

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

- Introduction
- Related Work
- Description of the Method
- Evaluation
- Conclusions



Identifying Topics in Social Media Posts using DBpedia

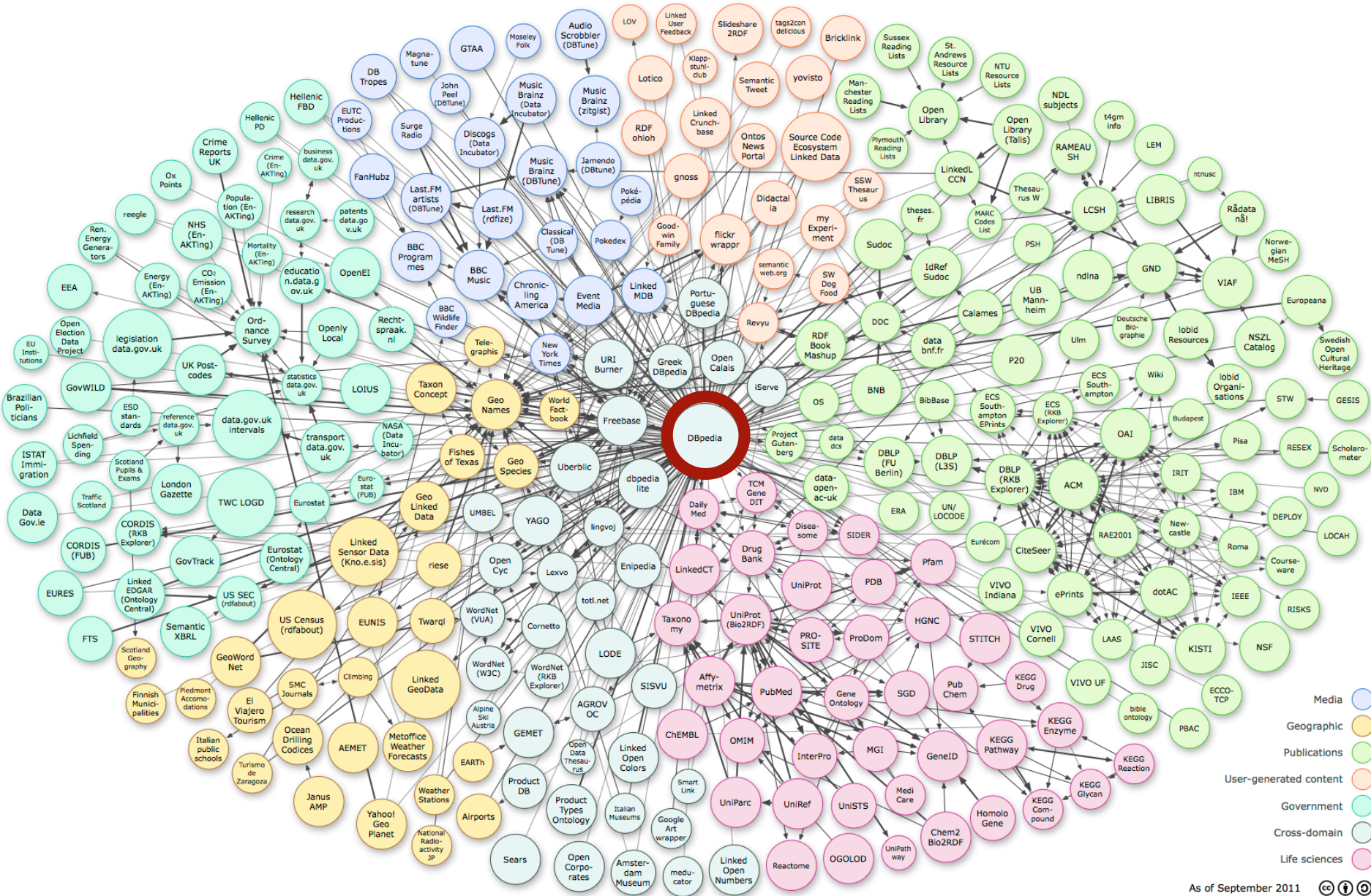
Introduction

- 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...)

