Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik
Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering

2017 - Volume: 18 Number: 2
Page: 346 - 359
DOI: 10.18038/aubtda.322136
Received: 12 October 2016 Revised: 19 January 2017 Accepted: 11 April 2017

STOP WORD DETECTION AS A BINARY CLASSIFICATION PROBLEM

Senem KUMOVA METİN [1, *], Bahar KARAOĞLAN [2]

[1] Department of Software Engineering, Faculty of Engineering, İzmir University of Economics,
No.156, 35330, Balçova-İzmir, Turkey
[2] Ege University, International Computer Institute (ICI), 35100 Bornova-İzmir, Turkey

ABSTRACT

In a wide group of languages, the stop words, which have only grammatical roles and not contributing to information content, may be simply exposed by their relatively higher occurrence frequencies. But, in agglutinative or inflectional languages, a stop word may be observed in several different surface forms due to the inflection producing noise.

In this study, some of the well-known binary classification methods are employed to overcome the inflectional noise problem in stop word detection. The experiments are conducted on corpora of an agglutinative language, Turkish, in which the amount of inflection is high and a non-agglutinative language, English, in which the inflection is lower for stop words. The evaluations demonstrated that in Turkish corpus, the classification methods improve stop word detection with respect to frequency-based method. On the other hand, the classification methods applied on English corpora showed no improvement in the performance of stop word detection.

Keywords: Stop word, Content word, Binary classification, tf-idf

1. INTRODUCTION

Starting with [1], it is common to classify words into two classes depending on how much they contribute to the meaning of a sentence. Some words rarely contribute to the meaning that they do not have semantic functions in language. They have merely functional or grammatical roles. These words are referred as 'function', 'grammatical' or 'stop' words. The other class of words; 'content' or 'content-bearing' words; supplies the bulk of the meaning in a text (or a sentence). Stop words indicate how to connect meanings of content words in a text among other uses. Content words are open-class words, since languages can freely add new words to the set. On the other hand, stop words belong to closed-class of word; that, it is very uncommon for new stop words to emerge. The class of content words involves nouns, verbs and adjectives whereas the class of stop words includes determiners, auxiliaries, conjunctions, degree adverbs, pronouns and prepositions.

The distinction of stop and content words has an important role in applications of information retrieval and natural language processing (NLP). The aim of information retrieval systems may be summarized as finding the related information by eliminating the unrelated. The systems simply generate a list of index terms to reduce information overload. Therefore, stop words which occupy a large proportion of written texts may increase both time and resource consumption for the systems by overloading texts and sometimes reduce performance by merging into index lists [2]. In NLP applications generally stop words are filtered out prior to, or after, processing of natural language data. Removal of stop words is a common preceding step especially in applications such as machine translation and automatic summarization [3-6]. In such applications, the mechanism that is sometimes used with the aim of omitting stop words is simply not indexing them at all using a predetermined stop word list or part of speech tags [7]. The other widely established approaches that do not require for a dictionary or a preprocessing step, detect stop words using the occurrence frequency.

*Corresponding Author: senem.kumova@ieu.edu.tr

The assumption behind frequency based approaches are two-fold: "A stop word occurs relatively more than a content word in a corpus" and "The number of different documents in which a stop word occurs (document frequency) is higher relative to the number of documents in which a content word resides". The assumptions based on occurrence frequencies are stronger in languages in which the stop words are rarely prone to inflectional suffixes. Briefly, in such languages, the surface forms of stop words in corpus are deformed rarely due to the inflections when compared to agglutinative or inflectional languages. This is why, the observed frequencies of stop words in corpus are higher and the frequency-based methods are effective in discriminating stop words. However in some agglutinative languages, the inflection may generate many surface forms of the same stop word in the corpus and the resulting disrupted frequency distribution may complicate the detection of stop words by the frequency.

In this article, stop word detection is accepted as a binary classification task in which each word in a text may be assigned either to the group of stop words or to the group of non-stop words (roughly, the content words). The classification algorithms are utilized to obtain a combined classification measure by merging several discriminating features. The important criterion for determination of discriminating features is that each feature must be a syntactical feature that does not require high computational effort to be obtained from the corpus. We exploited five features in the experiments: term frequency, collocative frequency, document frequency, word length and word position.

The classification methods are evaluated on corpora of two widely used languages, Turkish and English. In the study, Turkish is expected to hold higher inflectional noise on stop words compared to English. The classification methods are evaluated with respect to previously well-established frequency based method: tf-idf.

Following sections involve related work in literature to obtain stop word lists, mathematical background of classification methods employed in the study, binary classification in stop word detection, results and the conclusion.

2. RELATED WORK

Traditionally, stop word lists are constructed from most frequently occurring words in a corpus. There are two important methods in the literature based on the frequency of occurrence: term frequency and tf-idf (term frequency x inverse document frequency).

The term frequency method simply assigns most frequently occurring words as stop words in a frequency ranked list of words. In practice, the method has two disadvantages. First, some content words that are repeated to reinforce the topic of text may be listed among the stop words [8]. Second, it is difficult to determine an accurate cut-off (threshold) value for the frequency of stop words. Although, these disadvantages reside, the method is still used in applications both in English and in other languages [9-12]. This approach is refined by stemming, part of speech filtering and normalization of frequency values. In the study, since the preliminary test results of term frequency method over the training corpora were not promising, we decided to exclude the term frequency method from the list of competing methods.

