Actors of the R & D System as covered in the Spanish Press

Carlos G. Figuerola, Tamar Groves, Ana V. Pérez, Bruno Maltrás, Miguel Angel Quintanilla
- ECyT Institute
University of Salamanca
MACAS Workshop, Stellenbosch - 16-18/09/2015

The SCSC

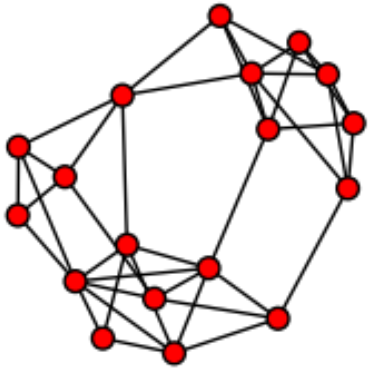
- The Spanish Corpus on Scientific Culture has about 50,000 news articles
- the articles come from three main spanish newspapers in their digital version, covering from 2002 until 2011

Topics in Science News

- high amount of news
- not suitable for manual analysis
- automatic procedures:
  - topic distillation
  - Social Network Analysis, community finding

Topics Discovering using SNA Techniques

- objects can have relationships between them
- we can map objects and relationships towards a network or graph
  - objets are nodes
  - relationships are edges or links between nodes



Establishing relationships between news

- we can compute semantic simmilarity between two documents
  - based on having words in common
  - (in reality is a bit more complicated)
- we can represent the corpus as a network
  - news articles are nodes
  - semantic simmilarity are links
  - links between two articles are more or less strong if news articles are more or less simmilar

Network of news

Detecting Communities

- in a network, a community is a bunch of nodes
  - strongly linked between themselves
  - weakly linked with nodes outside the bunch
- in our network of news, a community is a topic
- they are several algorithms to find communities in networks
- we use InfoMap: fast and efficient, accurate results
- we found 13 main communities or topics and we labelled them with a more or less expressive tag

Main Topics

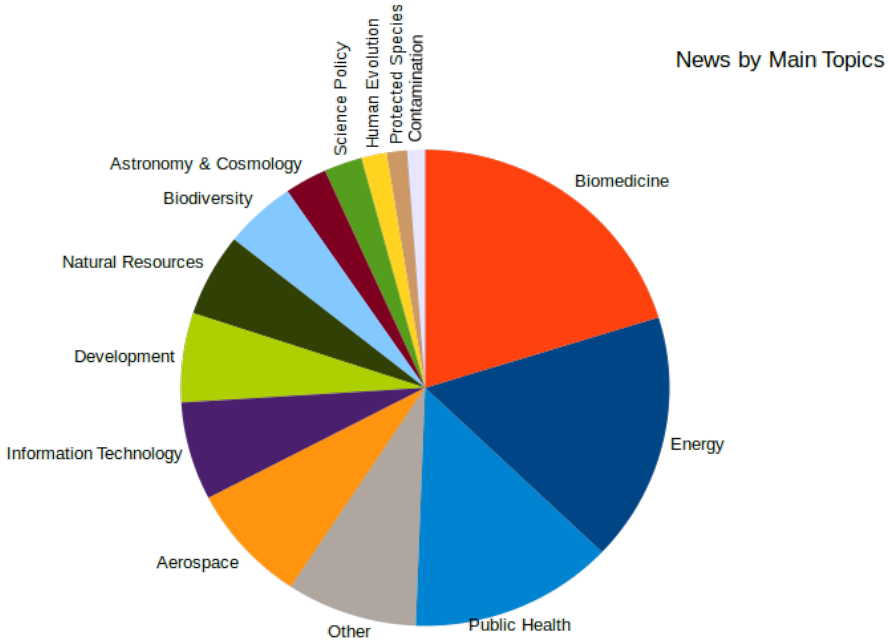
Communities listing

community	topic	community	topic
1	Public Health	8	Astronomy & Cosmology
2	Biomedicine	9	Information Technology
3	Energy	10	Science Policy
4	Human Development	11	Protected Species - Spain
5	Natural Resources	12	Human Evolution
6	Aerospace Research	13	Contamination
7	Biodiversity		

Subtopics also

Subcommunity	Topic	Subcommunity	Topic
1.1	influenza	1.11	infections, E. Coli,
1.2	AIDS	1.12	cholera
1.3	mortality	1.13	Legionella
1.4	drugs	1.14	polio
1.5	vaccines	1.15	mad cow disease
1.6	malaria	1.16	foot and mouth disease
1.7	SARS	1.17	dengue
1.8	tuberculosis	1.18	insect infections
1.9	hepatitis C	1.19	Chagas
1.10	antibiotics, bacteria	1.20	bio-bac

Amount of news per Topic

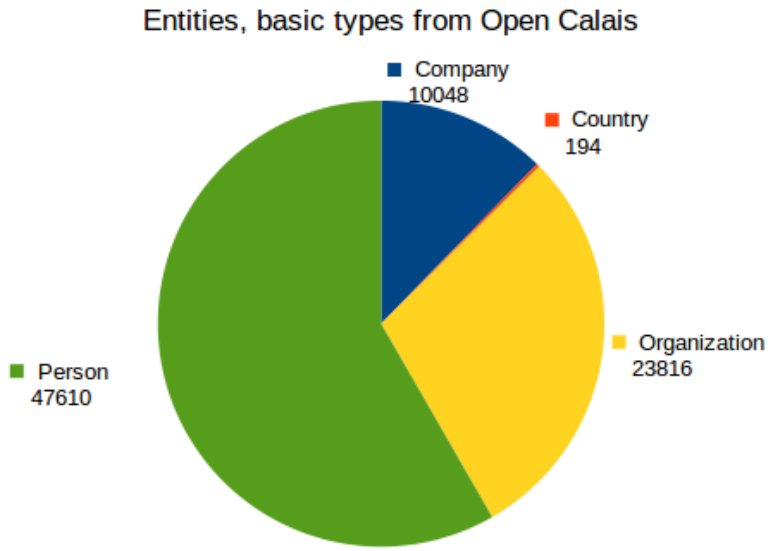


Actors

- we looked for individuals, organizations, companies in Science news
- again, too many documents to process them by hand
- NER (Named Entity Recognition) software techniques are available
- they involve patter recognition (use of uppercase, nouns linked by specific prepositions, etc.)
- also, knowledge about context to distinguish between persons, countries, organizations ...