Introduction





Identifying Topics in Social Media Posts using DBpedia

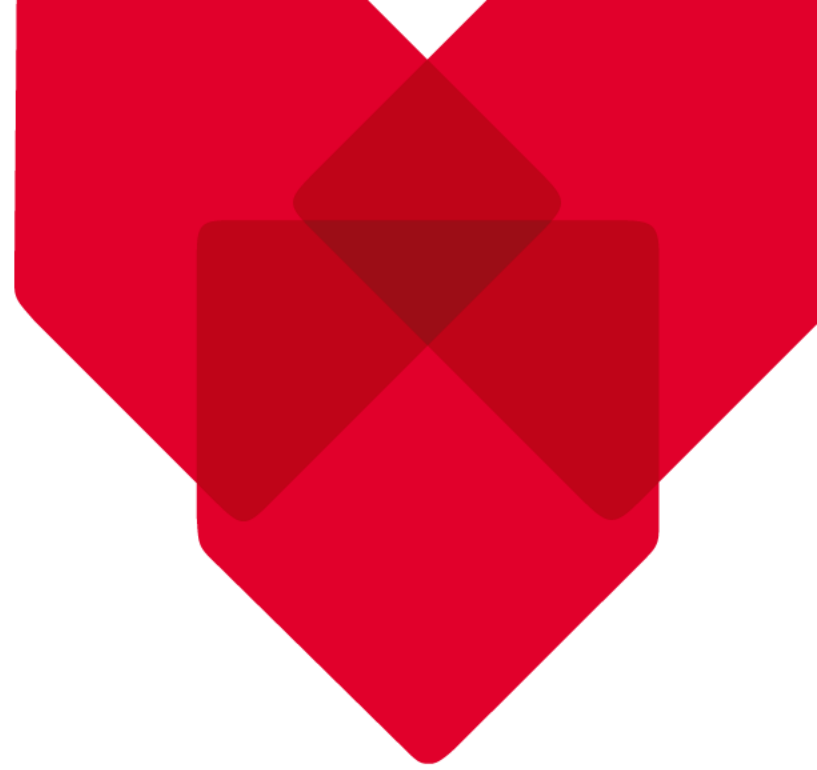
Related Work

- 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]

- 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]

- Supervised learning methods
 - [Bodo et al, 2007] [Gabrilovich et al, 2006] [Mendelyan et al, 2008]
- 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



ValiLalioti vali lalioti

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>
- [http://dbpedia.org/resource/User_\(computing\)](http://dbpedia.org/resource/User_(computing))

Language
Filtering

- “Torino”, “arte”, “utente”, “mezzo di comunicazione di massa”, ...



- Part-of-speech tagging

- $W_p = w_1, w_2, \dots, w_n$ \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
- $L = \{\text{common noun, proper noun, acronym...}\}$ \equiv meaningful lexical categories that we consider
- $\theta = \{\text{"RT", "/cc", ";", ...}\}$ \equiv stop words (lemmas excluded)
- $K_p = k_1, k_2, \dots, k_n$ \equiv list of keywords with meaning

def GetKeywords(W_p) :

$K_p \leftarrow \emptyset$

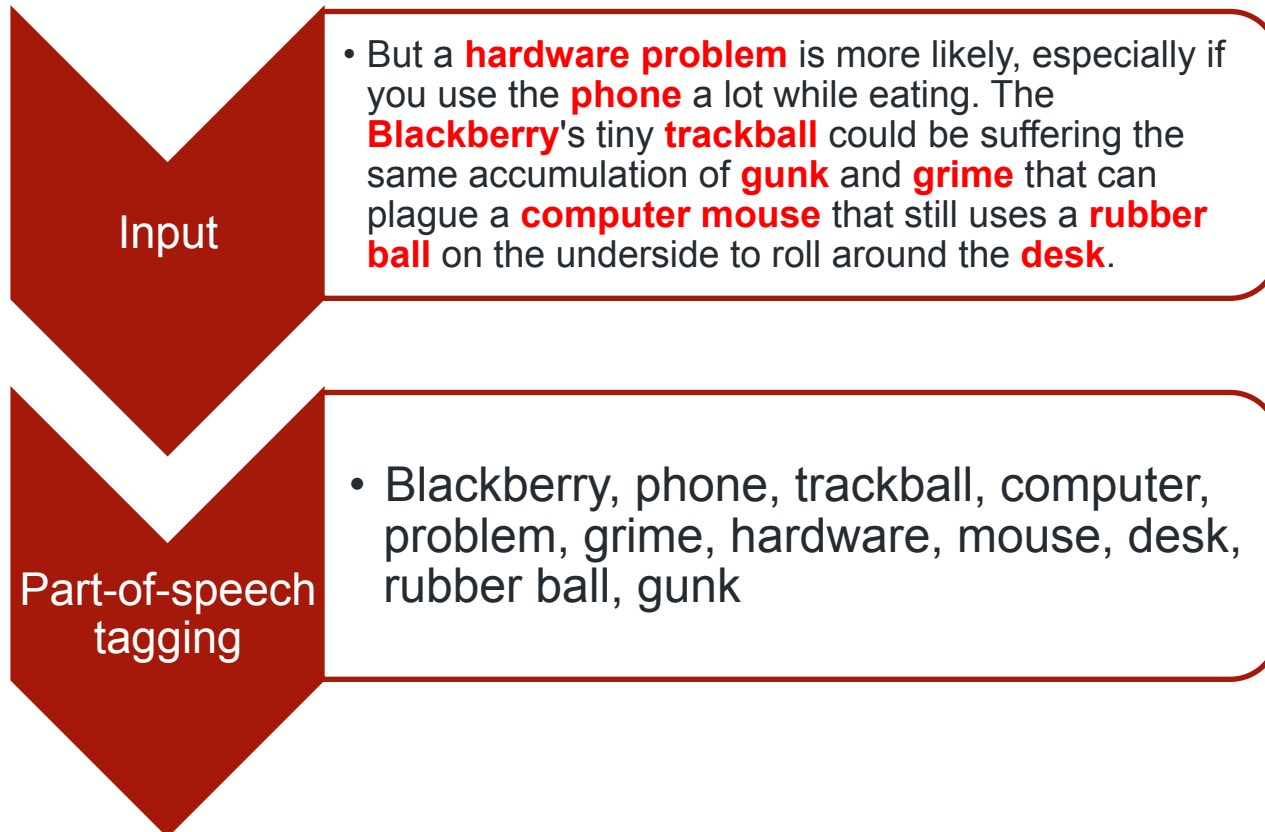
for each w_i **in** W_p :