The method of tf-idf is often used in information retrieval or text mining; the method gives a measure that evaluates how important a word is to a document in a collection or corpus. The tf-idf weight (w_{ij}) of the term T_j is computed as

$$w_{ij} = tf_{ij} \cdot \log_2(\frac{N}{n}) \tag{1}$$

where tf_{ij} is frequency of term T_j , N is total number of documents in collection or corpus and n is the number of documents where the term T_j occurs at least once.

The tf-idf weight is calculated for all words in a corpus in which document boundaries are clarified and a ranked tf-idf list is generated for each document. The words having lower values are accepted as stop words and the words having higher values are assigned as content words. Although, the method generates more accurate stop word lists compared to term frequency method, it has also some weaknesses. First of all, it may not be utilized in a text collection without document boundaries. A cut-off value for stop words must be determined. In addition, if the corpus or collection involves documents with similar topics, it is possible to label some content words erroneously as stop words since they may be used in a large group of documents in corpus.

Another common approach is using predefined stop word lists. In this approach, stop words are typically selected from a dictionary/corpus or may be defined by the researcher manually. Since predefined lists are limited by the vocabulary size of the dictionary/corpus or of the researcher, it may not be possible to detect some domain specific stop words by this approach. For example, in some indexing applications, such as indexing of web documents, it is important to label the words such as “e-mail” as stop word even though they are content words for other documents. Moreover, it is not always possible to give a complete list of possible stop words for agglutinative languages in which the inflectional suffixes are widely used with stop words.

Besides the commonly used frequency based methods, there are also some improved methods for distinction of stop words; such as the term based sampling method of [13] which depends on a refined measure of frequency. A more cognitive oriented method in distinction of stop words is proposed in [14]. The authors claimed that prosodic cues such as duration, intonation contour, intensity and formants contribute as a basis for identification of stop and content words. Unlike others, this method requires speech corpus to retrieve prosodic information. In [15], a method based on chi-square statistics to build stop word lists in Chinese is presented.

Currently, a large portion of information retrieval applications is on web documents. In [16], it is claimed that web specific stop word lists would be beneficial because the current lists are out of web-specific function words (such as email, contact). They propose a method to construct entropy based stop word lists from web documents. In a more recent study, [17], linguistic and syntactic information are aggregated to build stop-word list in Persian information retrieval systems. In [17], part of speech (POS) tags are employed together with statistical measures such as entropy and the method is assessed by precision. The precision values reported are in range [0.25 0.3] for the whole set of different POS tags. Another recent study is presented in [18]. In [18], three different types of filters (hash-filter, most recently-used filter and sequence-filter) are implemented to construct a customized Chinese-English stop-word list by utilizing a classical stop word list. A method to extract context-aware stop word lists for Twitter data is proposed in [19]. The performance of the method in [19] is measured in terms of increase in binary sentiment classification performance. The main difference in [19] is that they employed an extra information source: sentiment lexicon. It is reported that the proposed method outperformed the traditional method where predetermined stop word lists are used in terms of the sentiment classification performance and the reduction.

3. MATHEMATICAL BACKGROUND of CLASSIFICATION METHODS

Binary classification methods are supervised learning methods in which discriminating effect of different features of classes are learned from a training set in order to classify the testing data set in two classes. In this study, we used discriminant analysis, decision tree, naïve bayes and k-nearest neighbor algorithms to identify stop words in a given corpus.

3.1. Discriminant Analysis

Discriminant analysis is a multivariate statistical method intended to estimate the relation between categorical dependent variables and metric independent variables (features). The analysis has assumptions of normal distribution, homogeneity of variances and co-variances across groups and independence criteria for independent variables.

Discriminant analysis can be realized by linear or quadratic functions. In this study, both the linear and the quadratic discriminant analysis with two groups have been applied. The difference between quadratic discriminant and linear discriminant analysis is that the former permits each group distribution to have its own covariance matrix, whilst the latter assumes a common covariance matrix for all group distributions.

Briefly, discriminant analysis involves determining weight to be given to each of several features (x_i) in order that the resulting composite score will have a maximum utility in distinguishing between members of groups [20]. If p different features are given, in linear analysis the desired discriminant function has the form in below

$$y = a_1x_1 + a_2x_2 + \cdots + a_px_p \tag{2}$$

where $a_1, a_2 .... a_p$ are weighting coefficients to be applied to p original scores for each observation. The purpose of the analysis is to determine optimal values for weighting coefficients such that the difference between mean scores for two groups will be maximized relative to variation within groups. This is equivalent to saying that weighting coefficients are to be derived such that t statistics or F ratio between groups will be the maximum. The function to be maximized, defined by R.A. Fisher [21], is the ratio of between-groups variance to within-groups variance in linear discriminant analysis.

$$f(a_1, a_2,..., a_p) = \frac{n_1n_2}{n_1 + n_2}\frac{(a_1d_1 + a_2d_2 + ...+ a_pd_p)^2}{\sum\sum c_{ij}a_ia_j} \tag{3}$$

In this criterion function, $[d_1d_2........d_p]$ is the vector of mean differences on p original measures, c_{ij} constructs within-groups covariation matrix, n_1 and n_2 are the number of members belonging to group 1 and group 2 respectively. The denominator of criterion function $\sum\sum c_{ij}a_ia_j$ is the within-groups variance of a linear combination in which a_i are weighting coefficients. The numerator of criterion function is proportional to between-groups variance. The problem is to solve a set of weighting coefficients that will maximize criterion function given that one has estimates of the mean vectors and within-groups variance-covariance matrix obtained from reasonably large samples [20].

