Classifying Words: A Syllables-based Model

P. Warintarawej, A. Laurent, P. Pompidor  
LIIRMM - Univ. Montpellier 2  
CNRS UMR 5506 - Montpellier  
France  
{warintarawej,laurent,pompidor}@lirmm.fr  

A. Cassanas, B. Laurent  
Namae Concept  
Montpellier  
France  
{a.cassanas,b.laurent}@namaeconcept.com

Abstract—Text classification has been extensively studied by linguists and computer scientists. However, there are very few works on classification of words into classes or concepts (e.g., thesaurus). In this paper, we consider this topic, especially in the context of the classification of names like brand names or neologisms. The challenge is thus to provide automated tools to analyze new names by classifying them into concepts. Then, for example, a naming company customer can be informed about which concept a new name is closest to. As we argue that a word can belong to several concepts, we propose to consider the top-k classification approach. Moreover, we rely on syllables to build the classification model. The word corpus is collected from French thesaurus. All labeled-words are separated into syllables. Feature selection techniques are used to select discriminative syllables. We use a syllables frequency (SF) and mutual information (MI) performing with Naive Bayes classifier and K-nearest neighbor (KNN). Instead of selecting only one class, the model select top-k classes ranking them by a classifier score. The result shows the top-k classification model helps to analyze a new word by showing that it can be related to more than one concept. Moreover, the set of discriminative syllables can be used to explain the classification results which makes the results more meaningful.

Keywords—Text Classification; Words Classification; Feature Selection; Syllables; Discriminative Features.

I. INTRODUCTION

We consider words from dictionary or thesaurus, which are organized into groups or classes depending on their meaning. A thesaurus provides a semantic road map to individual fields and the relationships among fields. It maps out a concept space, relates concepts to terms [1]. The task has been done by linguists. Meanwhile the emerging of new names are increasing such as brand names which are created from various forms, i.e., they can be taken from dictionaries or be a mix of existing words and become neologisms [2].

Text classification is a supervised learning involved with the process of learning from examples (the so-called training set) and using their evidences to classify unseen texts into predefined-classes. The task has been accomplished using the context of the document. In other words, the keywords which appear in that document actually refer to the class they should belong to. For classifying words into classes (to match thesaurus, we will use the word “concept" instead), we consider the words syllables as the document keywords. We assume that some syllables can convey some meanings that relate to some concepts.

Due to the fact that a new word is often the result of a mix of existing words, it can refer to several concepts. Then, it is interesting for linguists and name creators to know and validate which concepts they are related to. We propose top-k classes approach for words classification. Most of classifiers can produce a numeric score for judging which class is the best, such as Naive Bayes, KNN, Decision Tree etc [3]. Instead of selecting only one class, as usual, we select top-k classes ranking them with a classifier score. For doing text classification, feature selection techniques need to reduce the dimensionality of features with no information and select only the most relevant features. Rather than using all features from training words, feature selection can perform more discriminatingly if knowing which syllables have to be provided into classification models.

In this paper, we first introduce the main idea of syllables-based model for words classification (section II). Our feature selection methods are described in section II-A: i.e. syllables frequency (SF) and mutual information (MI). In section II-B, we present Naive Bayes classifier and K-nearest neighbor (KNN) for top-k classification. In section III, we show the experiment and the result of Naive Bayes and KNN. We describe the effect of feature selection techniques and number of k neighbors. In section IV, we conclude on the results and look forward to future works that would give more semantically discriminative results for words classification.

II. WORDS CLASSIFICATION SYLLABLES-BASED MODEL

In our context, words classification is a text classification model, unlike words classification in a grammatical point of view (parts of speech) which aims at classifying words into predefined-classes. Segmentation of words can be made by cutting sequences of letters (n-grams) but the result gives a large set of features. Moreover some of them have no or poor meaning. Instead of using sequences of letters, we propose syllables-based representation. Syllables-based representation gives a shorter list of results than letters-based, but more importantly, it captures meaning: i.e. it seems to be related to the words semantics [4]. The syllabification algorithm is a
Most of text classification models treat a classified item as feature vector. A typical way to transform a classified item into feature vector is to use “bag-of-words”. We use text classification models to classify words into concepts based on their syllables. Generally, a bag of words is used to represent a document; in our case, we represent a word by its syllables. Another method would be to use n-grams. However, there is no easy and unique way to cut words into syllables.

To represent words as syllabled-based models, each word \( w \) is represented as a vector of length \( |S| \), where \( |S| \) is the size of the syllables from training set.

Let us define word

\[
  w = < s_1(w), s_2(w), s_3(w), ..., s_{|S|}(w) >
\]

where \( s_i(w) \) is the binary weight of the \( i \)th syllable; 1 if the syllable appears in the word and 0 otherwise.

