Identifying Topics in Social Media Posts using DBpedia

Óscar Muñoz-García, Manuel de la Higuera Hernández, Carlos Navarro (Havas Media)
Andrés García-Silva, Óscar Corcho (Ontology Engineering Group - UPM)
NEM Summit – 28 Sept. 2011
Identifying Topics in Social Media Posts using DBpedia

Introduction
Introduction

- **Topic Identification**
  - “The task of identifying the central ideas in a text” [Chin-Yew Lin, 1995]

- **Applications of Topic Identification for Social Media**
  - Automatically summarising the content published in a channel.
  - Mining the interest of a given user.
  - etc…

- **Benefits for Advertising Companies**
  - To focus the advertisement actions to the appropriate channels.
  - To serve ads to the users based in their interest.
Difficulties of Topic Identification in Social Media

- Different channels with heterogeneous texts
  - Different lengths
    - From short sentences on Twitter to medium-size articles in blogs
  - Misspellings
    - Posts completely written in uppercase (or lowercase) letters
      - Makes difficult the detection of proper nouns.
    - In Spanish, absence / presence of an accent in a word → different meanings
      - “té” = “tea” (common noun)
      - “te” = “you” (personal pronoun)
  - Use of set phrases
    - E.g., “too many cooks spoil the broth” (if too many people try to take charge at a task, the end product might be ruined)
    - E.g., “rain cats and dogs” (rains heavily)

It is important to take into account the context of the post
Why DBpedia?

- DBpedia is a structured Semantic Web representation of Wikipedia
  - Wikipedia is maintained by thousands of editors
  - Wikipedia evolves and adapts as knowledge change [Syed et al, 2008]
- Each topic identified is mapped with a DBpedia resource
  - E.g., The URI http://dbpedia.org/resource/Turin
    - Represents the city of Torino
    - Has about 45 attributes defined (population, area, latitude, longitude, etc.)
    - Has labels and definitions in 14 different languages.
    - It is linked with many semantic entities
      - E.g. Birth place of Amedeo Avogadro: http://dbpedia.org/resource/Amedeo_Avogadro
      - It is linked with its Wikipedia article: http://en.wikipedia.org/wiki/Torino
- It is a nucleus for the Web of Data [Bizer et al, 2009]
  - Data published on the Web according to Tim Berners-Lee’s Linked Data principles.
  - Several billion RDF triples (i.e. facts)
  - Multi-domain datasets (geographic information, people, companies, online communities, etc…)
Identifying Topics in Social Media Posts using DBpedia

Related Work
Wikipedia has been exploited for the following tasks:

- Topic identification and text categorization
  - [Bodo et al, 2007], [Coursey et al, 2009], [Gabrilovich et al., 2006], [Syed et al, 2008], [Schonhofen, 2009]
- Semantic Relatedness between fragments of text
  - [Gabrilovich et al, 2007]
- Keyword Extraction
  - [Mihalcea et al, 2007]
- Word sense disambiguation
  - [Mihalcea, 2007]
Related Work

Uses of Wikipedia data-structure:

- Relating words in text with articles using article **title** information
  - [Schonhofen, 2009]
- Exploiting **anchor text** in links
  - [Coursey et al, 2009] [Mihalcea et al, 2007] [Mihalcea, 2007]
- Exploiting the **whole articles**
  - [Syed et al, 2008] [Gabrilovich, 2007]
- Exploiting **categories** to measure **relatedness** between articles
  - [Coursey et al, 2009] [Syed et al, 2008]
- Exploiting **disambiguation pages** and **redirection links** to select candidate senses and alternative labels
  - [Mendelyan et al, 2008]
Related Work

- **Supervised learning methods**

- **Unsupervised techniques**
  - Based on a Vector Space Model
    - [Schonhofen, 2009]
  - Based in a Graph
    - [Coursey et al, 2009] [Syed et al, 2008]

- **Combined methods (supervised and unsupervised)**
  - Based on a Vector Space Model
    - [Mihalcea et al, 2007]
Our approach

- Exploits titles, disambiguation pages, redirection links and article text to select candidate senses and alternative labels
- Uses an unsupervised method
- Uses a vector space model

Main benefit in comparison with previous approaches:
- The interlinking of social media posts with the Web of data through DBpedia resources
Identifying Topics in Social Media Posts using DBpedia

