Classifying Words: A Syllables-based Model

Pattaraporn Warintarawej, Anne Laurent, Pierre Pompidor, Armelle Cassanas, Bénédicte Laurent

Classifying Words: A Syllables-based Model

P. Warintarawej, A. Laurent, P. Pompidor
LIRMM - Univ. Montpellier 2
CNRS UMR 5506 - Montpellier
France
{*warintaraw,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 jugding 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

copyright reserved of *Namae Concept company*[1] (we can not explain in detail). The syllabification process has been done as pre-processing so that it does not take too much time out of the classification phase itself. The example of the syllabification process in concept *"Nouveauté"* (Novelty) is shown in Table I.

Word	Syllables
nouveauté	_nou / veau / té_
nouvelle	_nou / vel / le_
nouvellement	_nou / vel / le / ment_
nouvelleté	_nou / vel / le / té_
nova	_no / va_
novale	_no / va / le_
novateur	_no / va / teur_
jeune	_jeu / ne_
jeunement	_jeu / ne / ment_
jeunesse	_jeu / nes / se_

Table I
AN EXAMPLE OF SYLLABIFICATION PROCESS OF WORDS IN CONCEPT "NOUVEAUTÉ".

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 woud 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 $s_1(w)$ is the binary weight of the ith 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 sucessful 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] :

$$\begin{aligned} I(U;C) &= \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} \\ &+ \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}} \end{aligned}$$

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.

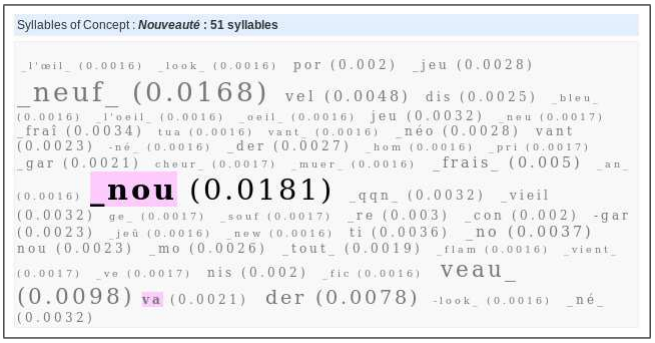


Figure 1. The discriminative syllables of concept "Nouveauté' by MI.

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 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.

[1]Namae Concept company is the naming company in France, http://www.namaeconcept.com

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_t) from training set; we use the CosSim function [10] which is particularly simple using the binary term weight approach:

$$CosSim(w_i, w_t) = \frac{D}{\sqrt{A * B}} \quad (1)$$

Where D is the number of syllables that an instance word (w_i) and a training word (w_t) 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_t). To select k neighbors of a word w_i, we calculate the confidence of a word (w_i) belonging to a class (c) as:

$$confidence(c, w_i) = \frac{\sum_{k_i\prime \in K|(Class(k_i\prime)=c)} Sim(k_i\prime, w_i)}{\sum_{k_i \in K} Sim(k_i, w_i)} \quad (2)$$

Where 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 coresponding 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.

Concept	#Num of words	Concept	#Num of words
Éventualité	138	Violence	355
Saisons	82	Distinction	168
Nouveauté	169	Droit	3,065
Humidité	195	Figures de discours	128
Terre	477	Architecture	1,539
Soleil	369	Posie	378
Lichens	52	Pain	325
Reptiles	124	Sucrerie	274
Goût	196	Boisson	595
Effort	163	Mode	169

Table II
THE 20 CONCEPTS FROM FRENCH LAROUSSE THESAURUS FOR THE EXPERIMENT.

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.

