Automatic Turkish Text Categorization in Terms of Author, Genre and Gender

M. Fatih Amasyalı and Banu Diri

Yıldız Technical University, Computer Engineering Department,
34349 Beşiktaş, İstanbul, Turkey
{mfatih, banu}@ce.yildiz.edu.tr

Abstract. In this study, a first comprehensive text classification using n-gram model has been realized for Turkish. We worked in 3 different areas such as determining the identification of a Turkish document's author, classifying documents according to text's genre and identifying a gender of an author, automatically. Naive Bayes, Support Vector Machine, C 4.5 and Random Forest were used as classification methods and the results were given comparatively. The success in determining the author of the text, genre of the text and gender of the author was obtained as 83%, 93% and 96%, respectively.

1 Introduction

Text categorization (TC) is the process of classifying documents into a certain number of predefined categories. One of the problems in TC is the authorship attribution which is identifying the author of an anonymous text or text whose authorship is in doubt [1]. The other problem is the identification of text genre which is becoming an important application in web information management. The other problem is also a need to classify texts according to author gender.

In the last 35 years there were several studies in the identification of the author of a text. Mosteller and Wallace took the Federalist Papers and determined a very credible attribution of authorship on the basis of a range of discriminates and used Bayesian analysis [2]. Burrows [3] focused on common high-frequency words. Amongst the pioneers of authorship attribution are Brinegar [4], who focused on word lengths, Morton [4], who focused on sentence lengths, and Brainerd [4], who focused on syllables per word. Stamatatos [4] have applied Multiple Regression and Discriminant Analysis using 22 style markers. They have measured these results on ten authors. Fürnkranz [5] described an algorithm for efficient generation and frequency-based pruning of n-gram features. Cavnar et all [6]described a n-gram based approach to text categorization is tolerant of textual errors.

The traditional literature on genre that uses quantitative methods is that of Biber [7], which draws on work on stylistic analysis readability indexing and differences between spoken and written language. Kessler et al [8], who developed a simple and confident method for genre detection.

C. Kop et al. (Eds.): NLDB 2006, LNCS 3999, pp. 221–226, 2006.

Some of the works for determining the gender of a document's author are performed by Mulac et al [9], Herring [10] and Palander-Collin [11]. Koppel et al. [12] employed machine learning algorithms on a genre-controlled corpus of 566 documents taken from the British National Corpus to construct models.

In this work we used character n-grams to achieve the TC. We have figured out the language bi-grams and tri-grams by using a corpus, which is composed of Turkish texts. Using this bi-gram and tri-gram, 6 different datasets were constructed for 3 different classification problems (author, genre and gender) and 2 different n-gram models (bi-gram, tri-gram).

The remainder of the paper is organized as follows: In section 2 a brief description of n-grams and our corpus are introduced. In Section 3 Classification and feature selection algorithms are presented. In Section 4, we have discussed empirical results in our experiments. Finally, we summarize our conclusions in section 5.

2 Testing Ground

2.1 N-Grams

An n-gram is an n-character fragment of a longer string. In literature, the n-gram term is included the notion of any co-occurring set of characters in a string [6].We have handled the text as a whole and we have extracted the bi-grams and the tri-grams.

2.2 Corpus

In the text categorization experiments we chose to deal with texts taken from newspapers, considering the variety of authors publishing their writings in the press. Our corpus consists of texts downloaded from 3 Turkish daily newspapers. This corpus consists of eighteen randomly selected authors writing on different subjects like political, popular interest and sport.

The dataset consists of 630 singly authored documents written by 18 different authors, with 35 different texts written by each author. Also, this dataset has been chosen from 3 different classes such as politic, popular interest and sport in order to be used to determine the genre of the document. Again, the same dataset is composed of 4 female and 14 male authors in order to determine the gender of the author. To determine the author of text, the genre of text and the gender of the author six different data set has been constructed as shown in the Table 1.

While forming the bi-grams and the tri-grams of the corpus the number of occurrences of each feature is counted. At the end of this process we have observed that the number of different bi-grams and tri-grams are too much. In order to avoid the combinatorial explosion in the feature vectors, which consist of bi-grams and tri-grams, we used a threshold value to reduce the number of features. Infrequent features are removed from the feature vectors. The dimensions of the bi-gram and tri-gram feature vectors are 446 and 913 respectively.

