natural languages, which have some inflection forms. Many valuable works on the learning of inflections of the verbs and on lemmatization of the inflected verbs have been made so far, however these methods do not classify the verbs according to their inflections. Most of the methods propose a way to learn an inflected form (such as past tense) of the verb from a training set, which includes for instance, present and past tenses of the verbs. The purpose of this paper is different than the mentioned works above. It does not try to learn the inflected form of the verb, but it aims to classify the inflected verb as the base form, past tense form, third person singular form or present participle form.

Mladenic (1997) says that a new approach is needed for lemmatizing the words for automatic word lemmatization. In earlier works, datasets were including the different inflections of the same verb, for instance *eating, eats, ate*. All these inflected forms were classified as *eat*, the stem of the verb. Mladenic claims that it is more appropriate to first classify the verbs according to their inflection verbs and then apply the stemming algorithms.

Ling (1994), in learning the past tense of the English verbs, proposes a classification method which makes use of C4.5 (Quinlan, 1993) algorithm. Since C4.5, an improved implementation of ID3, is an N-to-1 classifier, Ling combines the output of C4.5 and obtain a set of trees. Ling uses an input data representation called UNIBET which is introduced by MacWhinney (1990). In this representation the phonemes of a verb are assigned to numerical or alphabetical letters and each such letter stands for an attribute of the input data. Data representation used in this paper is also based on the same representation.

The approach used in this work is a symbolic method because it uses a symbolic representation and the acquired knowledge is in the form of a decision tree which is a symbolic structure. Ling discusses the advantages of the symbolic approach over the connectionist models in his work, Learning the Past Tense of the English Verbs: The Symbolic Pattern Associator vs. Connectionist Models, 1994.

The rest of the paper is organized as follows: In section 2, the problem description is given; in section 3 the data representation is introduced. Section 4 is about the classification algorithms used and section 5 explains the experimental setup. In section 6 we give the test results and conclusion comes in section 7.

2. Problem Description

Given a set of verbs which are inflected in several different forms like past tense, present participle, third person singular and base form, we have to classify them according to their inflection forms. The dataset may include different inflected forms of the same verb and irregular verbs may also exist in the dataset. A dataset may include the textual representation of the verbs or the phonetic representation, but not both at the same time. An example input and output schema is shown below:

In textual representation:

```
Eat => base form
Ate => past form
Smiling => present participle form
looks => third person singular form
```

OR

In phonetic representation:

```
IksEptId => past form
6dapt => base form
6kOrdIN => present participle form
6k1nts => third person singular form
```

We want to achieve the mapping shown above with a high prediction rate and for this purpose we want to find the appropriate data representation to be able to use the decision tree classification algorithms.

3. Data Representation

In the classification of verbs according to their inflection forms, the input data representation is a key point. The model to be used should adequately represent the regularities to be caught by the classification algorithm. It should also support both the textual format and phonetic format. Supporting textual representation is important because the input data will be most probably extracted from the text documents. It should not be necessary to convert textual representation into phonetic format to be able to apply the classification operation.

By considering the requirements above, we decided to use a data representation that takes every character of a verb (in textual or phonetic format) as an attribute. However, since the verbs are not at the same length, number of attributes change from verb to verb. To avoid this problem, we force a fixed length, (15 in our experiments) and for verbs which are shorter than 15, we put "-" characters at the end or beginning of the verb according to the assumed representation. This approach ensures that the number of attributes is same for every verb. In addition the last attributes will represent the class value. An example input data representation is given below:

Textual representation

```
E,a,t,-,-,-,-,-,-,-,-,-,-,-,-,b
a,t,e,-,-,-,-,-,-,-,-,-,-,-,-,d
s,m,i,l,i,n,g,-,-,-,-,-,-,-,-,g
l,o,o,k,s,-,-,-,-,-,-,-,-,-,-,s
```

Phonetic representation (in UNIBET)

```
I,k,s,E,p,t,I,d,-,-,-,-,-,-,-,d
6,d,a,p,t,-,-,-,-,-,-,-,-,-,-,b
6,k,O,r,d,I,N,-,-,-,-,-,-,-,-,g
6,k,1,n,t,s,-,-,-,-,-,-,-,-,-,s
```

The class values `b,d,g,s` stands for base form, past tense, present participle and third person singular form respectively.

