Short Text Categorization
Exploiting Contextual Enrichment and External Knowledge

Stefano Mizzaro, Marco Pavan, Ivan Scagnetto, Martino Valenti

University of Udine, Italy
Disclaimer

- “Keep it simple, keep it short, and nobody will complain” [Michael Buckland]

- The Good Presentation Gold Rule
#ShortTxtCateg...

SM, MP, IS, MV
uniud, IT
#Outline

- #pbm
- #approach
- #eval
- @home
The problem

- Short texts are growing
- (at least) 2 reasons
  - Twitter 140 limit
  - Mobile devices, input limitations
- Categorization of short texts, or #ShortTxtCateg
#ShortTxtCateg: why it is useful

- To understand what the txt is about
  - #socceroos: easy
  - Goalkeeper did a good job today: difficult (which team? Which “today”?)
  - “I hate that referee”
  - “I hate that referee... He did not understand my paper”
- We focus on Tweets, but not only (facebook status & comments, txt messages, …)
#ShortTxtCateg: why difficult

- Not enough data
  - Short sentences
  - Abbreviated words, new coined acronyms
- Typos, mispellings, grammar wrong is often
- Time, ephemeral content
- Ambiguity, Disambiguation is more difficult
Damiano Spina @damiano10 · 11h
POI mentions in Twitter are likely to be ambiguous (~50% unigrams) #microblog #SIGIR2014
Details

Damiano Spina @damiano10 · Jul 6
Jet lag is almost under control. Gold Coast is simply amazing! #sigir2014
Details
#ShortTxtCateg: why difficult

- Not enough data
  - Short sentences
  - Abbreviated words, new coined acronyms
- Typos, mispellings, grammar wrong is often
- Time, ephemeral content
- Ambiguity, Disambiguation is more difficult
- #hashtags: potentially useful, but not "normal words"
- Combination: #WFT?!
Combination: #WFT?!

- #WTF = Whom To Follow
- but also…
  - #WTF = What the F*%&%
- or, for IR researchers,
  - #WTF = Where is The F^%$#& data?
Aim

• Find categories/labels that describe the general topic of a short text

• More specifically:

  • Select the Wikipedia categories that best describe a tweet
<table>
<thead>
<tr>
<th>Number</th>
<th>English Category</th>
<th>Italian Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Literature</td>
<td>Letteratura</td>
</tr>
<tr>
<td>2</td>
<td>Economics</td>
<td>Economia</td>
</tr>
<tr>
<td>3</td>
<td>History</td>
<td>Storia</td>
</tr>
<tr>
<td>4</td>
<td>Philosophy</td>
<td>Filosofia</td>
</tr>
<tr>
<td>5</td>
<td>Science</td>
<td>Scienza</td>
</tr>
<tr>
<td>6</td>
<td>Entertainment (Hobby, Entertainment)</td>
<td>Intrattenimento (Hobby, Intrattenimento)</td>
</tr>
<tr>
<td>7</td>
<td>Finance</td>
<td>Finanza</td>
</tr>
<tr>
<td>8</td>
<td>Politics=(Politics, Law)</td>
<td>Politica=(Politica, Diritto)</td>
</tr>
<tr>
<td>9</td>
<td>Food and drink</td>
<td>Alimentazione</td>
</tr>
<tr>
<td>10</td>
<td>Video games</td>
<td>Videogiochi</td>
</tr>
<tr>
<td>11</td>
<td>Computer science</td>
<td>Informatica</td>
</tr>
<tr>
<td>12</td>
<td>Health and fitness=(Health, Physical fitness)</td>
<td>Salute e fitness=(Salute, Fitness)</td>
</tr>
<tr>
<td>13</td>
<td>Fashion</td>
<td>Moda</td>
</tr>
<tr>
<td>14</td>
<td>Medicine</td>
<td>Medicina</td>
</tr>
<tr>
<td>15</td>
<td>Music</td>
<td>Musica</td>
</tr>
<tr>
<td>16</td>
<td>Engines=(Automobiles, Auto racing, Motorcycle sport)</td>
<td>Motori=(Automobili, Automobilismo, Motociclismo)</td>
</tr>
<tr>
<td>17</td>
<td>Photo and Video=(Photography, Film)</td>
<td>Foto e Video=(Fotografia, Cinema)</td>
</tr>
<tr>
<td>18</td>
<td>Sports</td>
<td>Sport</td>
</tr>
<tr>
<td>19</td>
<td>Places=(Tourism, Geography, Travel)</td>
<td>Luoghi=(Turismo, Geografia, Viaggi)</td>
</tr>
<tr>
<td>20</td>
<td>Meteorology</td>
<td>Metereologia</td>
</tr>
</tbody>
</table>

Table 1: Wikipedia categories used in our systems, in English and Italian. The notation X=(Y,Z,...) denotes the labels we made to group categories about related topics.
Outline

- #pbm
- #approach
- #eval
- @home
Our approach

• Exploiting Wikipedia
  • Search engine
  • Article/category labels
  • Category relationships
• Enrichment
  • Exploiting search engines
• Time aware
Categories selection