if $\text{lexcat}(w_i) \in L$ **and** $\text{lemma}(w_i) \notin \theta$:

$K_p \leftarrow K_p \cup \{\text{lemma}(w_i)\}$

return K_p

- Part-of-speech tagging example



- 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/Computer>

● Context Selection

- For each keyword, a set of up to 4 related keywords that will help to disambiguate the 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]

Keyword	Relatedness	Keyword	Relatedness
phone	0.347	hardware	0.347
trackball	0.311	mouse	0.311
computer	0.288	desk	0.287
problem	0.246	rubber ball	0.246
grime	0.190	gunk	0.168

Active context selection for ***blackberry*** keyword

● 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

DBpedia resource	Definition	Similarity
BlackBerry	Is a line of mobile e-mail and smartphone	0.224
Blackberry	is an edible fruit	0.15
BlackBerry_(song)	is a song by the Black Crowes	0.0
BlackBerry_Township,_ _Itasca_County, _Minnesota	Is a township in ... Itasca County	0.0

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

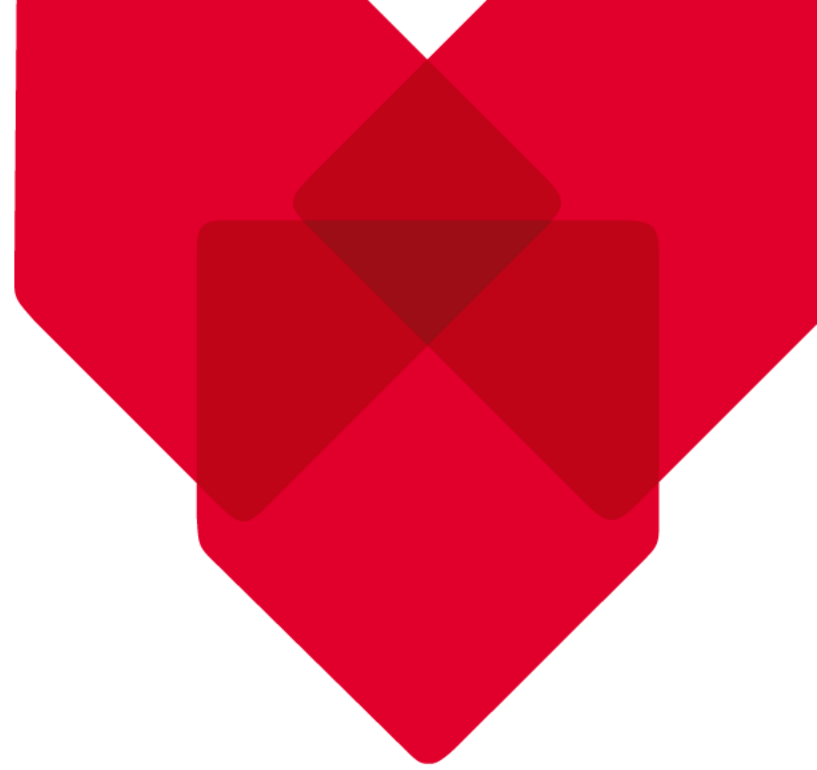
$$\text{idf}(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$

$$\text{cosine}(\mathbf{d}_1, \mathbf{d}_2) = (\mathbf{d}_1 \bullet \mathbf{d}_2) / \|\mathbf{d}_1\| \|\mathbf{d}_2\|$$

- Language Filtering

- $T_p = t_1, t_2, \dots, t_n \equiv$ 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
- $T_p^l \equiv$ set of topics with labels in language l

```
def FilterLanguage( $T_p, l$ ) :  
     $T_p^l \leftarrow \emptyset$   
    for each  $t_i$  in  $T_p$  :  
        if  $\exists b_j \in Labels(t_i) | lang(b_j) = l$  :  
             $T_p^l \leftarrow T_p^l \cup \{t_i\}$   
    return ( $T_p^l$ )
```



Identifying Topics in Social Media Posts using DBpedia

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

- Coverage

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

- 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

- 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)

Evaluación 1/10

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.

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

[Borrar datos](#)[Enviar valoración](#)

● Precision Results

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

- 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

- 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

