Classifying Words: A Syllables-based Model

8th International Workshop on Text-based Information Retrieval - TIR ’11 [DEXA 2011]

Pattaraporn Warintarawej[1], Anne Laurent[1], Pierre Pompidor[1], Armelle Cassanas[2], Bénédicte Laurent[2]

[1]Lirmm, University Montpellier 2

[2]Namae Concept

August,31 2011

Outline

1. Motivations
2. Syllabification
3. Words Classification: A syllables-based model
4. Top-k Classification
5. Experiment and Result
6. Conclusion and Future work

Plan

1. Motivations
2. Syllabification
3. Words Classification: A syllables-based model
4. Top-k Classification
5. Experiment and Result
6. Conclusion and Future work

Motivations

Brand names

VIVENDI NOUVALIA GOÛCOLAT AUREA
SEVEANE LYLIA SOLÉA EVARANDA ECONOVISTA

- Linguistics demands:
  - The liguists create new names regarding to business requirement
  - Methods to automatically analyse new names by saying which concepts they are related to

Motivations

Brand names

VIVENDI NOUVALIA GOÛCOLAT AUREA
SEVEANE LYLIA SOLÉA EVARANDA ECONOVISTA

- Linguistics demands:
  - The liguists create new names regarding to business requirement
  - Methods to automatically analyse new names by saying which concepts they are related to
- Syllabification approach:
  - To retrieve syllable boundaries in words
  - Takes syllables into account for analysing a new name

Motivations

Brand names

VIVENDI NOUVALIA GOÛCOLAT AUREA
SEVEANE LYLIA SOLÉA EVARANDA ECONOVISTA

- Linguistics demands:
  - The liguists create new names regarding to business requirement
  - Methods to automatically analyse new names by saying which concepts they are related to
- Syllabification approach:
  - To retrieve syllable boundaries in words
  - Takes syllables into account for analysing a new name

Motivations

Brand names

VIVENDI NOUVALIA GOÛCOLAT AUREA
SEVEANE LYLIA SOLÉA EVARANDA ECONOVISTA

- Linguistics demands:
  - The liguists create new names regarding to business requirement
  - Methods to automatically analyse new names by saying which concepts they are related to
- Syllabification approach:
  - To retrieve syllable boundaries in words
  - Takes syllables into account for analysing a new name

Text classification + Bag-of-syllables => Classifying Words: A Syllables-based Model

Plan

1 Motivations

2 Syllabification

3 Words Classification: A syllables-based model

4 Top-k Classification

5 Experiment and Result

6 Conclusion and Future work

Syllabification concept

- Syllabification (in french Syllabation):
  - Syllabification is the separation of a word into syllables
  - The syllabifier was created applying “Rule-based framework”,*from Namae Concept Company*
  - Syllabification algorithm implements the predefined rules to separate word

Syllabification concept

- Syllabification (in french Syllabation):
  - Syllabification is the separation of a word into syllables
  - The syllabifier was created applying “Rule-based framework”,*from Namae Concept Company*
  - Syllabification algorithm implements the predefined rules to separate word
- The example rule:
  - VCCV => V-CCV when V = any vowel, CC = either PH,CH,TH or GN
  - Ex. résignation => ré-si-gna-tion, marcher => mar-cher

Syllabification process

- The algorithm scans the word from left to right and reaches the second vowel to find the boundary of the first cut according to the syllabification rules
- The process goes on till the last letter is reached
- The algorithm performs recursively

Syllabification process

- The algorithm scans the word from left to right and reaches the second vowel to find the boundary of the first cut according to the syllabification rules
- The process goes on till the last letter is reached
- The algorithm performs recursively

Syllabification process of the word : nouvalia

Round	Curent stream	result syllable	Next stream	Rule
1	nouvalia	nou	valia	VCV => V-CV
2	valia	va	lia	VCV => V-CV
3	lia	lia	-	keep vowels together at the end of words

Plan

1. Motivations
2. Syllabification
3. Words Classification: A syllables-based model
4. Top-k Classification
5. Experiment and Result
6. Conclusion and Future work

Syllables-based model

Words Classification function: a syllables-based model

- Words classification is based-on a text classification model
- Let's define the function as:

$$c = f(w)$$

when c is the predefined concept, w is the word to classify

- To represent words as a syllables-based model, each word w is represented as a vector of weights length $|S|$,where $|S|$ is the number of syllables in domain
- Let's define a word as:

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

where $s_i(w)$ is the binary weight of the i^{th} syllable; 1 if the syllable appears in the word, 0 otherwise.

Feature Selection

- High dimensionality of the feature space
- Most of these features are not relevant and can slow down the classification process
- Feature selection is commonly used to reduce the dimensionality of feature space and improve the efficiency of classifier
- We propose Syllable frequency (SF) and Mutual Information (MI) for feature selection

Feature Selection

- *Syllable frequency (SF)*: the simple weightening of features calculate by its frequency in a class
- *Mutual Information (MI)*: the weight of feature represents the dependency of that feature in the regarding class

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

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 is the total number of words in domain.