A. Feature Selection

Feature selection is the process by which a subset of all features is selected from training set. In text classification, where the problem of huge dimensions of text data is encountered, a feature selection technique is crucial. Methods have been proposed to select the most relevant attributes and to improve classification effectiveness and computation efficiency [5]. To discriminate syllables for each concept, we consider two feature selection techniques. The first one is a syllable frequency (SF), based on document frequency (DF): the frequency of syllables is counted in each concept for computing the ranking. The second one uses mutual information (MI). There are many successful works in text classification relying on MI [6], [7]. MI measures how much information the presence/absence of a term gives to the correct classification decision.

Formally [8]:

\[
I(U; C) = \frac{N_{11}}{N} \log \frac{N N_{11}}{N_{1}N_{1}} + \frac{N_{01}}{N} \log \frac{N N_{01}}{N_{0}N_{1}} + \frac{N_{10}}{N} \log \frac{N N_{10}}{N_{1}N_{0}} + \frac{N_{00}}{N} \log \frac{N N_{00}}{N_{0}N_{0}}
\]

where the \( N_{10} \) is the number of words that contain syllable \( t \) and not in class \( c \) etc. \( N_{1} = N_{10} + N_{11} \) is the number of words that contain syllable \( t \), \( N = N_{00} + N_{01} + N_{10} + N_{11} \) is the total number of words in domain.

Figure 1 shows the example of discriminative syllables of concept “Nouveauté” by MI.

![Figure 1. The discriminative syllables of concept “Nouveauté” by MI.](image)

B. Top-k Classification by Naive Bayes and KNN

The Naive Bayes classifier is a simple classifier model that is based on a probabilistic theory called the Bayesian theorem including an independence assumption called Naive Bayes assumption. The assumption assumes that all attributes of the example are independent. In fact this assumption is not correct in real-world text classification, even though Naive Bayes performs classification very well [9]. The probability model for the classifier is a conditional probability model \( P(c_j|d_i) \) when \( c_j \) is a member of the set of classes \( C \) in classification domain and \( d_i \) is a testing document. The probabilistic of a document \( d_i \) being in class \( c_j \) by Bayes theorem is:

\[
P(c_j|d_i) = \frac{P(c_j)P(d_i|c_j)}{P(d_i)}
\]

In our task, words classification follows Naive Bayes formular:

\[
P(c_j|w_i) = \frac{P(c_j)P(w_i|c_j)}{P(w_i)}
\]

where word \( w_i \) is a vector space that consists of syllables occurring in a word as defined in section II.
To classify a word into top-k classes, we rank \( P(c_j|w_i) \) and select top-k classes.

For KNN, we compute similarity between an instance word \((w_i)\) and words \((w_j)\) from training set; we use the CosSim function [10] which is particularly simple using the binary term weight approach:

\[
\text{CosSim}(w_i, w_j) = \frac{D}{\sqrt{A + B}}
\]  

Where \( D \) is the number of syllables that an instance word \((w_i)\) and a training word \((w_j)\) have in common, \( A \) is the number of syllables in an instance word \((w_i)\) and \( B \) is the number of syllables in a training word \((w_j)\). To select \( k \) neighbors of a word \( w_i \), we calculate the confidence of a word \((w_i)\) belonging to a class \((c)\) as:

\[
\text{confidence}(c, w_i) = \frac{\sum_{k, t \in K|C(class(k,t)=c)} \text{Sim}(k, t, w_i)}{\sum_{k \in K} \text{Sim}(k, w_i)}
\]  

Where \( \text{Sim} \) is the similarity function as we define in (1). To compute the confidence of a word \( w_i \) belonging to a class \( c \), for each \( k \) neighbors of a word \( w_i \), the confidence is the summation of their similarities to words in class \( c \) and divide by the summation of all similarities of \( k \) neighbors with regard to \( w_i \).

After computing all confidence scores of \( w_i \) belonging to each class \( c \), we rank the confidence scores and select top-k classes to be the answer of classification models.

III. EXPERIMENT AND RESULTS

A. The corpus

The word corpus has been collected from 2 sources that are French Larousse thesaurus and JeuxDeMots[11] website: http://www.lirmm.fr/jeuxdemots. The word collection processing was divided into 2 steps. First, from French Larousse thesaurus which is organized with words in categories called concepts, the thesaurus contains 70,201 words from 873 concepts. In a second step, to enrich the concepts, we retrieve the related words from JDM corresponding to every concept name. JeuxDeMots is a free game which allows players to propose words related to a provided term. The purpose of this project is to construct French vocabulary semantic groups. After the collecting process, the total words in corpus is equal 229,855. The experiment is performed on 20 concepts (Table II) chosen by linguist, the syllabification algorithm is used to divide words into syllables, and then we remove stopwords from Snowball French stopword list[12]. The final data set for testing contains 8,961 words and 3,605 syllables. 10-fold cross validation is run to evaluate Naive Bayes and KNN.