Feature Selection	#Num of features	Accuracy (%)
MI	100	72.57
	500	75.50
	1000	74.37
	1500	72.88
SF	100	71.62
	500	76.54
	1000	77.22
	1500	75.70

Table III
EXPERIMENT RESULTS: CLASSIFICTION 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) performes 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

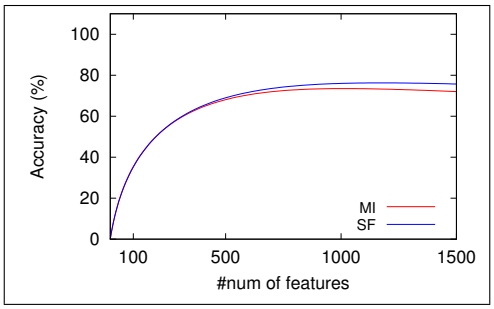


Figure 2. The classification accuracy of Naive Classifier comparing between MI and SF.

into account for each comparing of pair words. We run classification testing varying with numbers of k neighbors by computing the confidence of words againts 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.

#Num of k	Accuracy (%)
10	85.36
20	90.60
30	92.49
40	93.64
50	94.47
60	94.99

Table IV
EXPERIMENT RESULTS: CLASSIFICTION ACCURACY BY TOP 3 CLASSES OF KNN WITH VARIOUS #NUM K NEIGHBORS.

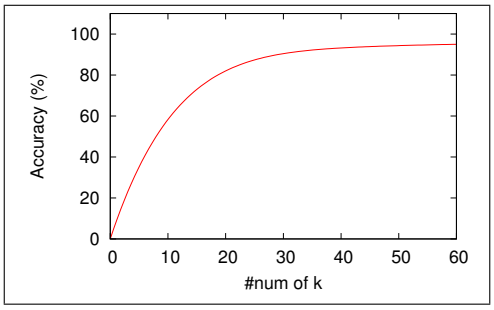


Figure 3. The accuracy percentage of KNN various by #num of k neighbors.

We run testing with various numbers of k neighbors, 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 similaries 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

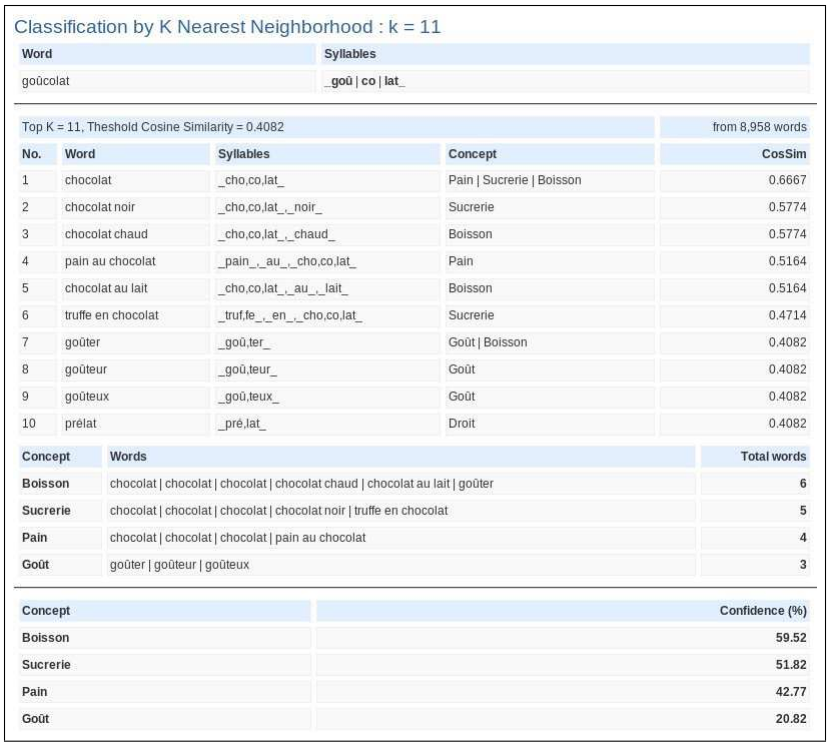
For analyzing a new brand name, the user needs to know how the new name can refer to predicted-concepts. The important issue is that the user needs meaningful explanation. Syllable-based model can serve this purpose. For example, the word *"nouvalia"* that can be used to name an exposition center for 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.