For phonetic representation we adopt to UNIBET representation which assigns for every phonetic unit a numerical or alphabetical character.

The basic data representation is as shown above; however it is a left-justified representation. In the experiment right-justified representation is also used and the results are compared. The results differ according to the representation because the values of the attributes change according to the representation. As seen in the following example, an attribute column includes different values in one representation and different in another.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
e	a	t	i	n	g	-	-	-	-	-	-	-	-	-	g
s	m	i	l	i	n	g	-	-	-	-	-	-	-	-	g

Table 1: Left-Justified representation

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
-	-	-	-	-	-	-	-	-	e	a	t	i	n	g	g
-	-	-	-	-	-	-	-	s	m	i	l	i	n	g	g

Table 2: Right-Justified representation

From the examples above, it is seen that *–ing* parts of the verbs are put in the same attributes in right-justified representation whereas in left-justified representation they are the values of different attributes. We will see how this changes the accuracy of the algorithm in the following sections.

4. Classification Algorithms

The algorithms used for classification operation are ID3 and C4.5 which are introduced by Quinlan. These two algorithms are based on decision trees and the latter one is an improved implementation of the former one. C4.5 accounts for unavailable attribute values, pruning of decision trees and it has some other extended features. We will also compare the algorithms to

find out which algorithm is more suitable for classifying English verbs according to the inflection forms.

5. Experimental Setup

The datasets used in our experiments are obtained from a large dataset which include nearly 7000 inflected forms of approximately 2000 different verbs. The dataset include both textual and phonetic representation of the verbs. The required datasets are extracted from the main dataset according to some criteria which depend on the aim of the experiment to be performed.

For testing the accuracy of classification task we used the WEKA which is a library of machine learning algorithms written in Java.

6. Test Results

The experiments are done on separated test sets for every class first to see the accuracy of classification for each class separately. In the end, test sets that include instances from all classes are used to see the overall accuracy. The training is always done with datasets that include instances from all of the classes. In all experiments, test and training sets are disjoint, if not explicitly told the opposite.

6.1. Testing Base Form

The test dataset used in this section includes only the base forms of the verbs. The results are shown in the following table:

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4444	1070	54.3%	76.8%	72.0.6%	78.9%
1480	1070	45.8%	68.9%	67.1%	76.2%
877	1070	39.0%	64.6%	65.4%	72.2%
472	1070	39.0%	72.0%	63.4%	68.1%
344	1070	31.1%	57.3%	60.2%	56.5%
196	1070	24.3%	58.3%	62.9%	54.0%

Table 3: Results of testing base form in phonetic representation

As seen from Table 3, accuracy for predicting the base form in phonetic representation is close to 75-80% for large training sets. Accuracy decreases slowly while the training size decreases sharply. An interesting observation is that while training size decreases from 344 to 196, accuracy for right-justified test data increased by 1 percent in ID3 case. In C4.5 case we cannot observe the similar situation. In C4.5 case, accuracy decreases while the training size decreases. Actually that is the expected result. It is obviously seen that R-justified representation is more suitable than L-justified representation. While the accuracy decrease according to representation is about 20% for ID3, it is less than 10% for C4.5. We can say that C4.5 is more robust to data representation change than ID3.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4444	1070	74.3%	87.4%	93.1%	95.0%
1480	1070	65.1%	82.1%	89.1%	94.2%
877	1070	64.5%	79.2%	92.1%	94.3%
472	1070	55.0%	82.6%	88.8%	93.5%
344	1070	41.8%	75.0%	80.5%	93.3%
196	1070	32.4%	63.7%	83.3%	89.7%

Table 4: Results of testing base form in textual representation

In textual representation, the results are much better than phonetic representation as seen from Table 4. Even for small training sizes, especially C4.5 performs well both for L-justified and R-Justified representation. In addition to these observations, when we analyze the confusion matrices, we see that base form is mostly confused with past tense form. The reason for this may be the irregular verbs because they look like the base forms since they do not take the suffix –d.

6.2. Testing Past Tense Form

Test procedure for past tense form is done by separately testing regular and irregular verbs. Results are shown in the tables 5,6,7 and 8.