3.2. Decision Tree

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree [21]. The decision tree learning may be induced by many alternating algorithms such as Hunt’s algorithm, ID3, C4.5, which are widely used in existing systems. In this study, we employed classification and regression trees (CART) which is a non-parametric decision tree learning technique that produces either classification or regression trees. If the dependent variable is categorical, the tree performs classification and if it is numerical, the regression tree is produced. While creating the classification tree, a recursive procedure that splits a node in two depending on the value of a feature is applied. The splitting of nodes stops when CART detects no further gain is obtained or the data are split as much as possible.

3.3. Naive Bayes

Naive Bayes classifier is a supervised classifier that is based on applying Bayes theorem with strong (naive) independence assumptions. The independence assumption is that the presence of a particular feature of a class is unrelated to any other feature. Simply, in Bayes classification, using the training data, the parameters of a probability distribution is estimated. For a testing instance, the posterior probability of the instance belonging to each class is computed and the instance is assigned to class with the largest probability. The further details on Naïve Bayes classification may be found in [22].

3.4. k-Nearest Neighbor (k-nm)

k-Nearest Neighbor algorithm is a non parametric lazy learning algorithm, originally proposed in [23], in which when an instance in testing set (whose class is unknown) is to be classified, the algorithm computes its k closest neighbors in the training set, and the class is assigned by voting among those neighbors. While categorizing instances in testing set based on their distance to instances in a training dataset, various metrics to determine the distance may be used. In this study, we have employed 4 different distance metrics with k ranging from 1 to 5: Euclidean, city block, cosine and correlation distance.

In k-nn algorithms, an arbitrary instance x may be described by the feature vector $f(x) = \langle a_1(x), a_2(x),...a_n(x) \rangle$ where $a_r(x)$ denotes the value of the *r*th attribute of instance *x*. Table 1 gives the formulas of distance metrics used in the study to calculate the distance between two instances x_i and x_j which is defined to be $d(x_i, x_j)$.

4. BINARY CLASSIFICATION IN STAP WORD DETECTION

Binary classification methods employ discriminating features (independents) to categorize the words in two groups. In the study, the ultimate goal is to present the contribution of syntactical features which do not require for high computational preprocessing (e.g. part of speech tagging, stemming) in distinction of stop words especially in languages which stop words may have a wide range of different forms. For this purpose, we have selected five features (word position in a sentence, word length, term frequency, document frequency and collocative frequency) that are measured with ease and expected to perform well for the corresponding languages in the experiments. Two corpora, one for training and one for testing, are utilized and evaluation of the methods is performed using testing lists obtained from testing corpora.

Following subsections involve the details and reasons to use each feature, the training and testing corpora, and the evaluation method.

4.1. Classification Features

4.1.1. Term frequences (TF)

Since stop words serve to construct the grammatical structure of the sentences, it can be accepted that each sentence contains at least one stop word. This leads stop words

Table 1. The distance metrics in k-nn classification experiments

Distance Metric	Formula
Euclidean Distance	$\sqrt{\sum_{r=1}^{n}(a_r(x_i)-a_r(x_j))^2}$
City Block Distance	$\sum_{r=1}^{n}\lvert a_r(x_i)-a_r(x_j)\rvert$
Cosine Distance	$1-\frac{\sum_{r=1}^{n}a_r(x_i)a_r(x_j)}{\sqrt{\sum_{r=1}^{n}a_r(x_i)^2}\sqrt{\sum_{r=1}^{n}a_r(x_j)^2}}$
Correlation Distance	$1-\frac{(f(x_i)-\overline{f(x_i)})(f(x_j)-\overline{f(x_j)})'}{\sqrt{(f(x_i)-\overline{f(x_i)})(f(x_i)-\overline{f(x_i)})'}\sqrt{(f(x_j)-\overline{f(x_j)})(f(x_j)-\overline{f(x_j)})'}}$ where $\overline{f(x_i)}=\frac{1}{n}\sum_{r=1}^{r=n}a_r(x_i)$ and $\overline{f(x_j)}=\frac{1}{n}\sum_{r=1}^{r=n}a_r(x_j)$

to be uniformly distributed in text resulting with a high occurrence frequency in the whole text/corpus. In this study, it is accepted that if a word is a stop word not only the total frequency but also the local frequency of the word must be higher compared to other words. The local frequency is the frequency within a given window size. For example, if the window size is set to 1000, for each word, frequency is calculated within a window of 500 on either side of the word. We believe that the window based frequency will not only discriminate stop words but it will also overcome the problem of varying corpus size in comparing the frequency based studies.

4.1.2. Collocative frequency (CF)

Collocations are groups of two or more lexical items that co-occur with a frequency greater than random probability [24]. The simplest method for finding two-worded collocations in a corpus is counting bigrams and selecting most frequently occurring bigrams. However, this method creates collocation candidate lists that involve also pairs of stop words [8]. When the lists are examined, it can be clearly seen that frequently occurring stop words are coupled with many other stop or non-stop words so they take place in several candidates. As a result, we considered that if a word is seen in many collocation candidates, there is a high probability that it is a stop word. We ranked the bigrams that occurred at least 4 times in text and calculate collocative frequency for any word by the list.

4.1.3. Document frequency (DF)

The words which occur in a variety of texts with different topics are accepted to be stop words in several methods such as tf-idf. In the study, the terms *text* and *document* will be used interchangeably. Hereby, the document frequency of a word is defined as the ratio of number of documents/texts in which the word occurs to the total number of documents/texts in the corpus.

