NLP for Decision Making

German Rigau
http://adimen.si.ehu.es/~rigau
Foreword

“Cuando creíamos que teníamos todas las respuestas, de pronto, cambiaron todas las preguntas.”

- Mario Benedetti

“When we thought we had all the answers, suddenly, they changed all the questions. ”

- Mario Benedetti
Foreword

- Where are the **answers** to the new (and old) questions?
  - Introspection? Experts? …
  - From many people? … “Wisdom of the Crowd”?  
  - Books, News, Tweets, … Textual Sources?
  - Multimedia sources? Images, Radio, TV …
  - Sensors? IoT? …
  - Anything? Everything?

- Information **overload** …
Foreword

- Information overload ...
  - infobesity, infoxication!
Foreword

- Information overload ...
  - infobesity, infoxication!
  - by Bertram Gross, *The Managing of Organizations: The administrative struggle* (1964)
Foreword

- Information overload ...

  - infobesity, infoxication!
  - by Bertram Gross, *The Managing of Organizations: The administrative struggle* (1964)
Foreword

- **Information overload ...**
  - infobesity, infoxication!
  - by Bertram Gross, *The Managing of Organizations: The administrative struggle* (1964)
  - Seneca complained that “the abundance of books is distraction” in the 1st century AD!
**Foreword**

- **Information overload** occurs when the amount of input to a system exceeds its processing capacity.
- Decision makers have fairly **limited** cognitive processing capacity.
- Consequently, when information overload occurs, it is likely that a **reduction** in decision quality will occur.
- From (Speier et al 1999)

- Always when **advances in technology** have increased a production of information.
Outline

- Big Data & NLP
- NewsReader
- Concluding Remarks
Big Data & NLP

- **Unstructured** digital content accounts for **90%** of all information [White paper IDC 2014] ...

- Usually in the form of **texts** and documents in **multiple languages** ...

- **Only** appropriate NLP tools can access this wealth of knowledge ...

- NLP among the **top 10** strategic technology trends for 2017 according to **Gartner**
Big Data & NLP

Because everybody knows that ...
Big Data & NLP

But in fact ...

KNOWLEDGE IS POWER
Big Data & NLP

e.g. IBM Watson ...

but also Google, Facebook, Amazon, Microsoft, ...
Big Data & NLP

- What happens in Internet every **second**? (July 2015)
Big Data & NLP

- What happens in Internet every **second**? (July 2015)
Big Data & NLP

http://tour-pedia.org/gui/demo/
The background in Google

Choi and Varian (2009): Predicting the Present with Google Trends

AR models with searches to predict the present beat those that do not include them

Examples: retail, vehicle and house sales, and tourism

Can searches help to forecast the future?
<table>
<thead>
<tr>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (today)</td>
<td>naruto: ending, ends, end</td>
<td>veterans day: closed, free, school</td>
<td>rosetta: landing, date, probe, register: join, free, closes</td>
<td>kyuhyun: concert, solo, release</td>
<td>galaxy note edge: coming, reaches, confirms</td>
<td>guangzhou: perform, attending, concert</td>
</tr>
<tr>
<td>11</td>
<td>veteran's day: closed, free, school</td>
<td>remembrance day: closed, ceremony, school</td>
<td>mexico: stoked, aired, broadcasted</td>
<td>concert hall tby: present #artday2014, diskony, enak</td>
<td>beyond the lights: see, watch, movies</td>
<td>city national civic auditorium:</td>
</tr>
<tr>
<td></td>
<td>school</td>
<td>t-mobile: coming, launch, selling</td>
<td>tahta: concert, follow @eventnamche, information</td>
<td>max day: give to, give, coming up</td>
<td>world diabetes day: join, wear, raise awareness</td>
<td>lkti:</td>
</tr>
<tr>
<td></td>
<td>livestream: school, watch, sleep</td>
<td>lumia: unveil, launch, reveal</td>
<td>monument valley: expanded, confirmed, receive</td>
<td>graha cakrawala: miss, ngajak,</td>
<td>balai sarbini: concert, buy, tiket</td>
<td>eeteuk: broadcasted, running, raichanxd</td>
</tr>
<tr>
<td>17</td>
<td>orlando: show, see, performing</td>
<td>meizu: launch, sends out invite, sent out</td>
<td>rochester: held, date, set</td>
<td>wii u: coming, launches, smash</td>
<td>america recycles day: recycling, celebrates, planning</td>
<td>janeiro brazil: <a href="http://t.co">http://t.co</a> /ckeaxwsgw for, tickets <a href="http://t.co">http://t.co</a> /6mkal8bkll, tickets <a href="http://t.co">http://t.co</a> /ckeaxng58</td>
</tr>
<tr>
<td>18</td>
<td>x-box: coming, game, play</td>
<td>boston: arena</td>
<td>19</td>
<td>meizu: launch, sends out invite, sent out</td>
<td>super smash bros:</td>
<td>open house: join, check out, campus</td>
</tr>
<tr>
<td>19</td>
<td>drake: listen, fall</td>
<td>baltimore: arena</td>
<td>20</td>
<td>rochester: held, date, set</td>
<td>set</td>
<td>21</td>
</tr>
</tbody>
</table>