B. Naive Bayes

Feature selection techniques; SF and MI were considered. We selected 100, 500, 1000 and 1500 syllables testing with Naive Bayes classifier. The result of Naive Bayes classification by selecting top-3 classes is displayed in Table III.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>#Num of features</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>100</td>
<td>72.57</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>75.20</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>74.37</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>72.88</td>
</tr>
<tr>
<td>SF</td>
<td>100</td>
<td>71.62</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>76.54</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>77.22</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>73.70</td>
</tr>
</tbody>
</table>

Table III
EXPERIMENT RESULTS: CLASSIFICATION ACCURACY BY TOP-3 CLASSES OF NAIVE BAYES CLASSIFIER WITH VARIOUS #NUM OF FEATURES.

The result of Naive Bayes shows that the syllable frequency (SF) with 1000 syllables achieves the best accuracy score at 77.22%. When comparing between feature selection techniques (Figure 2), syllable frequency (SF) performs better than mutual information (MI) on average. For both feature selection techniques, the accuracy slowly increases from small numbers of features (100) and gradually drops after 1000 features. When comparing between the size of features, the accuracy average of 500 syllables achieves the highest percentage than other sizes.

C. KNN

The result of KNN classification is displayed in Table IV. For KNN, the comparison task between test word and train word uses the CosSim function as in Equation (1). We do not use feature selection for KNN, the reason is to control member of the testing set to be the same in every round of the experiment. Due to sizes of the feature sets are different. For example, the small size of feature set (100), the more words in testing set will be ignored. To compare the accuracy between size of \( k \) neighbors, we took all syllables
into account for each comparing of pair words. We run classification testing varying with numbers of k neighbors by computing the confidence of words against every class. The result from confidence scores were ranked and top-3 classes were selected to be the answer of classification. The classification accuracy results is displayed in Table IV. The result reports the best classification accuracy obtained by k = 60 with 94.99%, we further tested running with increasing numbers of k neighbors but the accuracy does not change dramatically. The result shows that larger the number of k, the higher the accuracy (see Figure 3).

D. Top-k classification obtained more related concepts

Top-k classification models help to analyze a new word by showing which concepts it is more likely to be related to. For instance, a word “goûcolat” is a brand name made of the combination of 2 words: “goût” (taste) and “chocolat” (chocolate). In French, “goût” means “taste” and is a part of the word “goûter” meaning “afternoon snack”. The analyze of “goûcolat” with KNN by selecting top-4 classes shows that indeed “goûcolat” have the most similarities to “chocolat” and “goût” and then refers to the concepts “Boisson” (Beverage), “Sucrerie” (Candy), “Pain” (Bread) and “Goût” (Taste) see in Figure 4.

The result helps the user to perceive the meanings and evocations of the word (that can relate to more than one concept) and shows the most relevant concepts depending on the syllables the words contain. If the concept contains words that have a lot of common syllables with the new word, the confidence value will be large because it was computed from summation of CosSim of all words from the concept in set of k neighbors divided by the summation of CosSim of all words in set of k neighbors.

E. Syllables make more meaningful results

For analyzing a new brand name, the user needs to know how the new name can refer to predicted-concepts. The important issue is that the user needs meaningful explanation. Syllable-based model can serve this purpose. For example, the word “nouvalia” that can be used to name an exposition center for all the new objects of the year, the result from Naive Bayes says “nouvalia” belongs to concept “Nouveauté”. We explain the result by syllables-based model; the word “nouvalia” contains the syllables “nou” and “va’ which are the members of the set of discriminative syllables of concept “Nouveauté”. The model helps to track back the predicted classes by using discriminative syllables. In this example, the result thus gives more meaningful tips to perceive which concepts it can refer to, as shown by Figure 1.
the multiple evocations a word has. For example, “goûcolat” belongs to the concept “Boisson”, “Sucrerie”, “Pain” and “Goût” respectively ranking by confidence scores.

Future works include works on how to find the lexemes of words by using their syllables. Although some syllables have meaning, lexemes are considered by linguists to be more definite and certain source of meaning: a lexeme is the minimal set of letters containing the meaning of a word. We will study a way to find lexemes based on syllables. Instead of using syllables in classification model, lexemes will thus be used as a feature set.

REFERENCES


<table>
<thead>
<tr>
<th>Concept</th>
<th>Words</th>
<th>Total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boisson</td>
<td>chocolat, cho, co, lat</td>
<td>6</td>
</tr>
<tr>
<td>Sucrerie</td>
<td>chocolat, cho, co, lat</td>
<td>5</td>
</tr>
<tr>
<td>Pain</td>
<td>chocolat, cho, co, lat</td>
<td>4</td>
</tr>
<tr>
<td>Goût</td>
<td>goûter, goûteur</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept</th>
<th>Confidence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boisson</td>
<td>58.52</td>
</tr>
<tr>
<td>Sucrerie</td>
<td>51.62</td>
</tr>
<tr>
<td>Pain</td>
<td>42.77</td>
</tr>
<tr>
<td>Goût</td>
<td>20.82</td>
</tr>
</tbody>
</table>