4.1.4. Word length (WL)

Abiding by the least effort principle, frequently used words are shorter than less frequently used words. Due to the assumption of stop words to occur more than the content words, it may be stated that shorter words are more likely to be stop words. For this reason, the word length that is simply the number of characters in the word is accepted to be a discriminating feature for stop words.

4.1.5. Word position (WP)

Due to the grammatical rules to construct valid sentences, the words in a sentence are not arranged randomly. Though the word alignment rules vary in different languages, the position of a word in a sentence may still give a clue on the type of the word. For example, in Turkish, the main constituent of a sentence is the predicate. The predicate is the constituent that gives information; indicates some state or action, therefore it is impossible to construct a sentence without one. In a sentence where constituents are arranged according to standard grammar rules, the predicate is located at the end of the sentence with other members of the sentence arranged according to their degree of importance, with the closest to the predicate considered the most important. As a result, the information content of words in a sentence increases towards the end of the sentence. This enables to use the position of a word in a Turkish sentence as a discriminating feature. Statistics gathered from Turkish corpus used in the study support this idea. ~64.5% of final words of sentences in corpus is verbs and ~19.5% is nouns, strengthening the idea that the last word is likely to be a content word.

Similar grammatical rules about word alignment in a sentence also exist in other languages. In English; in statements, the subject generally precedes the predicate; in questions, the subject usually comes after the whole or part of predicate; in exclamations, the subject is occasionally placed after the predicate. Therefore, if not definitely, heuristically, we can say that predicates reside in mid-position in English sentences.

In the study, word position is the order of a word in a sentence that is normalized by the sentence length. For example in a 5-word sentence, the word position for the third word is calculated as 3/5.

4.2. The Training and Testing Corpora

Part of Brown corpus is utilized to construct the training and testing corpora for English. The Brown corpus was first compiled by Henry Kucera and W. Nelson Francis [25]. It involves sentence boundary tagged texts/documents in 15 different genres and part of speech tags in each text. Two texts from each genre are selected randomly to create English training corpus, of totally 60074 words and 3717 sentences. English testing corpus (16337 words, 868 sentences) is constructed from 8 texts that are from different genres but not involved in the training corpus. No preprocessing was performed on both corpora.

We have selected a Turkish corpus that has been previously compiled collaboratively by Sabancı and Middle East Technical University [26-27] as Turkish training corpus. The corpus (46532 words, 5665 sentences) comprises of documents from different genres such as scientific text, research, news. Morphological analysis of the corpus and annotation of sentence boundaries has been performed manually [26-27]. In this study, we used the surface-formed corpus without correcting other errors on part of speech tags and punctuation. Turkish testing corpus (13378 words, 889 sentences) is constructed from a group of eight texts that are randomly retrieved from web by queries on different topics.

The training phase of classification methods requires the labeling of stop and non-stop words on the training corpora. Being aware of the weaknesses, we decided to use part of speech information to assign words as stop or as non-stop in order to decrease the effort required for manual labeling. In the study, nouns, verbs and adjectives; which are considered in open class vocabulary, are regarded as non-stop words. And all the rest are considered as stop words.

4.3. Evaluation

The evaluation of stop word detecting methods may be based on precision, recall or combined measures that are commonly used in information retrieval. For stop word extraction, precision may be simply defined as the fraction of true stop words in the lists of stop word candidates. Recall is the proportion of retrieved true stop words over the whole set of true stop words in the corpus.

In the study, we used a combined measure, F-measure, which is simply a weighted average of the precision *p* and the recall *r*. The F-score reaches its best value at 1, worst score at 0 and it is formulized as follows.

$$F = 2 \cdot \frac{p \cdot r}{p + r} \tag{4}$$

We derived testing lists of stop words from best candidates that are offered by tf-idf and classification methods for both languages. For each language, firstly, all the words in testing corpus are listed according to the tf-idf values in descending order. Approximately first half of the words in the sorted list (8000 words for English and 6000 for Turkish) are unified to build the testing list of stop word candidates. The testing lists of stop word candidates are named as TFIDF_T for Turkish and TFIDF_E for English. Secondly, for each language, classification methods are applied on the regarding testing corpus. The words that are predicted as stop words by at least half of the classification methods are merged to build testing list of stop word candidates. The testing lists of stop word candidates that are built by classification methods are named as CL_T for Turkish and CL_E for English. Finally, four testing lists (TFIDF_T, TFIDF_E, CL_T, CL_E) are tagged (each candidate is labeled as stop word or non-stop word) based on their POS tags. Afterwards, the class (stop and non-stop) predictions of classification methods and tf-idf over testing lists are evaluated. During evaluation, by any classification method, if the same word appearing in the testing list is predicted to be in different classes in different occurrings in the corpus, the class (stop word or non-stop word) in which the word is predicted more is taken to be the class of the word. That is if the word “this” is classified 100 times as stop word and 10 times as non-stop word by Naïve Bayes method, it is accepted to be in stop word class by Naive Bayes since 100>10. To rank the words in the testing lists tf-idf scores are used.