6.2.1 Testing Regular Past Tense Form

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4876	638	65.2%	80.4%	88.2%	91.7%
1616	638	38.1%	75.7%	70.2%	77.3%
992	638	24.9%	72.7%	64.9%	77.3%
513	638	21.0%	63.2%	45.8%	77.3%
387	638	19.3%	59.6%	46.1%	77.3%
214	638	19.0%	58.5%	9.9%	77.3%

Table 5: Results of testing regular past tense form in phonetic representation

Prediction rate for regular past tense form in phonetic representation is 80-90% for large training sets. C4.5 performs with accuracy around 80% even with small training sets. As seen from table 5, L-Justified representation is not suitable for regular verbs.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4876	638	87.3%	98.6%	97.6%	100.0%
1616	638	84.5%	96.7%	94.7%	100.0%
992	638	69.7%	85.6%	94.4%	100.0%
513	638	63.8%	100.0%	91.4%	100.0%
387	638	43.4%	92.6%	86.7%	100.0%
214	68	61.3%	100.0%	85.7%	100.0%

Table 6: Results of testing regular past tense form in textual representation

For textual representation we see that C4.5 correctly classifies all the regular verbs in R-Justified representation. This is because of the *–d (-ed, -ied)* suffix in textual representation. This is not the case in phonetic representation because some *–d*'s are pronounced differently, so the endings of the regular past tense verbs are not same in phonetic representation. ID3 also performs well with accuracy close to 95% on the average for R-justified representation.

6.2.2 Testing Irregular Past Tense Form

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
5446	68	10.3%	23.5%	26.5%	31.0%
1814	68	11.8%	30.9%	25.0%	44.1%
1098	68	17.6%	24.9%	33.8%	42.6%
586	68	16.2%	28.0%	17.6%	44.1%
423	68	14.7%	28.0%	11.8%	35.3%
238	68	21.0%	28.0%	10.3%	36.7%

Table 7: Results of testing irregular past tense form in phonetic representation

As seen from the table 7 and table 8, our approach is not suitable for irregular verbs for both phonetic and textual representation. When we analyze the confusion matrices we see that irregular verbs are mostly confused with base form, because they do not any suffix or some similar structure which distinguish them from base form. None of the algorithms perform well enough to be used in real applications. However, it can be said that in phonetic representation we can catch the regularities better than the textual representation for irregular verbs, because results for phonetic representation are nearly 20% better than textual representation.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
5446	68	16.2%	20.6%	5.9%	8.8%
1814	68	7.4%	11.8%	10.3%	10.3%
1098	68	10.3%	19.1%	7.4%	8.8%
586	68	11.8%	20.6%	4.4%	11.8%
423	68	10.3%	17.6%	2.9%	11.8%
238	68	5.9%	8.8%	8.8%	14.7%

Table 8: Results of testing irregular past tense form in textual representation

6.3 Testing Present Participle Form

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4844	670	80.1%	95.8%	98.2%	100.0%
1608	670	77.2%	96.1%	92.8%	100.0%
969	670	51.5%	96.9%	92.8%	100.0%
518	670	70.6%	96.0%	93.6%	100.0%
368	670	37.0%	89.7%	75.8%	100.0%
214	670	8.8%	56.3%	17.2%	100.0%

Table 9: Results of testing present participle form in phonetic representation

As seen from tables 9 and 10, both ID3 and C4.5 algorithms are very successful at classifying the present participle form of the verbs. The reason for this is the *–ing* suffix which appears at the end of every present participle form verb. Results are a bit better for textual representation compared to phonetic representation. In addition it is remarkable that C4.5 algorithm classifies present participle form with 100% success even with small training sets.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4844	670	91.0%	97.0%	99.0%	100.0%
1608	670	92.5%	97.9%	95.1%	100.0%
969	670	82.1%	91.5%	97.5%	100.0%
518	670	84.8%	96.6%	93.4%	100.0%
368	670	53.7%	89.3%	95.5%	100.0%
214	670	39.3%	100.0%	90.0%	100.0%

Table 10: Results of testing present participle form in textual representation

6.4 Testing Third Person Singular Form

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
5196	318	56.9%	85.8%	85.8%	90.3%
1747	318	36.5%	87.7%	74.2%	89.6%
1043	318	22.0%	74.8%	66.4%	90.3%
565	318	29.9%	70.1%	39.3%	67.9%
394	318	20.7%	56.9%	35.2%	67.9%
229	318	9.4%	50.3%	28.3%	100.0%