Table 1. Datasets

	Bi-grams (446 features)	Tri-grams (913 features)
Author of document (18 classes)	Dataset I	Dataset II
Genre of document (3 classes)	Dataset III	Dataset IV
Gender of author (2 classes)	Dataset V	Dataset VI

3 Classification and Feature Selection Algorithms

In this work, we used 4 different classification methods such as Naive Bayes, Support Vector Machines, C 4.5 and Random Forest that are used in identification of an author, determination of a text genre and gender of an author.

Naive Bayes (NB): Classical Naive Bayes is a probabilistic classifier that uses joint probabilities of words and categories to calculate the category of a given document. In our study, the features are not word frequencies. They have continuous distributions. For this reason, Naive Bayes's WEKA (available at www.cs.waikato.ac.nz/ml/weka) implementation was used for our experiments.

Support Vector Machines (SVM): Support Vector Machines which are based on the structural risk minimization principle and mapping of input vectors in high-dimensional feature space also avoids over fitting and does not need a feature selection. We used WEKA's SVM implementation.

C4.5: The main idea of the classic decision tree algorithm is the division of the feature space into two regions. The division process is repeated until each region contains single-class data or a predefined criterion is achieved. C4.5 is univariate decision tree algorithm. At each node, only one attribute of instances are used for decision making.

Random Forest (RF): The forest consists of several different multivariate decision trees which are trained by different training sets. Different training sets are constructed from original training set by bootstrap and random feature selection. Multivariate decision trees are constructed with CART algorithm. Although, each tree has its own decision, the maximum voted class in the forest is accepted as final decision.

Correlation-based Feature Selection (CFS): CFS is one of the feature selection methods which place in WEKA. Each feature's independent prediction ability is calculated. Subsets of features that are highly correlated with the class while having low inter-correlation are selected.

4 Experimental Results

In our experiments, we showed whether the modeling of Turkish texts with n-grams is successful approach or not for in determining the author of text, genre of the text and gender of the author. For each classification problem, 4 different classifiers were trained on 2 different n-gram models. 5-fold cross validation was used for the evaluation of success ratio. Default performance (dp) is the success ratio when all instances were classified as majority class.

4.1 Author Identification

DataSet-I and DataSet-II were constructed by using bi-gram and tri-gram models for author identification, respectively. Classifiers were trained with all n-gram features and n-gram feature subsets that selected by CFS. Each author has 35 texts of which 28 were used as training set and 7 were used as test set. The success of 4 different classification methods used in determining the author of the text is given Table 2.

Table 2. Author Identification Results

dp : 5.5%	C4.5%	NB%	SVM%	RF%
Dataset-I (446 features)	59.5	76.2	72.2	74.4
Dataset-I CFS: (34 subset feature)	66.2	83.3	79	80.2
Dataset-II (913 feature)	54.4	71.4	69.7	70.1
Dataset-II CFS : (29 subset feature)	61.2	76.2	74.1	74.3

When the feature selective method is not used, Naive Bayes classifier gave the best result in identifying the author of text. Feature selection's improvement on success can be seen for all classifiers and n-gram models. Again, in both two datasets, Naive Bayes gave the best result. Bi-gram model is more successful than tri-gram in determining the author of the text.

4.2 Genre Identification

DataSet-III and DataSet-IV were constructed by using bi-gram and tri-gram models for genre identification, respectively. Classifiers were trained with all n-gram features and n-gram feature subsets that selected by CFS. Each genre has 210 texts of which 168 were used as training set and 42 were used as test set. The success of 4 different classification methods used in determining the genre of the text is given Table 3.

Table 3. Genre Identification Results

dp : 33.3%	C4.5%	NB%	SVM%	RF%
Dataset-III (446 features)	84.9	84.1	90.6	87.9
Dataset-III CFS: (34 subset feature)	88.8	86.2	92.5	92.2
Dataset-IV (913 feature)	79.8	84.3	88.5	88.9
Dataset-IV CFS : (29 subset feature)	84.2	89.8	93.6	92.7

When the feature selective method is not used, while SVM gives good result in DataSet-III, RF gave almost the same result with SVM in DataSet-IV. Feature selection's improvement on success can be seen for all classifiers and n-gram models. SVM gave good result in DataSet-III and DataSet-IV. There is no significant difference between bi-gram and tri-gram models.