5. RESULTS

In the study, as mentioned before, the stop word detection problem is accepted as a binary classification problem in which the words in a corpus are categorized as stop or non-stop words. The utility of the methods used in corresponding classification problem are assessed within the training corpus. In training, the classification accuracy is measured by the correct classification rate. Since a certain percentage of samples in any data set is expected to be correctly classified by chance, regardless of the classification model, to assess classification accuracy relative to chance, maximum chance criterion and proportional chance criterion are taken as the basis. Morrison [28] states that maximum chance criterion, $C_{\max}$, is the minimum expected correct classification for a selected group of interest and is measured simply assigning all samples to larger group; proportional chance criterion (C_{pro}) with two groups is

$$C_{pro} = p^2 + (1 - p)^2 \tag{5}$$

where p is the proportion of samples in first group and $1 - p$ is the proportion of samples in second group.

Table 2 gives the correct classification rates of competing methods together with maximum and proportional chance values obtained from training corpora. In order to observe the effect of aforementioned occurrence frequency based features (term and document frequency), the methods are employed with and without those features. Table 2 shows the classification rates on each corpus with and without term frequency and document frequency features employed. The correct classification rates of k-nn method are presented with an interval of minimum and maximum values obtained from k values varying from 1 to 5. For example, in Turkish corpus, correct classification rate of k-nn method using city block distance varies between minimum 93.15% and maximum 95.07% if all the features are employed in the classification. The correct classification rates of all methods for both corpora are higher than the maximum chance and the proportional chance criteria meaning that classification methods provide better classification rates than a random one. Another important outcome of the training is examined when the rates with and without the term and document frequency features are compared. Briefly, for both languages, it is observed that the absence of occurrence based features not drastically but sufficiently decreases the correct classification rates for all the methods.

Table 2. The correct classification rates (%) of classification methods in training corpora

	Turkish Training Corpus	**Turkish Training Corpus** (without TF & DF)	**English Training Corpus**	**English Training Corpus** (without TF & DF)
Linear Discriminant	86.28%	82.34%	85.60%	78.52%
Quadratic Discriminant	86.71%	84.39%	79.57%	75.02%
Naive Bayes	86.77%	84.65%	78.65%	76.09%
Decision Tree	94.94%	90.18%	97.09%	90.74%
Knn -City Block	[93.15 - 95.07%]	[86.27%-88.62%]	[93.06 - 98.62%]	[86.74%- 89.87%]
Knn-Euclidean	[92.97 - 95.07%]	[86.27%-88.63%]	[92.61 - 98.61%]	[86.57%- 89.87%]
Knn-Cosine	[91.90 - 94.92%]	[85.20%-87.90%]	[91.61 - 98.61%]	[86.14%- 89.19%]
Knn-Correlation	[91.56 - 94.96%]	[85.41%-87.16%]	[91.24 - 98.61%]	[82.25%- 88.98%]
$C_{\max}$	75.89%		50.10%	
C_{pro}	63.41%		50.00%	

The testing lists, TFIDF_T and TFIDF_E, derived by tf-idf method, include 523 and 105 unique candidates respectively. Since the inflectional variety disturbs the surface forms of words more in Turkish, the size of Turkish subset is significantly bigger than the size of English set as expected. The ratios of true stop words in corresponding subsets are 26.20% and 75.24% in order for Turkish and English.

Figures 1 and 2 depict the F-measure curves of the methods over TFIDF_T, TFIDF_E testing lists. The horizontal axis in the graphics is the ratio of completed stop word candidates to the total number of candidates. In Figures 1 and 2, the only k-nn method that generates the best F scores consistently is depicted in order to simplify the graphs. For example, in Figure 1, it is depicted that k-nn method using the city block distance with k=1 is performing better than the other k-nn alternatives.

As shown in Figure 1, the classification methods k-nn and linear discriminant analysis give consistently higher F-scores after the first quarter of TFIDF_T. On the other hand, on English testing list, TFIDF_E, tf-idf method generates similar F-scores with classification methods inferring that the features that are expected to improve the stop word detection are not sufficiently contributive.

When similar testing experiments are employed over the testing lists, CL_T and CL_E, the lists of 866 candidates for Turkish (18.24% are true stop words) and 408 candidates (36.52% are true stop words) for English, are obtained. F-measure graphs are given in Figure 3 and 4. The horizontal axis in the graphics is the ratio of completed candidates to the total number of candidates