Table 11: Results of testing third person singular form in phonetic representation

Results of testing third person singular form in phonetic and textual representation are shown in tables 11 and 12 respectively. C4.5 algorithm performs with 100% success in textual representation for all training sets. In phonetic representation success rate is less than textual representation due to similarities with base form. Verb stems which end with “s” causes incorrect classification and also in phonetic representation pronunciation of some third person singular forms are different than the regular behavior. Nevertheless, it can be said that both algorithms are successful at classification of third person singular form of English verbs.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
5196	318	87.1%	100.0%	88.1%	100.0%
1747	318	63.2%	99.1%	80.5%	100.0%
1043	318	56.6%	95.3%	88.4%	100.0%
565	318	33.0%	97.8%	64.2%	100.0%
394	318	19.8%	87.7%	64.8%	100.0%
229	318	26.4%	53.5%	62.3%	100.0%

Table 12: Results of testing third person singular form in textual representation

6.5 Testing a Mixed Random Dataset

Having applied separate tests for all classes, we also run an experiment to see how our approach performs on a mixed test data. For this purpose a test data set with 1107 instances are chosen randomly and classification algorithms are applied. The results are shown in the tables 13 and 14 for phonetic and textual representation respectively.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4407	1107	70.5%	86.4%	88.3%	92.8%
2208	1107	62.1%	85.5%	87.0%	91.4%
1470	1107	55.5%	83.6%	86.0%	91.1%
731	1107	51.6%	79.2%	82.1%	90.8%
348	1107	39.0%	76.3%	78.3%	90.2%
173	1107	28.2%	61.1%	67.1%	87.0%

Table 13: Results of testing a random dataset that includes instances from all forms in phonetic representation

For phonetic representation ID3 algorithm performs good with large training sets but with small training sets it is not so successful. On the other hand, C4.5 outperforms ID3 and it also classifies with nearly 90% success rate with small training sets. For textual representation results are 7-8% better than the phonetic representation for both of the algorithms.

Training Size	Test Size	ID3		C4.5	
		L-Justified	R-Justified	L-Justified	R-Justified
4407	1107	86.5%	93.0%	93.1%	97.0%
2208	1107	81.7%	92.0%	90.6%	97.1%
1470	1107	77.9%	92.3%	90.7%	96.7%
731	1107	70.9%	88.1%	88.5%	96.7%
348	1107	52.1%	89.6%	87.5%	95.8%
173	1107	46.7%	79.0%	82.3%	95.3%

Table 14: Results of testing a random dataset that includes instances from all forms in textual representation

7. Conclusion

Data representation and bias of learning algorithms are two key points that affect the generalization ability of a machine learning approach. In this paper we proposed a data representation model and used decision tree based learning algorithms. Experiment results show that our data representation and bias of the learning algorithms are adaquate for classification of English verbs according to the inflection forms in most cases. However, the model is not successful at classifying irregular verbs. Some previous works (Ling, MacWhinney and Leinbach) also suffer from the irregular verbs and they are also far from being successful at generalizing irregular verbs.

Interpretation of experiment results shows that most suitable data representation is the right-justified representation of the verbs, because it allows us to catch the regularities which are mostly at the end of the inflected verbs. This is specific to English obviously, for other natural

languages some other representation may be more convenient. Nevertheless, we can say that taking the letters or phonetic units as attributes is a good approach for this classification task. In addition C4.5 algorithm is better than ID3 algorithm for almost all cases. It is also more robust to data representation changes. In conclusion we can say that C4.5 with right justified representation can be used to classify both phonetic and textual forms of inflected verbs with high accuracy.

8. References

Ling, C. (1994). Learning the Past Tense of English Verbs: The Symbolic Pattern Associator vs. Connectionist Models. In Journal of Artificial Intelligence Research Vol 1, 209-229.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revisiting the verb model. Cognition,40, 121-157

MacWhinney, B.(1990). The CHILDES Project: Tools for Analyzing Talk. Hillsdale, NJ: Erlbaum.

Mladenic, D. (1997) Automatic Word Lemmatization

Quinlan, J. (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA.