Automatic Ontology Population from News
ATCM Lab Work

Hernani Costa
hpcosta@dei.uc.pt
Supervisor: Alexandre Miguel Pinto
ampinto@dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra

Coimbra, 13 July, 2012
We are experiencing a huge growth in the available information, most of the time provided as natural language text (e.g., daily news).

Keyphrases provide a brief summary of a document’s contents:
- usually chosen manually
- less prone to errors
- hardly repeatable
- time-consuming
- sometimes subjective
A system capable of:
▶ extract keyphrases automatically from heterogeneous Web sources
▶ representing them in an ontology
▶ identify main topics in the ontology

The goals are:
▶ ascertain how well-suited are machine learning approaches for keyphrases extraction task
▶ represent the resulting knowledge into a well-defined knowledge base, that will provide support to a News Recommender System
Keyphrase Extraction Approaches

Four existing approaches:

- **Statistical**
  - simple (e.g., TF-IDF (Salton and Buckley (1988)))
  - non-linguistic features

- **Linguistic**
  - add value, e.g., by filtering ‘undesirable’ words (van der Plas et al. (2004))
  - normally uses mixed methods (linguistic + statistics)

- **Machine Learning**
  - usually uses supervised learning (Witten et al. (1999))

- **Hybrid**
  - combines the aforementioned methods (Medelyan (2009))
KEA & Neo4j

KEA¹ (*Keyword Extraction Algorithm*)

- includes a cut-down version of the Weka
- it can be either used
  - for free indexing
  - for indexing with controlled vocabulary

Neo4j²

- is a graph database
- nodes are interconnected with an undetermined number of relations between them

¹[http://www.nzdl.org/Kea](http://www.nzdl.org/Kea)
²[http://neo4j.org](http://neo4j.org)
System Overview

Keyphrases

Keyphrase Extraction Algorithm (KEA)

Stopwords

Aggregator

RSS feeds

Data

Graph Clustering

Triple Store

Database (MySql)

getInfo

saveInfo

Hernani Costa (CISUC)

ATCM

Coimbra, 13 July, 2012

6 / 13
# KEA’s Model Evaluation

<table>
<thead>
<tr>
<th>S</th>
<th>E</th>
<th>D(%)</th>
<th>T(%)</th>
<th>P(%)</th>
<th>R(%)</th>
<th>F₁(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Science</strong></td>
<td>85</td>
<td>20</td>
<td>70.0</td>
<td>30.0</td>
<td>85.4</td>
<td>57.1</td>
</tr>
<tr>
<td><strong>World</strong></td>
<td>632</td>
<td>41</td>
<td>70.7</td>
<td>29.3</td>
<td>66.6</td>
<td>63.0</td>
</tr>
<tr>
<td><strong>Economy</strong></td>
<td>35</td>
<td>24</td>
<td>70.8</td>
<td>29.2</td>
<td>60.7</td>
<td>48.7</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>30</td>
<td>22</td>
<td>72.3</td>
<td>27.3</td>
<td>47.9</td>
<td>72.2</td>
</tr>
<tr>
<td><strong>Macintosh</strong></td>
<td>13</td>
<td>10</td>
<td>70.0</td>
<td>30.0</td>
<td>46.8</td>
<td>68.3</td>
</tr>
<tr>
<td><strong>Video-DVD</strong></td>
<td>26</td>
<td>16</td>
<td>68.8</td>
<td>31.2</td>
<td>37.5</td>
<td>56.4</td>
</tr>
<tr>
<td><strong>Cinema</strong></td>
<td>19</td>
<td>16</td>
<td>68.8</td>
<td>31.2</td>
<td>27.5</td>
<td>39.1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>840</td>
<td>149³</td>
<td>70.3</td>
<td>29.7</td>
<td>53.2</td>
<td>57.8</td>
</tr>
</tbody>
</table>

S - news in the system  
E - news manually evaluated  
D - % used in the training process  
T - % used to test the algorithm  
P - precision  
R - recall  
F₁ - F measure

³ Sample set has a confidence level of 95% and a confidence interval of 7.3%.
Ontology Definition & Population

Neo4j as a Triple Store

Ontology definition

- RDF schema created in Protégé\(^a\)

Ontology Population

- triples
  - \( t_n = \{ \text{keyphrase}_n, \text{occursIn}, \text{news}_n \} \), where \( n \in N \)
  - gathering news from one week (i.e., 7 days * 24 hours)
    - 4683 nodes (keyphrases)
    - 18569 edges (connections between the keyphrases and the news)

\(^a\)http://protege.stanford.edu
Graph-Clustering
Concluding Remarks

- Some scores in the extraction process revealed $F_1$ scores higher than 60%
- Graph-Clustering can be considered as a successful proof of concept
- The prototype can be seen as a first step in the construction of an automatic keyphrase extract system
- Made for Portuguese, but easily adapted to other languages

Prototype system capable of:
- extract keyphrases from different sources
- formally represent the extracted knowledge into an ontology
- identify main topics
Future Work

- Extraction algorithm can be improved by using
  - **pre-filtering**
    - increasing the stopword list
    - the usage of a POS tagger + a grammar to create rules before the extraction process (Costa (2011))
  - **post-filtering**
    - identify the keyphrases POS (e.g., to discard verbs)
    - rate the keyphrases to improve the algorithm accuracy (Costa et al. (2010))
    - taking advantage of other resources, e.g., Onto.PT⁴ or BDpedia⁵

- Use more news to study the effect in the system accuracy

- Explore other keyphrase extraction algorithms and perform a comparison between them

⁴http://ontopt.dei.uc.pt
⁵http://dbpedia.org


The end