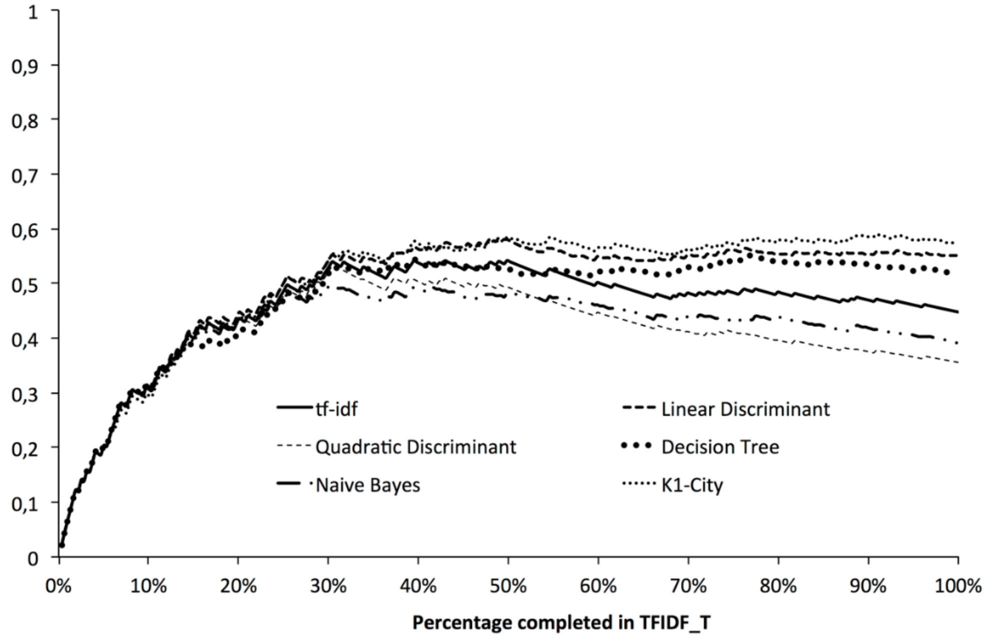

Figure 1. F-measure graph of Turkish testing list created by tf-idf method (TFIDF_T)

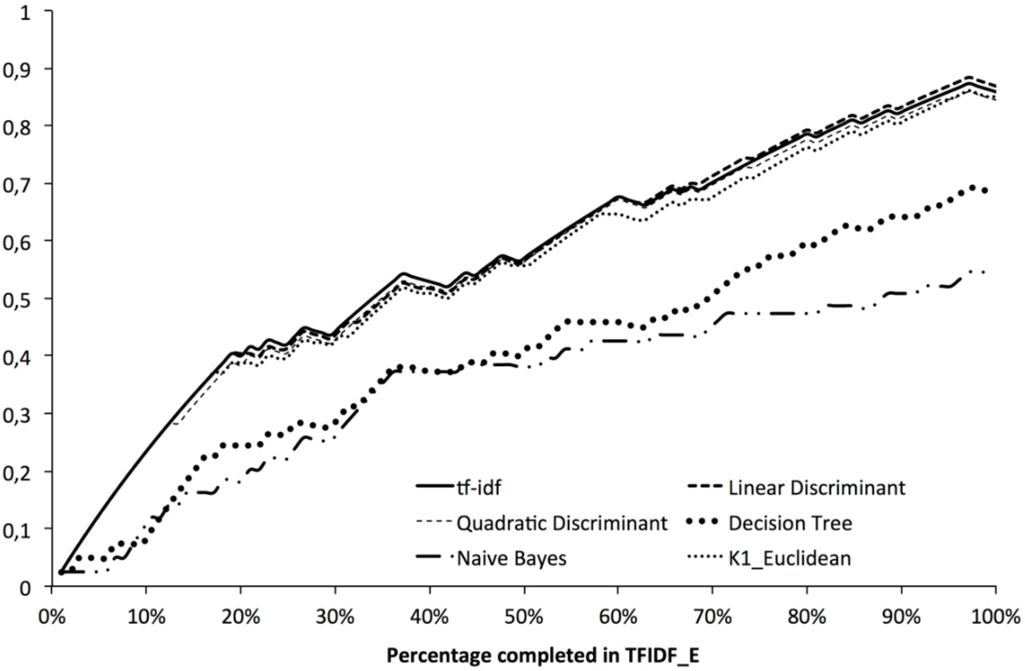

Figure 2. F-measure graph of English testing list created by tf-idf method (TFIDF_E)

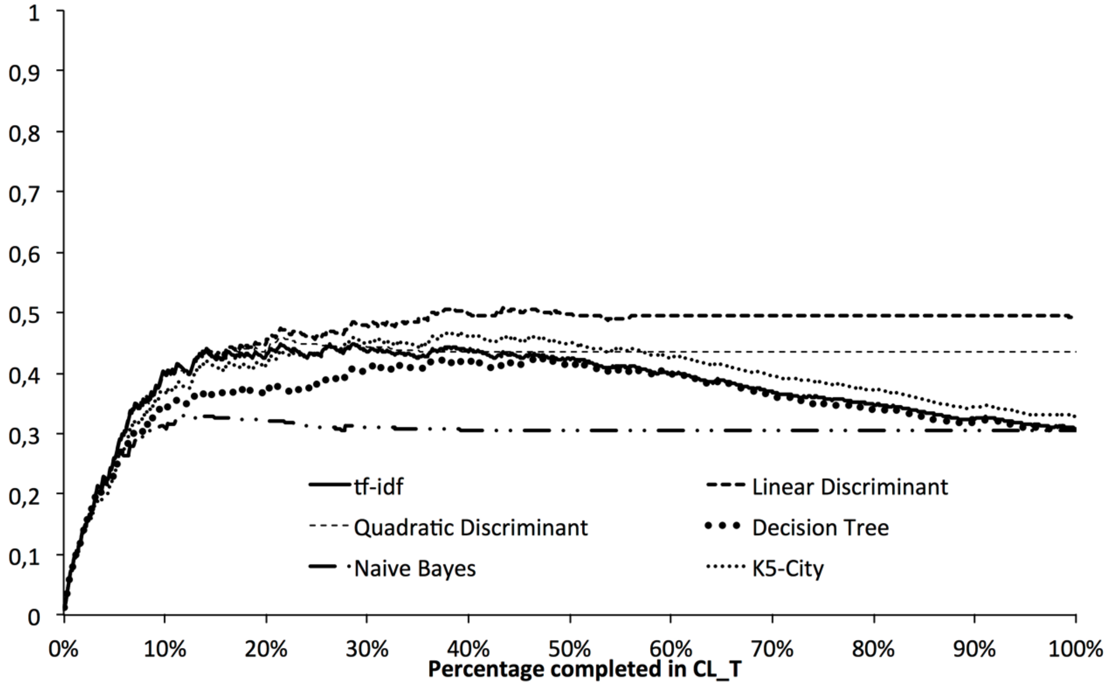

Figure 3. F-measure graph of Turkish testing list created by classification methods (CL_T)