Naive Bayes Classifier

Naive Bayes Classifier

- The multi-variate Bernoulli Event Model
- Given a word w_i, the probability of each class c_j is calculated as

$$P(c_j|w_i) = \frac{P(c_j)P(w_i|c_j)}{P(w_i)}$$

- where a set of syllables S is given from feature selection
- a word w_i is represented with a vector of $|S|$ dimensions as

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

- $P(w_i|c_j)$ can be calulated under the Naive Bayes assumption as:

$$P(w_i|c_j) = \prod_{1 \leq k \leq |S|} P(s_k|c_j)^{(s_k(w))}(1 - P(s_k|c_j))^{(1-s_k(w))}$$

KNN Classifier

KNN Classifier

- **Step 1**: Calculate the similarity between a testing word (w_i) and a word (w_t) in domain, define by CosSim function as:

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

Where D is the number of syllables that a testing word (w_i) and a word in domain (w_t) have in common A is the number of syllables in a testing word (w_i) and B is the number of syllables in a word in domain (w_t).

- **Step 2**: Select k neighbors of w_i by ranking the similarity values
- **Step 3**: Calculate the confidence of a word (w_i) belonging to a class (c) as:

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

Where *Sim* is the CosSim function

Plan

1. Motivations
2. Syllabification
3. Words Classification: A syllables-based model
4. Top-k Classification
5. Experiment and Result
6. Conclusion and Future work

Top-k Classification

- The idea of Top-k classification is to select more than one class for classification result
- Both of *NaiveBayes* and *KNN* produce the score to measure how much the word belongs to the class
- Ranking top scores from classifier and selecting k classes

Plan

1. Motivations
2. Syllabification
3. Words Classification: A syllables-based model
4. Top-k Classification
5. Experiment and Result
6. Conclusion and Future work

Experiment and Result

- The corpus

Concept(Class)	#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	Poésie	378
Lichens	52	Pain	325
Reptiles	124	Sucrerie	274
Goût	196	Boisson	595
Effort	163	Mode	169

- Collect words from French Larousse thesaurus and JeuxDeMots [M.Lafourcade]
- Select 20 concepts containing 8,961 words and 3,605 syllables. (after removing stopwords)
- Evaluate Naive Bayes and KNN by 10-fold cross validation

Naive Bayes result

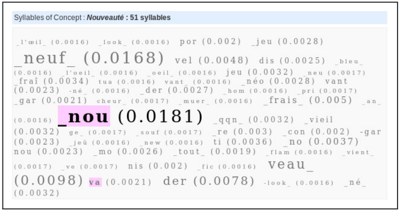
- *SF* and *MI* were considered as 100, 500, 1000 and 1500 syllables
- Experiment Results: Classification Accuracy by Top-3 classes of Naive Bayes Classifier with various #num of features

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

Naive Bayes Example result

Syllables make more meaningful results

- User needs meaningful explanation for classification results
- Syllables-based model can serve this purpose : *"nouvalia"* is studied to be the name of an exposition center for all the new objects of the year, Naive Bayes says *"nouvalia"* belongs to the concept *"Nouveauté"* because it contains the syllables "*nou*" and "*va*' which are parts of the set of discriminative syllables from concept "*Nouveauté*".



Syllables of Concept : ***Nouveauté* : 51 syllables**

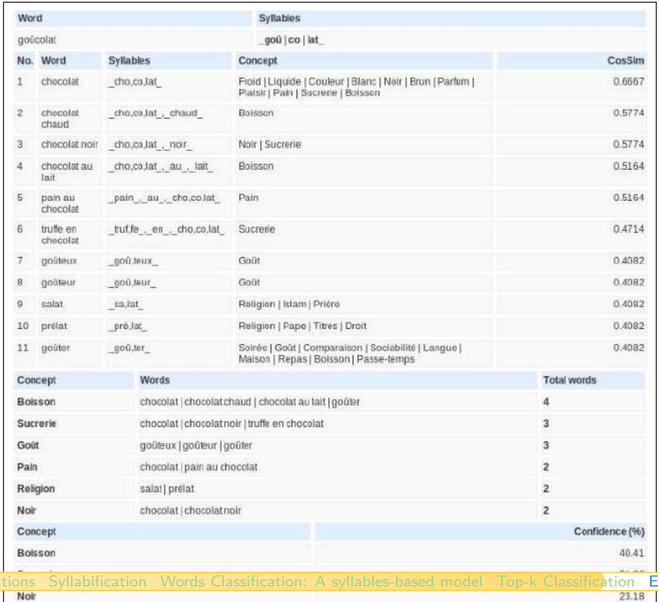
l'œil (0.0016) _look_ (0.0016) por (0.002) _jeu (0.0028) _neuf_ (0.0168) vel (0.0048) dis (0.0025) _bleu_ (0.0016) _l'oeil_ (0.0016) _oeil_ (0.0016) jeu (0.0032) _neu (0.0017) _fraî (0.0034) tua (0.0016) vant_ (0.0016) _néo (0.0028) vant (0.0023) -né_ (0.0016) _der (0.0027) _hom (0.0016) _pri (0.0017) _gar (0.0021) cheur_ (0.0017) _muer_ (0.0016) _frais_ (0.005) _an_ (0.0016) **_nou** (0.0181) _qqn_ (0.0032) _vieil (0.0032) ge_ (0.0017) _souf (0.0017) _re (0.003) _con (0.002) -gar (0.0023) _jeû (0.0016) _new (0.0016) ti (0.0036) _no (0.0037) nou (0.0023) _mo (0.0026) _tout_ (0.0019) _flam (0.0016) _vient_ (0.0017) _ve (0.0017) nis (0.002) _fic (0.0016) veau_ (0.0098) **va** (0.0021) der (0.0078) -look_ (0.0016) _né_ (0.0032)

