EuroVoc classifier

Guido Boella
Dipartimento di Informatica
Università di Torino

FP7-ICT-2013-SME-DCA
Overview

• Introduction
• Background
• Our approach
• Pre-processing of the texts
• Evaluation
• Classification of legal text
  – deals with large amount of documents
  – it usually involves intensive manual work (slow and costly)

Need of automatization
Eurovoc thesaurus

• Eurovoc, a multilingual, multidisciplinary thesaurus with about 7,000 categories (also called classes, labels, or descriptors from now on) covering the activities of the EU, the European Parliament in particular.

• It contains terms in several languages and it is managed by the Publications Office of the European Union, an interinstitutional office whose task is to publish the publications of the institutions of European Union.

• Eurovoc is an ontology-based information collector that groups and links concepts through different types of relationships. The top level of the scheme is defined by 21 general concepts.
• Multi-label Text Classification
  – Each document can belong to more than one label / category

• Problems
  – Most of the algorithms only support mono-labeled datasets

• Solutions
  – Adaptation of existing algorithms to deal with multi-labels
  – Transformation of multi-labeled datasets into mono-labeled
• Background on transformation algorithms
  – Removal of all the documents that have more than one label from the dataset
  – Random selection of one of the multiple labels for each document, discarding the rest

Very naïve solutions! (with bad results)
• **Background on transformation algorithms**
  
  – Each different set of labels is considered as a single label *(power set)*
    
    • Example: *if the labels of a document are ‘A’, ‘B’, and ‘C’, the system transforms the labels of the document in a single label ‘ABC’.*
    
    • **Weakness:** *it may lead to datasets with a large number of classes and few examples per class.*

  – Learning one binary classifier for each label in the data
    
    • Classification procedure: *to classify a new document, it needs to pass over all the classifiers to determine its associated set of labels*
    
    • **Weakness:** *in case of thousands of categories (as in the data that we will use), this strategy becomes unsustainable.*
Our approach (1)

• Main idea...
  – ...each \( n \)-labeled document becomes a collection of \( n \) minor documents (each one associated to only one label),
  – and then use a state-of-the-art classification technique for mono-labeled datasets

• Problem...
  – ...how to segment the original document, that is how to choose the features to maintain for each of the new mono-label documents?
Our approach (2)

Category A

Category B

Category C

state-of-the-art technique for mono-labeled datasets
Our approach (3)

• Segmentation

  – We compute the Pointwise Mutual Information (PMI) between categories and features (terms)

    \[ M'_{i,j} = \frac{P_{i,j}}{P_i \times P_j} \]

    – \( P_{i,j} \) is the probability of having a non-zero co-occurrence value for the \( i \)-th feature and the \( j \)-th category in the whole corpus
    – \( P_i \) and \( P_j \) are the individual probabilities

• The utility of \( M' \) is to capture the strength of the associations between features and categories.
Our approach (4)

• Segmentation
  – for each original document vector $d$ to be segmented, given the set of categories $S_d$ to which it belongs, the system creates $n = |S_d|$ new document vectors $d_k$ (each one associated to exactly one class) in the following way:
    $$\vec{d}_k = < sel(f_1), sel(f_2), ...., sel(f_{|F|}) >$$
  – where $k \in S_d$ (it represents the category associated to the new vector), and where $sel(f_i)$ is a selection function that can assume the following values:

$$sel(f_i) = \begin{cases} f_i, & \text{if } M_{i,k} \geq M_{i,\hat{k}} \forall \hat{k} \in S_d \text{ and } \hat{k} \neq k \\ 0, & \text{otherwise} \end{cases}$$
Our approach (5)

• Segmentation
  – variant: selection parameter $Q$

\[
\text{sel}_Q(f_i) = \begin{cases} 
  f_i, & \text{if } \exists S'_d \subseteq S_d \text{ and } |S'_d| = Q \text{ and } \forall x \forall y [((x \in S'_d) \text{ and } \forall y \forall x [((x \in S'_d) \text{ and } \\
  (y \in (S_d/S'_d))) \rightarrow ((M_{i,x} \geq (M_{i,y})))]]} \\
  0, & \text{otherwise}
\end{cases}
\]

that is, $\text{sel}_Q(f_i)$ is equal to $f_i$ if there exists a subset of $S_d$ named $S’d$ of cardinality $Q$ such that each one of its element has a PMI-value with feature $f_i$ greater than (or equal to) all the elements outside $S’d$ (but in $S_d$). This way, the system allows the use of feature $f_i$ for exactly $Q$ segmented vectors.
## Pre-processing of the texts

<table>
<thead>
<tr>
<th>Language</th>
<th>Level of word analysis</th>
<th>Time unit - consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Stemming</td>
<td>1</td>
</tr>
<tr>
<td>Italian</td>
<td>Lemmatization and Part-Of-Speech</td>
<td>15</td>
</tr>
<tr>
<td>German</td>
<td>Stemming</td>
<td>1</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>Lemmatization</td>
<td>10</td>
</tr>
<tr>
<td>French</td>
<td>Stemming</td>
<td>1</td>
</tr>
</tbody>
</table>
Evaluation

• Measures
  – *Precision and Recall (and F-Measure).*

• Data
  – 23,472 documents (5 languages version)
### Evaluation

Evaluation of the classification module using 10-folds cross validation (values averaged on the 5 languages: IT, EN, BG, DE, FR)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steinberger Ralf et al (JEX)</td>
<td>47.13%</td>
<td>54.64%</td>
<td>50.61%</td>
</tr>
<tr>
<td>EuCases classification module</td>
<td>67.38%</td>
<td>74.81%</td>
<td>70.73%</td>
</tr>
</tbody>
</table>
## Evaluation

<table>
<thead>
<tr>
<th>Language</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>+0.8%</td>
</tr>
<tr>
<td>Italian</td>
<td>+1.1%</td>
</tr>
<tr>
<td>German</td>
<td>-0.2%</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>-1.2%</td>
</tr>
<tr>
<td>French</td>
<td>-0.5%</td>
</tr>
</tbody>
</table>
accordo CE (1474) [COMUNITÀ EUROPEE -> costruzione europea -> allargamento dell'Unione europea]
protocollo di accordo (2850) [RELAZIONI INTERNAZIONALI]
accordo di pesca (2556) [AGRICOLTURA - SILVICOLTURA - PESCA - COOPERAZIONE]
accordo provvisorio CE (5407) [COMUNITÀ EUROPEE -> costruzione europea -> allargamento dell'Unione europea]
trasporto aereo (4505) [TRASPORTO -> trasporti aerei e spaziali]
adesione all'Unione europea (12) [COMUNITÀ EUROPEE -> costruzione europea -> allargamento dell'Unione europea]

Accordo in forma di scambio di lettere relativo alla proroga del protocollo che fissa le possibilità di pesca e la contropartita finanziaria previste dall'accordo tra la Comunità economica europea e il governo della Repubblica democratica di São Tomé e Príncipe sulla pesca al largo di São Tomé e Príncipe per il periodo dal 10 giugno 2005 al 31 maggio 2006

A. Lettera della Comunità

Egregio signore,