IV. CONCLUSION

In this paper, we address the word classification topic and we show that KNN performes better than Naive Bayes in this framework when considering syllables. This can be explained by the fact that KNN does not consider the probability of classes when comparing words based on their syllables. Meanwhile, when considering Naive Bayes, if the value of $P(w_i|c_j)$ is not enough to overcome $P(c_j)$ value, the class which contains a large number of words will get high score of $P(c_j|w_i)$.

We argue that syllables that appear more often in the class play an important role in the classification model, leading to the fact that syllable frequency (SF) performs better than mutual information (MI) when considering the Naive Bayes classification model.

Finally, by considering top-k approach, the classification accuracy increases and more valuable results are provided to the user. It is indeed important for the user to be aware of



Classification by K Nearest Neighborhood : k = 11

Word	Syllables
goûcolat	_goû | co | lat_

Top K = 11, Theshold Cosine Similarity = 0.4082 — from 8,958 words

No.	Word	Syllables	Concept	CosSim
1	chocolat	_cho,co,lat_	Pain | Sucrerie | Boisson	0.6667
2	chocolat noir	_cho,co,lat_,_noir_	Sucrerie	0.5774
3	chocolat chaud	_cho,co,lat_,_chaud_	Boisson	0.5774
4	pain au chocolat	_pain_,_au_,_cho,co,lat_	Pain	0.5164
5	chocolat au lait	_cho,co,lat_,_au_,_lait_	Boisson	0.5164
6	truffe en chocolat	_truf,fe_,_en_,_cho,co,lat_	Sucrerie	0.4714
7	goûter	_goû,ter_	Goût | Boisson	0.4082
8	goûteur	_goû,teur_	Goût	0.4082
9	goûteux	_goû,teux_	Goût	0.4082
10	prélat	_pré,lat_	Droit	0.4082

Concept	Words	Total words
Boisson	chocolat | chocolat | chocolat | chocolat chaud | chocolat au lait | goûter	6
Sucrerie	chocolat | chocolat | chocolat | chocolat noir | truffe en chocolat	5
Pain	chocolat | chocolat | chocolat | pain au chocolat	4
Goût	goûter | goûteur | goûteux	3

Concept	Confidence (%)
Boisson	59.52
Sucrerie	51.82
Pain	42.77
Goût	20.82

Figure 4. The result of KNN top-4 classes with k=11 for a new word *"goûcolat"*

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

[1] D. Soergel, "Functions of a thesaurus / classification / ontological knowledge base," October 1997.

[2] A. Laurent, B. Laurent, D. Brouillet, S. Martin, and M. Roche, "Embedding emotions within automatically generated brand names," in *Proc. of the Int. Conference on Kansei Engineering and Emotional Research*, 2010.

[3] C. X. Ling, J. Huang, and H. Zhang, "Auc: a statistically consistent and more discriminating measure than accuracy," in *Proceedings of the 18th international joint conference on Artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 519–524.

[4] H. Tschach, "Syllables and other string kernel extensions," in *In Proceedings of 19 th International Conference on Machine Learning*. Morgan Kaufmann, 2002, pp. 530–537.

[5] F. George, *Feature Selection for Text Classification*, ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, 2007, pp. 257–276.

[6] A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng, "Improving text classification by shrinkage in a hierarchy of classes," in *Proc. of the int. conf. on Machine Learning*, 1998, pp. 359–367.

[7] S. T. Dumais and H. Chen, "Hierarchical classification of web content," in *SIGIR*, 2000, pp. 256–263.

[8] I. C. Mogotsi, "Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval," *Information Retrieval*, vol. 13, pp. 252–253, 2010.

[9] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, 1997.

[10] P. Soucy and G. W. Mineau, "A simple knn algorithm for text categorization," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, ser. ICDM '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 647–648.

[11] L. Mathieu and Z. Virginie, "Jeuxdemots and pticlic: games for vocabulary assessment and lexical acquisition," in *In proc of Computer Games, Multimedia & Allied technology 09 (CGAT'09)*, May 2010.

[12] S. Project. [Online]. Available: http://snowball.tartarus.org/algorithms/french/stop.txt