http://ec2-54-170-89-29.eu-west-1.compute.amazonaws.com:8000
Big Data & NLP

- ... not only from Social Media / Internet.
- LexisNexis receives **daily** 1.5M news.
- CENDOJ stores 6M judicial sentences (0.6M/year)
- 5M Electronic Health Records (EHR) ...
- 0.2M Patents ...
- ?? Project proposals ...
- ...

- ... all kinds of e-documents ...
Pakistan plane carrying 48 crashes, reportedly on fire

Saddha Batohi (Pakistan) (AFP) - A Pakistani plane carrying 48 people crashed Wednesday in the country's mountainous north and burst into flames, authorities said, as rescue workers pulled dozens of bodies from the wreckage and officials expressed little hope for survivors.

http://emm.newsexplorer.eu/
Big Data & NLP

http://emm.newsbrief.eu/
Big Data & NLP

http://medisys.newsbrief.eu

NewsReader

Text Mining for Policy Making, 13th December 2016
Plan de Impulso de las Tecnologías del Lenguaje

Presentación del Plan

Video de presentación del Plan de Impulso de las Tecnologías del Lenguaje.

Big Data & NLP

- Deep and up-to-date knowledge of the ICT and BIO research and industrial sectors for policy makers design **effective public policies**

- Improve the procedure for granting and **monitoring public aid** to R+D+i in ICT and BIO sectors (more than a half research public aids)
Big Data & NLP

- Help designing Spanish R&D&i strategy

- Deep and up-to-date knowledge of the ICT and BIO research and industrial sectors for policy makers design effective public policies

- Improve the procedure for granting and monitoring public aid to R&D&i in ICT and BIO sectors (more than a half research public aids)
Big Data & NLP

- ICT health
- Oncology
- Pharmacology
- Hepatitis C
- Grants

Proyectos Sanidad Concedidos Plan Estatal 2013-2015

Proyectos ISCIII
Proyectos NO ISCIII
Big Data & NLP

World-Wide Integrated Crisis Early Warning System

Big Data & NLP

https://sites.google.com/site/distributedlittleredhen
Outline

- Big Data & NLP
- NewsReader
- Concluding Remarks
NewsReader

- Building structured event Indexes of large volumes of financial and economic Data for Decision Making
- VUA, UPV/EHU, FBK, LexisNexis, ScraperWiki.
- Main goal:
  - **StoryLines** in 4 languages at large scale
  - From NLP to Semantic Web (& Inference!)
- Focusing on:
  - Economy & Finance
  - Cross-lingual Event Mining
  - Monitoring car industry (10 years, 2.5M documents).
- http://www.newsreader-project.eu
- http://github.com/newsreader
NewsReader Project Objectives

- Event Detection in English, Dutch, Spanish and Italian
  - events in terms of **who** did **what**, **when** and **where**
  - **provenance:** who tells what and when
  - factuality / non-factual or speculative
  - **relations** between events
- Cross-lingual & Cross-document
- **Inferencing** about events & situations
- Narrative schemas, **timelines** & **storylines**
  - over long periods of time
- **Large-scale** processing, storage and retrieval of events (integrate the new with the old)
Storylines

LORD OF THE RINGS

NewsReader

Text Mining for Policy Making, 13th December 2016
End-to-end system

NLP-pipeline → Knowledge Store → Synerscope → Simple API

deep reading

NewsReader

Text Mining for Policy Making, 13th December 2016
In this week’s interview on Fox News Sunday, former U.S. President Bill Clinton described a question by the host Chris Wallace as a “conservative Hoax.”

According to the introduction given by Wallace, a pre-interview agreement to split the interview equally between questions on Clinton’s Global Initiative and any topic of the host’s choice.