Description of the Method
Description of the Method

Input

- Go to #NEMSummit 2011 for #Social #User #Immersive #Pervasive & #Cloud #media, stay for Sandpit and Art & enjoy #Torino

Part-of-speech tagging

- “torino”, “art”, “media”, “user”, “cloud”

Topic Recognition

- http://dbpedia.org/resource/Turin
- http://dbpedia.org/resource/Art

Language Filtering

- “Torino”, “arte”, “utente”, “mezzo di comunicazione di massa”, ...
**Part-of-speech tagging**
- \( W_p = w_1, w_2, \ldots, w_n \equiv \) list of lexical units contained in the post
- \( \text{lexcat}(w) \equiv \) lexical category of the lexical unit \( w \)
- \( \text{lemma}(w) \equiv \) lemma of \( w \)
- \( L = \{ \text{common noun, proper noun, acronym} \ldots \} \equiv \) meaningful lexical categories that we consider
- \( \theta = \{ \text{“RT”, “/cc”, “;”), } \ldots \} \equiv \) stop words (lemmas excluded)
- \( K_p = k_1, k_2, \ldots, k_n \equiv \) list of keywords with meaning

```python
def GetKeywords(W_p):
    K_p \leftarrow \emptyset
    for each \ w_i \ in \ W_p :
        if \ lexcat(w_i) \in \ L \ and \ lemma(w_i) \notin \theta :
            K_p \leftarrow K_p \cup \{\text{lemma}(w_i)\}
    return K_p
```
Part-of-speech tagging example

- But a **hardware problem** is more likely, especially if you use the **phone** a lot while eating. The **Blackberry**’s tiny **trackball** could be suffering the same accumulation of **gunk** and **grime** that can plague a **computer mouse** that still uses a **rubber ball** on the underside to roll around the **desk**.

Description of the Method

**Topic Recognition** (Sem4Tags [García-Silva et al, 2010])

### POS Tagging
- Blackberry, phone, trackball, computer, problem, grime, hardware, mouse, desk, rubber ball, gunk

### Context Selection
- Blackberry, {phone, hardware, trackball, mouse}
- Computer, {hardware, mouse, problem, desk}
- ...

### Disambiguation
- [http://dbpedia.org/resource/BlackBerry](http://dbpedia.org/resource/BlackBerry)
Context Selection

- For each keyword, a set of up to 4 related keywords that will help to disambiguate its meaning.
- 4 is the number of words above which the context does not add more resolving power to disambiguation [Kaplan, 1955].
- We compute semantic relatedness (active context) taking into account the co-occurrence of words in web pages [Gracia et al, 2009].

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Relatedness</th>
<th>Keyword</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone</td>
<td>0.347</td>
<td>hardware</td>
<td>0.347</td>
</tr>
<tr>
<td>trackball</td>
<td>0.311</td>
<td>mouse</td>
<td>0.311</td>
</tr>
<tr>
<td>computer</td>
<td>0.288</td>
<td>desk</td>
<td>0.287</td>
</tr>
<tr>
<td>problem</td>
<td>0.246</td>
<td>rubber ball</td>
<td>0.246</td>
</tr>
<tr>
<td>grime</td>
<td>0.190</td>
<td>gunk</td>
<td>0.168</td>
</tr>
</tbody>
</table>

Active context selection for **blackberry** keyword.
Description of the Method

● **Disambiguation Criteria**

  ● **OPTION 1**: Most frequent sense for the ambiguous word
    - Determined by Wikipedia editors (the first link in a disambiguation page)
  
  ● **OPTION 2**: Vector space model
    1. A vector containing the keyword and its context
    2. A vector containing top N terms is created from each candidate sense is created using TF-IDF (Term Frequency and Inverse Document Frequency)
    3. The cosine similarity is used to determine which vectorised sense is more similar to the vector associated to the keyword

<table>
<thead>
<tr>
<th>DBpedia resource</th>
<th>Definition</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlackBerry</td>
<td>Is a line of mobile e-mail and smartphone</td>
<td>0.224</td>
</tr>
<tr>
<td>Blackberry</td>
<td>Is an edible fruit</td>
<td>0.15</td>
</tr>
<tr>
<td>BlackBerry_(song)</td>
<td>Is a song by the Black Crowes</td>
<td>0.0</td>
</tr>
<tr>
<td>BlackBerry_Township, _Itasca_County, _Minnesota</td>
<td>Is a township in … Itasca County</td>
<td>0.0</td>
</tr>
</tbody>
</table>

\[
\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t)
\]

\[
\text{idf}(t) = \log \frac{|D|}{|\{d : t \in d\}|}
\]

\[
\text{cosine}(d_1, d_2) = \frac{(d_1 \cdot d_2)}{\|d_1\| \cdot \|d_2\|}
\]
Description of the Method