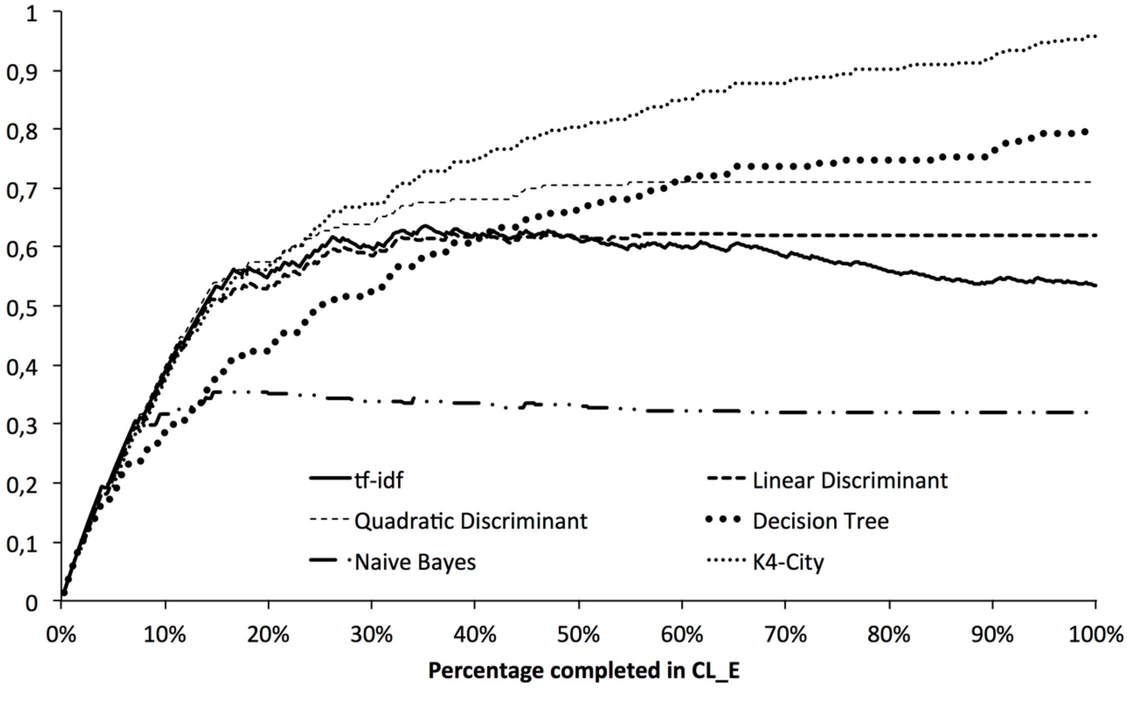

Figure 4. F-measure graph of English testing list created by classification methods (CL_E)

Over Turkish testing list CL_T, the linear and quadratic discriminant analyses perform better than tf-idf method almost for all proportions of the set. On the other hand, in English testing list CL_E, k-nn, quadratic discriminant analysis and decision tree algorithms are relatively successful in assigning words as stop or non-stop in the second half of the list.

One other important result which strengthens the basis of the study may be observed by examining Figures 1, 2, 3 and 4 all together. When F scores obtained by using the testing lists of English and Turkish are compared, it is seen that the scores are higher for English due to the lower inflectional noise in stop words. This also supports the dominancy of term frequency in stop word detection.

Table 3 gives the words from TFIDF and/or CL lists that are tagged as non-stop words due to the POS tagging but assigned as stop words by the whole set of classification methods. These words (in Table 3) may be accepted as the stop words proposed by the classification methods. Examining the words that are classified as stop words, it is seen that Turkish list includes much more words compared to English list. In addition, it is realized that both lists include different surface forms of frequently used verbs (e.g. "ol" in Turkish and "have" in English). On the other hand, in Table 3, there exist some words, such as "would", that are assigned as stop words by classification methods but tagged as non-stop due to POS tagging errors. As a result, it may be stated that classification methods also overcome the weakness of POS tagging.

Table 3. The stop words proposed by classification methods

Turkish	ağır	alan	aynı	az	biçimde	büyük	çeşitli	değil	devam	durum	eski
	gelen	geniş	gerçek	güçlü	gün	iki	iyi	kitap	küçük	madde	olan
	olarak	olduğu	on	önemli	sahip	son	sonra	söz	tek	tür	uzak
	var	yapmak	yaşam	yeni	yer	yerine	yok	zaman			
English	found	had	has	like	made	period	place	take	use	work	would

6. CONCLUSION

Even though occurrence frequency based features; term and document frequency; have been found sufficient to decide whether a word is a stop word or not in many studies, in this study, we propose to empower frequency based stop word detection by a combination of several other syntactical word features. The binary classification methods are employed to merge the features and are applied on two types of languages: Turkish and English.

In the study, the features that affect distinction between stop and non-stop words are taken as word position in sentence, word length, term frequency in a window, document frequency and collocative frequency. Experimental results showed that although the term frequency and document frequency are still contributive, there is not enough evidence to neglect other features in discrimination.