After a few opening questions, Wallace raised the issue of Clinton’s efforts to deal with Osama bin Laden. "Why didn’t you do more to get bin Laden and al-Qaeda out of bush record?" He asked about the efforts to kill bin Laden and compering them with the record of the current administration before September 11, 2001. "They had no meetings on bin Laden."

Clinton defended his efforts to bin Laden as an "obsession," saying they were put in place because of the threat posed by terrorists. He also noted that the current administration had not moved fast enough to attack the Taliban in Afghanistan.

In the wake of the 9/11 attacks, Clinton stated that the U.S. should look to the future and focus on the threat posed by al-Qaeda and other terrorist groups. He emphasized the importance of strong leadership and a clear strategy to combat terrorism. His remarks were met with mixed reactions, with some praising his candor and others questioning his apparent lack of concern for the issue.

Entity: Osama_bin_Laden

https://knowledgestore.fbk.eu/
Large-scale processing

- LN car company news (EN) : ~ 6M (10 years)
  - 1st year: 63K ~ 2 weeks
  - 2nd year: 1.3M documents ~ 1 week
  - 3rd year: **2.5M** documents ~ 1 week
- FIFA world Cup (EN): ~200K
- TechCrunch (EN) : 43K articles
- Kiem (EN) : 212K documents
- Dutch House of Representatives (NL) : 1M documents
- Europarl (EN, ES) : ~19K EN, ~19K ES
- WikiNews (EN, ES, IT, NL) : 19K EN, 8K IT, 7K ES and 1K NL
- ECB+ (EN) : 984 articles
- Spanish Ministry of Industry (EN, ES): 200K patents
## Knowledge Store

<table>
<thead>
<tr>
<th>Topic</th>
<th>MEANTIME</th>
<th>WikiNews (Ver. 1)</th>
<th>WikiNews (Ver. 2)</th>
<th>Dutch Parliament</th>
<th>Cars (Ver. 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>Dutch</td>
<td>English</td>
</tr>
<tr>
<td>Pipeline Version</td>
<td>3.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0-dutch</td>
<td>3.0</td>
</tr>
<tr>
<td>News Articles</td>
<td>120</td>
<td>18,510</td>
<td>19,755</td>
<td>597,530</td>
<td>2,316,158</td>
</tr>
<tr>
<td>Mentions</td>
<td>35,237</td>
<td>2,629,176</td>
<td>5,206,202</td>
<td>9,231,113</td>
<td>842,639,827</td>
</tr>
<tr>
<td>Events</td>
<td>3,333</td>
<td>624,439</td>
<td>632,704</td>
<td>5,383,498</td>
<td>42,296,287</td>
</tr>
<tr>
<td>Entities</td>
<td>339</td>
<td>45,592</td>
<td>40,314</td>
<td>111,579</td>
<td>2,263,156</td>
</tr>
<tr>
<td>Persons</td>
<td>82</td>
<td>19,677</td>
<td>17,617</td>
<td>43,546</td>
<td>895,541</td>
</tr>
<tr>
<td>in DBpedia</td>
<td>46</td>
<td>9,744</td>
<td>10,784</td>
<td>13,942</td>
<td>126,140</td>
</tr>
<tr>
<td>Organizations</td>
<td>172</td>
<td>15,559</td>
<td>14,358</td>
<td>44,139</td>
<td>1,139,170</td>
</tr>
<tr>
<td>in DBpedia</td>
<td>115</td>
<td>6,317</td>
<td>4,940</td>
<td>12,907</td>
<td>44,458</td>
</tr>
<tr>
<td>Locations</td>
<td>85</td>
<td>10,356</td>
<td>8,339</td>
<td>23,894</td>
<td>228,445</td>
</tr>
<tr>
<td>in DBpedia</td>
<td>81</td>
<td>7,773</td>
<td>7,369</td>
<td>11,167</td>
<td>76,341</td>
</tr>
<tr>
<td>Triples</td>
<td>95,219,534</td>
<td>105,675,519</td>
<td>110,861,823</td>
<td>188,296,316</td>
<td>1,240,774,944</td>
</tr>
<tr>
<td>from Mentions</td>
<td>1,046,544</td>
<td>9,700,585</td>
<td>16,688,833</td>
<td>65,631,222</td>
<td>1,146,601,954</td>
</tr>
<tr>
<td>from DBpedia</td>
<td>94,172,990</td>
<td>95,974,934</td>
<td>94,172,990</td>
<td>122,665,094</td>
<td>94,172,990</td>
</tr>
<tr>
<td>distilled from</td>
<td>DBpedia 2015</td>
<td>DBpedia 2015</td>
<td>DBpedia 2015</td>
<td>DBpedia 2014</td>
<td>DBpedia 2015</td>
</tr>
</tbody>
</table>
Micro & Macro vision
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced
Event Detection: Example

Mariano Rajoy *announced* yesterday in Madrid that the budget cuts will continue next year.