- **Language Filtering**
  - $T_p = t_1, t_2, ..., t_n \equiv$ set of topics identified
  - $l \equiv$ language to filter
  - $Labels(t) \equiv$ set of labels associated to a given topic (value of rdfs:label property)
  - $lang(b) \equiv$ language of a given label
  - $T^l_p \equiv$ set of topics with labels in language $l$

```python
def FilterLanguage($T_p$, $l$):
    $T^l_p \leftarrow \emptyset$
    for each $t_i$ in $T_p$:
        if $\exists b_j \in Labels(t_i) | lang(b_j) = l$:
            $T^l_p \leftarrow T^l_p \cup \{t_i\}$
    return ($T^l_p$)
```
Identifying Topics in Social Media Posts using DBpedia

Evaluation
Evaluation

- Evaluated with a corpora of 10,000 posts in Spanish extracted from
  - Blogs
  - Forums
  - Microblogs (e.g., Twitter)
  - Social networks (e.g., Facebook, MySpace, LinkedIn and Xing)
  - Review sites (e.g., Ciao and Dooyoo)
  - Audiovisual sites (e.g., YouTube and Flickr)
  - News publishing sites (e.g., elpais.com, elmundo.es)
  - Others (web pages not classified in the categories above)

- Variants evaluated
  1. Without considering any context
     - Default Wikipedia sense assigned for a given keyword
  2. Considering as context all the other keywords found in the same post
  3. Active context selection technique
     - Selecting the 4 most relevant topics from the keywords in the same post
Evaluation

Coverage

<table>
<thead>
<tr>
<th></th>
<th>Blogs</th>
<th>Forums</th>
<th>Microblogs</th>
<th>Social Networks</th>
<th>Others</th>
<th>Reviews</th>
<th>Audiovisual</th>
<th>News</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS Tagging</td>
<td>99.63%</td>
<td>96.64%</td>
<td>99.01%</td>
<td>98.14%</td>
<td>98.77%</td>
<td>98.20%</td>
<td>97.20%</td>
<td>99.62%</td>
<td>98.32%</td>
</tr>
<tr>
<td>Topic identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without context</td>
<td>96.7%</td>
<td>87.68%</td>
<td>94.22%</td>
<td>93.54%</td>
<td>92.71%</td>
<td>88.81%</td>
<td>90.29%</td>
<td>96.67%</td>
<td>92.35%</td>
</tr>
<tr>
<td>With context</td>
<td>96.64%</td>
<td>93.07%</td>
<td>95.54%</td>
<td>94.99%</td>
<td>95.13%</td>
<td>92.67%</td>
<td>97.41%</td>
<td>98.54%</td>
<td>95.02%</td>
</tr>
<tr>
<td>Active context</td>
<td>99.24%</td>
<td>89.71%</td>
<td>94.43%</td>
<td>96.40%</td>
<td>94.75%</td>
<td>93.81%</td>
<td>92.23%</td>
<td>97.4%</td>
<td>94.72%</td>
</tr>
<tr>
<td>Topic identification after language filtering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without context</td>
<td>91.21%</td>
<td>79.04%</td>
<td>87.54%</td>
<td>82.64%</td>
<td>86.93%</td>
<td>70.15%</td>
<td>82.52%</td>
<td>90.71%</td>
<td>82.74%</td>
</tr>
<tr>
<td>With context</td>
<td>88.43%</td>
<td>80.84%</td>
<td>86.31%</td>
<td>85.24%</td>
<td>88.72%</td>
<td>76.19%</td>
<td>89.66%</td>
<td>92.46%</td>
<td>84.85%</td>
</tr>
<tr>
<td>Active context</td>
<td>89.69%</td>
<td>80.51%</td>
<td>86.51%</td>
<td>86.78%</td>
<td>89.78%</td>
<td>75.59%</td>
<td>80.58%</td>
<td>90.54%</td>
<td>84.73%</td>
</tr>
</tbody>
</table>