The evaluation is performed on two testing lists that are obtained by applying different approaches. The results show that the classification methods give more precise lists of stop words compared to tf-idf method on Turkish. The results though as not successful but support our initial claim of lower inflectional noise on English stop words.

In summary, we believe that the considered classification methods offer a way to merge different features that may contribute to stop word detection. The experiments showed that this claim is promising especially with languages that have inflectional variances on stop words.

REFERENCES

[1] Herdan G. Type-token mathematics. A textbook of mathematical linguistics. ‘S-Gravenhage, the Netherlands: Mouton, 1960.

[2] Dolamic L, Savoy J. When stopword lists make the difference. Journal of the American Society for Information Science and Technology 2009.

[3] Bond F, Shirai S. A Hybrid Rule and Example-based Method for Machine Translation. In M. Carl & A. Way (Eds.), Recent Advances in Example-based Translation 2003:211-224.

[4] Martínez-Santiago F, García-Cumbreras MA., Díaz-Galiano MC, Ureña LA. Using Machine Translation Resources with a Mixed 2-Step RSV Merging Algorithm. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, & B. Magnini (Eds.), Multilingual Information Access for Text, Speech and Images (SINAI at CLEF 2004); 2004; pp. 156-164; Heidelberg: Springer Berlin.

[5] Jiang X, Fan X , Wang Y, Jia K. Improving the Performance of Text Categorization Using Automatic Summarization. In S. S. Mahmoud, K. Jusoff and K. Li (Eds.), International Conference on Computer Modeling and Simulation (ICCMS); 2009; pp.347-351. Danvers

[6] Radev DR, Fan W. Automatic summarization of search engine hit lists. Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics; 2004; pp.99-109; Morristown: Association for Computational Linguistics.

[7] Witten IH, Moffat A, Bell TC. Managing Gigabytes: Compressing and Indexing Documents and Images. San Francisco: Morgan Kaufmann, 1999.

[8] Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. London: The MIT Press, 2000.

[9] Lazarinis F. Engineering and utilizing a stopword list in Greek web retrieval. Journal of the American Society for Information Science and Technology 2007;58(11): 1645–1652.

[10] Savoy J. A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science and Technology 1999; 50(10): 944-952.

[11] Can F, Koçberber S, Balçık E, Kaynak C, Öcalan HÇ, Vursavas O. Information Retrieval on Turkish Texts. Journal of the American Society for Information Science and Technology 2008; 59 (3): 407-421.

[12] Zou F, Wang FL, Deng X, Han S. Automatic identification of Chinese stop words. Research on Computing Science 2006; 18: 151–162.

[13] Tsz-Wai Lo, R., He, B. and Ounis, I. Automatically building a stopword list for an information retrieval system. Journal of Digital Information Management: Special Issue On The 5th Dutch-Belgian Information Retrieval Workshop; 2005; pp 3-8.

[14] Blanc JM, Dominey PF. Using prosodic information to discriminate between function and content words. In B. Bel & I. Marlien (Eds.), Speech Prosody; 2004;293-296. Lisbon: International Speech Communication Association.

[15] Hao L, Hao L. Automatic Identification of Stop Words in Chinese Text Classification. In H. Zhou (Ed.), Proceedings of the 2008 International Conference on Computer Science and Software Engineering (CSSE); 2008; pp. 718-722. Danvers: IEEE Computer Society Conference Publishing Services.

[16] Sinka MP, Corne DW. Towards modernised and Web-specific stoplists for Web document analysis. Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence; 2003; pp. 396–402. Washington: IEEE Computer Society.

[17] Yaghoub-Zadeh-Fard, M. A., Minaei-Bidgoli, B., Rahmani, S., Shahrivari. S. PSWG: An automatic stop-word list generator for Persian information retrieval systems based on similarity function & POS information, in Proceedings of 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran; 2015; pp. 111-117.

[18] Z. Yao Z., Ze-wen, C. Research on the Construction and Filter Method of Stop-word List in Text Preprocessing. In Proceedings of 2011 Fourth International Conference on Intelligent Computation Technology and Automation, Shenzhen, Guangdong; 2011; pp. 217-221.

[19] Saif H, Feranandez M, Alani H. Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter, in Proceedings of ISWC 2014 Posters & Demonstrations Track; 2014; pp. 281-284

[20] Overall JE, Klett C. J. Applied Multivariate Analysis. McGraw Hill Book Company. 1972.

[21] Fisher, R. A. The Use of Multiple Measurements In Taxonomic Problems. Annals of Eugenics, 7: 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. 1936

[22] Mitchell T. Machine Learning. McGraw Hill, 1997.

[23] Fix E, Hodges J. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report No. 4, Project No. 21-49-004, 1951; School of Aviation Medicine, Randolph Field, TX.

[24] Hoey M. Pattern of Lexis in Text. Oxford: Oxford University Press. 1991

[25] Francis WN, Kucera H. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Providence, Rhode Island: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

[26] Oflazer K, Say B, Hakkani-Tür, DZ, Tür G. Building a Turkish Treebank, Invited chapter in Building and Exploiting Syntactically-annotated Corpora, 2003; Anne Abeille Editor, Kluwer Academic Publishers.

[27] Atalay NB, Oflazer K, Say, B. The Annotation Process in the Turkish Treebank, in Proceedings of the EACL Workshop on Linguistically Interpreted Corpora – LINC; 2003; Budapest, Hungary.

[28] Morrison DG. On Interpretation in Discriminant Analvsis. Journal of Marketing Research 1969; 6(2): 156-163.