**event1:** *announced*
- **who:** Mariano Rajoy
- **what:** *that the budget cuts will continue next year*
- **when:** yesterday
- **where:** in Madrid
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced
  who: Mariano Rajoy
  what: that the budget cuts will continue next year
  when: yesterday
  where: in Madrid

event2: cuts
  what: budget

event3: continue
  what: the budget cuts
  when: next year
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced
  who: Mariano Rajoy
  what: that the budget cuts will continue next year
  when: yesterday
  where: in Madrid

event2: cuts
  what: budget
  who? when? where?

event3: continue
  what: the budget cuts
  when: next year
  who? where?
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced
  who: Mariano Rajoy NERC (PER)
  what: that the budget cuts will continue next year
  when: yesterday
  where: in Madrid NERC (LOC)
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced
  who: Mariano Rajoy
    - http://dbpedia.org/resource/Mariano_Rajoy NED
  what: that the budget cuts will continue next year
  when: yesterday
  where: in Madrid
    - http://dbpedia.org/resource/Madrid NED
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced
  who: Mariano Rajoy
  what: that the budget cuts will continue next year
  when: yesterday
    ▪ 2016-11-30 TIMEX3
  where: in Madrid
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

event1: announced => announce.01
  who: Mariano Rajoy
    ▪ A0 = announcer = vn.Agent = fn.Speaker
  what: that the budget cuts will continue next year
    ▪ A1 = utterance = vn.Topic = fn.Message
  when: yesterday
    ▪ AM-TMP
  where: in Madrid
    ▪ AM-LOC

SRL + Predicate Matrix!
Event Detection: Example

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

Topics:

“economic convergence” : 0.030
“public finance” : 0.023
“budget estimate” : 0.022
“stability programme” : 0.021
“economic forecasting” : 0.021
“financial aid” : 0.019
Standards for inter-operability

- **NAF**
  - NLP Annotation Format
  - Basic format for inter-document NLP analysis
  - Stand-off XML, multi-layered annotation format
  - Allows parallel processing
  - Covers many linguistic levels
  - All NLP modules read and write NAF
IXA pipes: ready to use NLP tools

IXA pipes is a modular set of Natural Language Processing tools (or pipes) which provide easy access to NLP technology for several languages. It offers robust and efficient linguistic annotation to both researchers and non-NLP experts with the aim of lowering the barriers of using NLP technology either for research purposes or for small industrial developers and SMEs. The ixa pipes can be used or exploit its modularity to pick and change different components. The tools are developed by the IXA NLP Group of the University of the Basque Country.

ixa pipes

If you use the ixa pipes tools or the models, please cite this paper:


ixa-pipe-tok: Tokenizer and Segmenter for several languages.

ixa-pipe-pos: Statistical POS tagging and Lemmatizer for Basque, Dutch, English,

http://ixa2.si.ehu.es/ixa-pipes/
English generic pipeline v3.0

1) NAF libraries (EHU, VUA)
2, 3, 4) **ixa-pipes** (EHU): tok, pos, nerc
5, 6) ukb (EHU), vua-svm-wsd (VUA)
7) ixa-pipe-srl (mate tools + predicate-matrix):
   dependency parsing + SRL (N+V) + event-classification
8) ixa-pipe-parse (EHU): constituent parsing
9) corefgraph (EHU): nominal coreference
10) ixa-pipe-ned (DBpedia spotlight): Entity Linking
11) fbk-timepro (FBK): timex3
12) vua-factuality (VUA)
13) vua-opinion-miner (VUA)
14) vua-event-coref (VUA): intra-document event coreference
15) fbk-temprel (FBK)
16) fbk-causalrel (FBK)

http://ixa2.si.ehu.es/nrdemo/demo.php
English generic pipeline v3.0
Cross-lingual interoperability

- **Named entities**
  - Cross-lingual links from DBpedia

- **Events**
  - Predicate Matrix: Interoperable event models
    - VN, FN, PB, WN (SUMO, etc.), ESO
  - Also **nominal** predicates
  - Cross-lingual PM (EN, ES, NL, IT)

- **Time**
  - Cross-lingual time normalization

- **Concepts**
  - Cross-lingual wordnets: MCR, MultiWordNet, OMW

- **Same** semantic representations (ECKGs) for the same events in **different** languages!
### Cross-lingual Predicate Matrix

The image illustrates a cross-lingual predicate matrix, which is language-independent. The matrix is used to represent the structure of sentences in different languages with consistent categories and relations.

