Classifying Words: A Syllables-based Model

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

Pattaraporn Warintaraweja, Anne Laurenta, Pierre Pompidorda, Armelle Cassanasb, Bénédicte Laurentb

aLirmm, University Montpellier 2
bNamae Concept

August, 31 2011
Motivations

Syllabification

Words Classification: A syllables-based model

Top-k Classification

Experiment and Result

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 linguists 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 linguists 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

<table>
<thead>
<tr>
<th>Brand names</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIVENDI NOUVALIA GOÛCOLAT AUREA</td>
</tr>
<tr>
<td>SEVEANE LYLIA SOLÉA EVARANDA ECONOVISTA</td>
</tr>
</tbody>
</table>

- **Linguistics demands:**
  - The linguists 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 linguists 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 $\Rightarrow$ 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**

<table>
<thead>
<tr>
<th>Round</th>
<th>Current stream</th>
<th>Result syllable</th>
<th>Next stream</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nouvalia</td>
<td>nou</td>
<td>valia</td>
<td>VCV → V-CV</td>
</tr>
<tr>
<td>2</td>
<td>valia</td>
<td>va</td>
<td>lia</td>
<td>VCV → V-CV</td>
</tr>
<tr>
<td>3</td>
<td>lia</td>
<td>lia</td>
<td>-</td>
<td>keep vowels together at the end of words</td>
</tr>
</tbody>
</table>
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
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:

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

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

\[
I(U; C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_1.N.1} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_0.N.1}
\]

\[
+ \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_1.N.0} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_0.N.0}
\]

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

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 calculated 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

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

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

(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:

\[
\text{confidence}(c, w_i) = \frac{\sum_{k_i \in K | (\text{Class}(k_i) = c)} \text{Sim}(k_i, w_i)}{\sum_{k_j \in K} \text{Sim}(k_j, w_i)}
\]  

(2)

Where \( \text{Sim} \) is the CosSim function.
1 Motivations

2 Syllabification

3 Words Classification: A syllables-based model

4 Top-k Classification

5 Experiment and Result

6 Conclusion and Future work
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.
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

<table>
<thead>
<tr>
<th>Concept(Class)</th>
<th>#Num of words</th>
<th>Concept</th>
<th>#Num of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Éventualité</td>
<td>138</td>
<td>Violence</td>
<td>355</td>
</tr>
<tr>
<td>Saisons</td>
<td>82</td>
<td>Distinction</td>
<td>168</td>
</tr>
<tr>
<td>Nouveauté</td>
<td>169</td>
<td>Droit</td>
<td>3,065</td>
</tr>
<tr>
<td>Humidité</td>
<td>195</td>
<td>Figures de discours</td>
<td>128</td>
</tr>
<tr>
<td>Terre</td>
<td>477</td>
<td>Architecture</td>
<td>1,539</td>
</tr>
<tr>
<td>Soleil</td>
<td>369</td>
<td>Poésie</td>
<td>378</td>
</tr>
<tr>
<td>Lichens</td>
<td>52</td>
<td>Pain</td>
<td>325</td>
</tr>
<tr>
<td>Reptiles</td>
<td>124</td>
<td>Sucrerie</td>
<td>274</td>
</tr>
<tr>
<td>Goût</td>
<td>196</td>
<td>Boisson</td>
<td>595</td>
</tr>
<tr>
<td>Effort</td>
<td>163</td>
<td>Mode</td>
<td>169</td>
</tr>
</tbody>
</table>