4.3 Gender Identification

DataSet-V and DataSet-VI were constructed by using bi-gram and tri-gram models for determining gender of an author, respectively. Classifiers were trained with all n-gram features and n-gram feature subsets that selected by CFS. Female authors have 140 texts of which 112 were used as training set and 28 were used as test set. Male authors have 490 texts of which 392 were used as training set and 98 were used as test set. The success of 4 different classification methods used in determining the gender of the author is given Table 4.

Table 4. Gender Identification Results

dp : 77.7%	C4.5%	NB%	SVM%	RF%
Dataset-V (446 features)	84.9	83.8	88.2	88.3
Dataset-V CFS: (34 subset feature)	90.5	91.6	94.1	92.2
Dataset-VI (913 feature)	86.8	85.1	90.1	87.8
Dataset-VI CFS : (29 subset feature)	88.9	92.2	96.3	92.2

When the feature selective method is not used, while SVM and RF give the best results in DataSet-V, SVM gave better result in DataSet-VI. Feature selection's improvement on success can be seen for all classifiers and n-gram models. There is no significant difference between bi-gram and tri-gram models. SVM again over performed from all other classifiers in both datasets.

5 Conclusion

N-gram models are common and successful approach for text classification. This work is the first comprehensive study on classification of Turkish texts modeled with n-grams. Turkish texts were classified in terms of author, genre and gender. Our study shows that n-grams are suitable models solving different Turkish texts classification problems.

Using the same corpus, 6 different classification problems were constructed by labeling text according to their authors, genres and author genders. 4 different classification algorithms and a feature selection algorithm were used in this work. Default parameters in WEKA were used in all classification problems. It is observed that feature selection increased the classification accuracy on all datasets. Three classification problems are compared in terms of feature selection effect, best classifier and best n-gram model in Table 5.

Table 5. Comparing classification problems

	Author	Genre	Gender
Feature Selection Effect	Success increased	Success increased	Success increased
Best Model	Bi-gram	No difference	No difference
Best Classifier	Naive Bayes	SVM	SVM

As summarized in Table 5, when feature selective model is used we observed that the success of 3 classification problem is increased. Bi-gram model is more successful than tri-gram model in determining the author of the text. Both n-gram models gave the same success in identifying the genre of the text and gender of the author. While the Naive Bayes classifier is successful in determining the author of the text, SVM is more successful in determining the genre of the text and gender of the author.

For future work, combining other text features (lexical, syntactic annotation, stylometry, word n-grams eg.) with character n-grams is planning.

References

1. Love H.: Attributing Authorship: An Introduction, Cambridge Univ. Press (2002)
2. Dale R., Moisl H., Somers H.: Handbook of NLP, Marcel Dekker (2000)
3. Burrows J.F.: Not unless you ask nicely: the interpretative nexus between analysis and information. Literary Linguist Comput (7):pp.91-109 (1992)
4. Stamatatos E., Fakotakis N., Kokkinakis G.: Automatic Text Categorization in Terms of Genre and Author, Computational Linguistics, pp.471-495 (2000)
5. Fürnkranz J.: A Study using n-gram Features for Text Categorization, Austrian Research Institute for Artifical Intelligence (1998)
6. Cavnar W.B.: Using an n-gram-based Document Representation with a Vector Processing Retrieval Model. In Proceedings of the Third Text Retrieval Conference(TREC-3) (1994)
7. Biber D.: Dimensions of Register Variation: A Cross-Linguistic Comparison Cambridge Univ.Press (1995)
8. Kessler B., Nunberg G., Schütze H.: Automatic Detection of Text Genre, Proc. of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL'97), pp.32-38 (1997)
9. Mulac A., Studley L.B, Blau S.: The Gender-linked Language Effect in Primary and Secondary Students' impromptu Essays, Sex Roles, 9/10 (1990)
10. Herring S.: Two Variants of an Electronic Message Schema, in S.Herring ed., Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives, pp.81-106 (1996)
11. Palander C. M.: Male and Female Styles in 17th Century Correspondence, Language Variation and Change 11, pp. 123-141 (1999)
12. Koppel M., Argamon S., Shimoni A.R.: Automatically Categorizing Written Texts by Author Gender Literary and Linguistic Computing 17(4) pp.401-412 (2002)