KNN result

- Take all syllables into account for each comparing of pair words
- The result from confidence scores were ranked and top-3 classes were selected
- Experiment Results: Classification Accuracy by top-3 classes of KNN with various #num k neighbors.

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

KNN Result Example



Word	Syllables
goûcolat	_goû | co | lat_

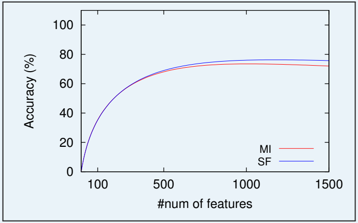
No.	Word	Syllables	Concept	CosSim
1	chocolat	_cho,co,lat_	Froid | Liquide | Couleur | Blanc | Noir | Brun | Parfum | Plaisir | Pain | Sucrerie | Boisson	0.6667
2	chocolat chaud	_cho,co,lat_,_chaud_	Boisson	0.5774
3	chocolat noir	_cho,co,lat_,_noir_	Noir | Sucrerie	0.5774
4	chocolat au lait	_cho,co,lat_,_au_,_lait_	Boisson	0.5164
5	pain au chocolat	_pain_,_au_,_cho,co,lat_	Pain	0.5164
6	truffe en chocolat	_truf,fe_,_en_,_cho,co,lat_	Sucrerie	0.4714
7	goûteux	_goû,teux_	Goût	0.4082
8	goûteur	_goû,teur_	Goût	0.4082
9	salat	_sa,lat_	Religion | Islam | Prière	0.4082
10	prélat	_pré,lat_	Religion | Pape | Titres | Droit	0.4082
11	goûter	_goû,ter_	Soirée | Goût | Comparaison | Sociabilité | Langue | Maison | Repas | Boisson | Passe-temps	0.4082

Concept	Words	Total words
Boisson	chocolat | chocolat chaud | chocolat au lait | goûter	4
Sucrerie	chocolat | chocolat noir | truffe en chocolat	3
Goût	goûteux | goûteur | goûter	3
Pain	chocolat | pain au chocolat	2
Religion	salat | prélat	2
Noir	chocolat | chocolat noir	2

Concept	Confidence (%)
Boisson	40.41
[illegible]	[illegible]
Noir	23.18

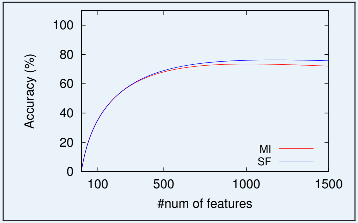
Compare Naive Bayes and KNN

The Naive Classifier

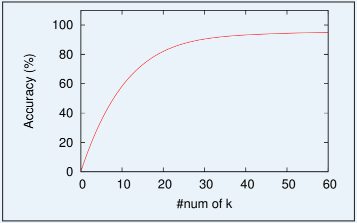


Compare Naive Bayes and KNN

The Naive Classifier



KNN



Plan

1. Motivations
2. Syllabification
3. Words Classification: A syllables-based model
4. Top-k Classification
5. Experiment and Result
6. Conclusion and Future work

Conclusion

Conclusion:

- KNN performed better than Naive Bayes
- Syllable Frequency(SF) archived the higher percentage of classification accuracy than Mutual Information(MI)
- Top-k classes helps user see more relevant concepts
- The syllables-based model helps to track back to explain why the word related to the concepts by using discriminative syllables set(Naive Bayes)

Future work

Future work:

- Although some syllables have meaning, but it is not enough for the linguists. The linguists need to know what are the lexemes in a word
- A lexeme is the minimal set of letters containing the meaning of a word
- Consider the way to find lexemes based on syllables. Instead of using syllables in classification model, lexemes will be used as a feature set

Thank you for your attention.