- 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.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>#Num of features</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MI</strong></td>
<td>100</td>
<td>72.57</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>75.50</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><strong>SF</strong></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><strong>77.22</strong></td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>75.70</td>
</tr>
</tbody>
</table>
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**

```
neuf (0.0168) veh (0.0048) dis (0.0025) bleu
(neuf (0.0016) )rest (0.0016) )seul (0.0016) )jeu (0.0022) )mae (0.0017)
fray (0.0041) )mae (0.0016) )nao (0.0028) )vant
(nou (0.0023) )ne (0.0016) )der (0.0027) )ban (0.0016) )pri (0.0017)
gar (0.0021) almost (0.0017) )maur (0.0016) )frais (0.005) _mn_
(nou (0.0016) )_nou (0.0181) )_wm (0.0032) )vioil
(nou (0.0032) )_wm (0.0016) )neuf (0.0017) )_mje (0.003) )con (0.002) )_gar
(nou (0.0023) )_wm (0.0016) )_new (0.0016) )ti (0.0036) )_na (0.0037)
(nou (0.0023) )_ne (0.0026) )tou (0.0019) )_jam (0.0016) )_viel_
(nou (0.0017) )_me (0.0017) )dis (0.002) )_ric (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.

<table>
<thead>
<tr>
<th>#Num of k</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>85.36</td>
</tr>
<tr>
<td>20</td>
<td>90.60</td>
</tr>
<tr>
<td>30</td>
<td>92.49</td>
</tr>
<tr>
<td>40</td>
<td>93.64</td>
</tr>
<tr>
<td>50</td>
<td>94.47</td>
</tr>
<tr>
<td>60</td>
<td>94.99</td>
</tr>
</tbody>
</table>
KNN Result Example

<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Syllables</th>
<th>Concept</th>
<th>CosSim</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chocolat</td>
<td><em>cho,co,lat</em></td>
<td>Froid</td>
<td>Liquide</td>
</tr>
<tr>
<td>2</td>
<td>chocolat chaud</td>
<td><em>cho,co,lat</em></td>
<td>Boisson</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>chocolat noir</td>
<td><em>cho,co,lat</em></td>
<td>Noir</td>
<td>Sucrerie</td>
</tr>
<tr>
<td>4</td>
<td>chocolat au lait</td>
<td><em>cho,co,lat</em></td>
<td>Boisson</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>pain au chocolat</td>
<td><em>pain__au__cho,co,lat</em></td>
<td>Pain</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>truffe en chocolat</td>
<td><em>tuf,fe__en__cho,co,lat</em></td>
<td>Sucrier</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>goûteux</td>
<td><em>goû,teux</em></td>
<td>Goût</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>goûteur</td>
<td><em>goû,leur</em></td>
<td>Goût</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>salat</td>
<td><em>sa,la</em></td>
<td>Religion</td>
<td>Islam</td>
</tr>
<tr>
<td>10</td>
<td>prélat</td>
<td><em>pré,la</em></td>
<td>Religion</td>
<td>Pape</td>
</tr>
<tr>
<td>11</td>
<td>goûter</td>
<td><em>goû,ter</em></td>
<td>Soirée</td>
<td>Goût</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept</th>
<th>Words</th>
<th>Total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boisson</td>
<td>chocolat</td>
<td>chocolat chaud</td>
</tr>
<tr>
<td>Sucrier</td>
<td>chocolat</td>
<td>chocolat noir</td>
</tr>
<tr>
<td>Goût</td>
<td>goûteux</td>
<td>goûteur</td>
</tr>
<tr>
<td>Pain</td>
<td>chocolat</td>
<td>pain au chocolat</td>
</tr>
<tr>
<td>Religion</td>
<td>salat</td>
<td>prél at</td>
</tr>
<tr>
<td>Noir</td>
<td>chocolat</td>
<td>chocolat noir</td>
</tr>
</tbody>
</table>
Compare Naive Bayes and KNN

The Naive Classifier

![Graph comparing accuracy and number of features for Naive Bayes (MI) and SF classifiers](image)

Motivations
Syllabification
Words Classification: A syllables-based model
Top-k Classification
Experiment and Results
Compare Naive Bayes and KNN

**The Naive Classifier**

![Graph](#)

**KNN**

![Graph](#)

Motivations
Syllabification
Words Classification: A syllables-based model
Top-k Classification
Experiment and Results
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:

- **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:

- 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.