#### English Analysis

- **Predicate Matrix**
  - **Agent**: Steve Jobs
  - **Recipient**: WWDC at Moscone Center
  - **Communication**: Speech
  - **Place**: Moscone Center
  - **Time**: Monday

#### Spanish Analysis

- **Predicate Matrix**
  - **Agent**: Steve Jobs
  - **Recipient**: WWDC en el Moscone Center
  - **Communication**: Conferencia
  - **Place**: Moscone Center
  - **Time**: Lunes

---

### Language Independence

The predicate matrix is designed to be language-independent, allowing for the representation of similar structures across different languages.
Event detection highlights

- **Most complete and advanced** NLP pipelines
  - EN, ES, NL and IT
  - Basic NLP processing (**ixa-pipes**)
  - More **advanced** analysis ...
  - 64 NLP modules: EN **19**, ES **13+2**, NL **14**, IT **16**

- Cross-lingual semantic interoperability

- **Most complete multilingual evaluation** of NLP pipelines

- **Most robust** large-scale processing

- **Open source** modules and ready-to-use **VM images**
Big NLP

- SURFsara **Hadoop** cluster https://www.surfsara.nl
  - 170 data/computer nodes.
  - **1370** CPU-cores for parallel processing using YARN (MapReduce, Spark, Giraph)
  - 2.3 PB.
Big NLP

- SURFsara **Hadoop** cluster
  - https://hadoop.apache.org/
  - http://www.cascading.org
  - to support complex workflows on Hadoop

- LexisNexis provided **11 years** of financial news about the **car industry**
  - 2.5M news (a subset of the 6M)
  - 198,134 core hours (22.6 years on a single CPU!)
  - 141.5 hours (5.9 days)
  - 17,700 news per hour : 425,000 news per day
  - 4.4 min. per document
  - https://github.com/sara-nl/newsreader-hadoop
Outline

- Big Data & NLP
- NewsReader
- Concluding Remarks
Superhuman results

From Andy Steinbach (NVIDIA)
Superhuman results

Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores \( (\omega^T s_t) \). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

Deep visual-semantic alignments for generating image descriptions (2014)
Superhuman results

LipNet: Sentence Level Lipreading (2016)
Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas
TextRunner Search

Retrieved 377 results for Who built the pyramids?

Grouping results by predicate. Group by: argument 1 | argument 2

built - 35 results:

Egyptians (97), aliens (70), Pharaohs (40), 77 more... built the pyramids

Egyptians (26), Khufu (16), ancient Mayans (9), 32 more... built the Great Pyramid

Imhotep (8), Pharaoh Zoser (4), Zoser (2), 3 more... built the Step Pyramid

ancient king Sneferu (3), 12th dynasty kings (3), Snefru (3) built two pyramids

Egyptians (8) built the Great Pyramid

Sneferu (3), Snefru (3), pharaoh Sneferu (2) built three pyramids

ancient Egyptians (6) built more than 90 royal pyramids

quite drama (3), three pharaohs (2) building the Great Pyramid of Giza

colonial silver city of Taxco (3), Explore (2) built the gigantic pyramids of the Sun

Mercy (4) builds glass pyramid

Search again:

Who built the pyramids?

Search

Jump to:

built (35)
who built (9)
to build (11)
were building (4)
that built (3)
did not build (2)
had built (2)
could build (1)
depicts building (2)
NLU

- Current AI challenges
  - Natural Language Understanding
  - Image/video Understanding
  - Process/agents/services Understanding
  - Database Understanding
  - Web Understanding
  - Driving Understanding
  - ...

Text Mining for Policy Making, 13th December 2016
Applying for a job at IKEA

Make a chair and take a seat.
Final words?

- **Unstructured** digital content accounts for 90% of all information [White paper IDC 2014] ...
- **Only** appropriate NLP tools can access this wealth of knowledge ...
- NLP among the **top** 10 strategic technology trends for 2017 according to Gartner

- NLP is **NOT** solved!
- Much more **R+D investment** on NLP is needed to address this challenge
- NLP is **NOT** in the EU research agenda!
NLP for Decision Making

German Rigau
http://adimen.si.ehu.es/~rigau