- We select the Wikipedia articles by search
- We extract their categories
- We browse the category graph
- We pick the nearest ones
3 versions of a system

1. W2C
2. FEL
3. WEL
### 3 systems

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. W2C</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>2. FEL</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>3. WEL</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>
1. W2C

- Step 1: Article selection
  - Query definition, by using **bi-grams** from short text
  - Article retrieval process (ranked by Wikipedia search engine)
  - Article re-weighting process, (exploiting their positions in the ranking)
  - Final articles list with distinct entries (by performing all queries and summing the scores)

- Step 2: Label selection
  - Wikipedia categories extraction (for each article)
  - Article-Macro-category relationship definition (based on **shortest paths**)
  - Wikipedia Macro-categories selection (based on our ranking function)
  - Final set of **5 labels**, based on selected Macro-categories
Figure 1: Workflow of the labelling process.
2. FEL

- Enters (short) text enrichment
- The short txt is augmented with some other terms
Workflow

Figure 1: Workflow of the labelling process.
Figure 1: Workflow of the labelling process.
Text enrichment

Figure 3: Workflow of the enrichment process.
Now, Time

• To be timely is important. I should have said that earlier…
Now, Time

• To be timely is important. I should have said that earlier...

• We query google right after the tweet

• Well actually a few hours (6) after the tweet.
3. WEL

Relevance value

CUT-OFF

Words (ordered by Rank)

$\text{h}_1$ (jump at least 0.1 relevance value)

$\text{h}_2$

Words selected for enrichment
Outline

• #pbm
• #approach
• #eval
• @home
Experimental evaluation

- 3 versions of the system (W2C, FEL, WEL), which is better?
- 20 labels/categories
- 10 twitter accounts
  - 30 tweets
- Assessments by 66 people
Assessing

• Participant was shown a set of labels generated by a system
  
  • “Is this set of labels good for describing the topic of the tweet?”

• 5 levels scale (1=worst, 5=best)

• Usual random shuffling, avoiding learning effects, etc.
Results

• Statistically significant
• High variance over tweets

Figure 4: Average rating for each short text
Results

- Statistically significant
- High variance over tweets

Figure 4: Average rating for each short text
6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented our short text categorization system. We developed a first version, W2C, that exploits Wikipedia as external knowledge source. Then, we improved it with an enrichment approach by developing FEL and WEL. The labels proposed by FEL have been evaluated better than the other solutions. In general, the enrichment improved the topic detection, but the cut-off function still needs to be enhanced to better exploit the semantic relations between words in the final rank list. Despite our observations in Section 3.2.2, the parameters used in the algorithm need to be tuned up to refine the cut-off index.

Our system represents a new proposal for short text categorization that does not need the help of URLs inside the text, or hashtags, or other social media features. With this approach it can be used also for general short texts, such as text messages, or vocal messages, on mobile phones. On this basis, we have planned to run other experiments to test new settings for the enrichment process with the goal to better emphasize the semantic relations between extracted words. We can also select different sets of macro-categories from Wikipedia for the W2C module, to test the system with other levels of granularity for topics. Another future work is related to user modelling; we planned to run this system on a set of short texts extracted from a single user social network account. Thus, we can try to detect the main topics discussed by the given user. This work can be a new approach for the development of a new proposal for user modelling based on social data.

7. REFERENCES

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented our short text categorization system. We developed a first version, W2C, that exploits Wikipedia as external knowledge source. Then, we improved it with an enrichment approach by developing FEL and WEL. The labels proposed by FEL have been evaluated better than the other solutions. In general, the enrichment improved the topic detection, but the cut-off function still needs to be enhanced to better exploit the semantic relations between words in the final rank list. Despite our observations in Section 3.2.2, the parameters used in the algorithm need to be tuned up to refine the cut-off index.

Our system represents a new proposal for short text categorization that does not need the help of URLs inside the text, or hashtags, or other social media features. With this approach it can be used also for general short texts, such as text messages, or vocal messages, on mobile phones. On this basis, we have planned to run other experiments to test new settings for the enrichment process with the goal to better emphasize the semantic relations between extracted words. We can also select different sets of macro-categories from Wikipedia for the W2C module, to test the system with other levels of granularity for topics. Another future work is related to user modelling; we planned to run this system on a set of short texts extracted from a single user social network account. Thus, we can try to detect the main topics discussed by the given user. This work can be a new approach for the development of a new proposal for user modelling based on social data.

7. REFERENCES


Outline

• #pbm
• #approach
• #eval
• @home
Conclusions

• #ShortTxtCateg
• @timeaware
• w/ or w\ txt enrichment
• txt enrichm seems useful
  • 2. FEL better than 3. WEL
Future work

• #WTF?
• Too much to be listed here
• Plenty of space for improvement
Any plans for the summer, Prof. Smith?

Yes, I'm going on vacation with my family.

However, I'll still be checking e-mail every day, having regular conference calls and working on grant proposals while I'm there.

Do you really want to go?

I'd rather send one of my postdocs but my spouse won't let me.