- Part-of-speech tagging: nearly 100%
- Topic recognition: over 90% for almost all the cases
- After language filtering coverage is reduced in about 10% because not all DBpedia resources have a label defined for Spanish language
Evaluation

- **Precision**
  - Evaluated a random sample of 1,816 posts (18.16%)
  - 47 human evaluator
  - Each post and topics identified shown to 3 different evaluators
  - Evaluation options:
    1. The topic is not related with the post
    2. The topic is somehow related with the post
    3. The topic is closely related with the post
    4. The evaluator has not enough information for taking a decision

- **Fleiss’ kappa test**
  - Strength of agreement for 2 evaluators = 0.826 (very good)
  - Strength of agreement for 3 evaluators = 0.493 (moderate)
Con ese plan, la compañía comenzó a desarrollar su oferta comercial en España y Latinoamérica bajo la marca Movistar, y en Europa con O2, convirtiéndose así Telefónica en la marca institucional de la compañía.

<table>
<thead>
<tr>
<th>Nada relacionado</th>
<th>Algo relacionado</th>
<th>Muy relacionado</th>
<th>NS/NC</th>
<th>Categoría/Tema</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Marca</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Una Marca es un título que concede el derecho exclusivo a la</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>utilización de un signo para la identificación de un producto</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>o un servicio en el mercado. Pueden ser marcas las palabras</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>o combinaciones de palabras, imágenes, figuras, símbolos,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>gráficos, letras, cifras, formas tridimensionales (envoltor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>iones, envases, formas del producto o su representación)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>América Latina</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>El término América Latina o Latinoamérica se refiere a las</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>regiones de América donde se hablan lenguas latinas,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>específicamente español, francés y portugués. Tiene varios</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>usos y connotaciones divergentes:.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>España</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>España, también denominado Reino de España, es un país</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>soberano miembro de la Unión Europea, constituido en Estado</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>social y democrático de Derecho y cuya forma de gobierno es</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>la monarquía parlamentaria.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Categoría: Marcas comerciales</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Categoría: Marca</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Categoría: España</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Categoría: Países de Europa</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Categoría: Países transcontinentales</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Categoría: Arquitectos de España</strong></td>
</tr>
</tbody>
</table>
Precision Results

<table>
<thead>
<tr>
<th></th>
<th>Blogs</th>
<th>Forum</th>
<th>Microblogs</th>
<th>Other</th>
<th>Social Networks</th>
<th>Reviews</th>
<th>Video</th>
<th>News</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without context</td>
<td>67.48%</td>
<td>66.67%</td>
<td>59.72%</td>
<td>72.32%</td>
<td>59.19%</td>
<td>79.17%</td>
<td>84.44%</td>
<td>71.95%</td>
<td>68.42%</td>
</tr>
<tr>
<td>With context</td>
<td>75.61%</td>
<td>59.35%</td>
<td>54.88%</td>
<td>65.71%</td>
<td>53.52%</td>
<td>83.87%</td>
<td>77.78%</td>
<td>64.37%</td>
<td>63.11%</td>
</tr>
<tr>
<td>Active context</td>
<td>67.71%</td>
<td>64.45%</td>
<td>65.58%</td>
<td>70.1%</td>
<td>49.15%</td>
<td>88.89%</td>
<td>79.07%</td>
<td>71.93%</td>
<td>66.59%</td>
</tr>
</tbody>
</table>

- Precision depends on the channel
  - From 59.19% for social networks
    - More misspellings
    - More common nouns
  - To 88.89% for review sites
    - Concrete products and brands
    - Proper nouns tend to have a Wikipedia entry

- Context selection criteria also depends on the channel
  - Active context selection better for **microblogs** and **review sites**
  - Considering all the post keywords as context better for **blogs**
  - Without context selection is better for the rest of the cases (almost all the channels)
    - Naïve default sense selection is effective
Identifying Topics in Social Media Posts using DBpedia

Conclusions
Conclusions

- We have achieved good results of coverage
- The precision depends on the channel (better for review sites, worst for social networks)
- With respect to considering context or not, there is not a variant that provide the best results for all the channels.

Future lines of work:
  - Improve Natural Language Processing
    - Dealing with slang
    - Detect set phrases
    - Improve n-gram detection
    - Dealing with microblogs’ specifics (e.g., hashtag expansion)
  - Combine broad-domain topic identification with knowledge about specific domains
    - Use of domain ontologies in combination with DBpedia ontology
Thank you!

oscar.munoz@havasmedia.com