OpenCalais

- we used OpenCalais a web service available from Thomson-Reuters
- it works for english, french and spanish, at the moment
- for spanish, it identifies entities and their basic type



However ...

We got a number of errors:

- entities misclassified. Example:
  - **Aurora Boreal** (*aurora borealis*) is not a woman, although Aurora is common name for a woman
- the same entity in a different form is not recognizede. Example:
  - **Consejo Superior de Investigación Científica**, **Consejo de Investigación Científica**, **Consejo**, **CSIC**

We were forced to perform a manual revision

Manual revision

- we got 86,000 distinct entities
- but only a small part of these are significantly frequent
  - only 637 persons appear in more than 15 news
  - 2, 695 organizations appear in more than 4 news
  - 808 companies appear in more than 2 news
- this manual revision allowed us to re-structure the basic types, to normalize names and to add some extra-information
- at the moment we have revised Organizations and Companies, as well as Countries
- revision, normalization and better categorization of persons is still in progress

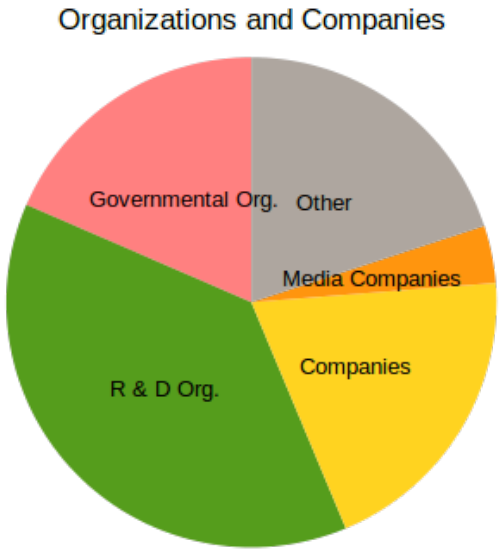
Countries


Europe
19864
N. America
14935
Spain
13465
Middle East
1598
Asia
13295
S.
America
4674
Africa
4674
Oceania
1762

Countries

- Spain and European Contries are the most cited en the news
- North America is also very relevant
- some geographical areas are more present in relation to specific topics
  - for example, bird flu is the topic of a lot of news with asian countries

Organizations & Companies



Organizations & Companies

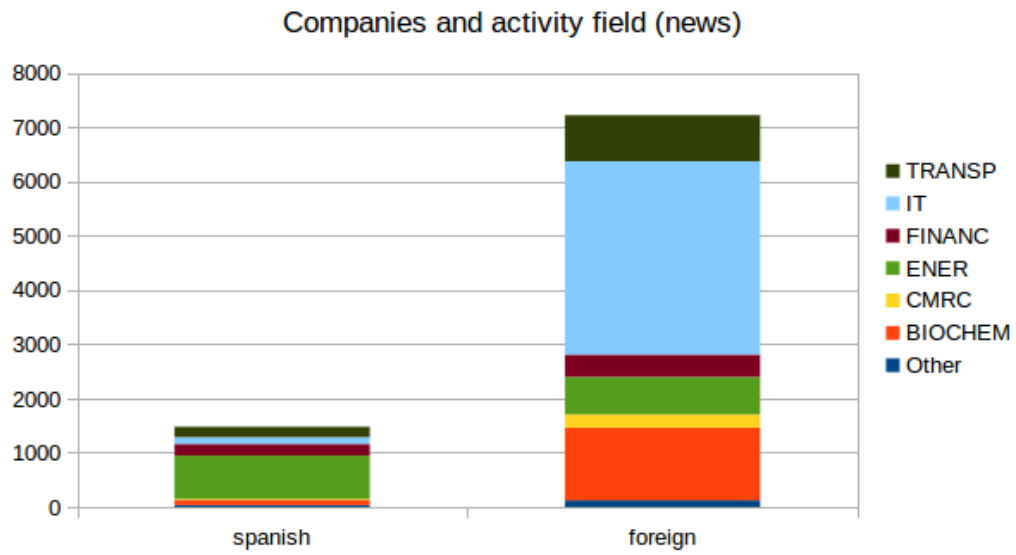
- R & D organizations are the big part (as expected)
  - R & D organizations do not include only research orgs.
  - also universities, hospitals and so on
  - most of them are public organizations
- Governmental organizations are an important part of the actors
- Companies are as many as Governmental organizations
- Media Companies refer in most cases to the source of news

Organizations & Companies

The Top Ten of Orgs. and Companies

Governmental Orgs	**R & D Orgs.**	**Companies**
World Health Organization	NASA	Microsoft
United Nat. Industrial Development Org.	European Space Agency	Google
European Union	Spanish Council for Scientific Res.	Iberdrola
United Nations	University of California	Sony
European Comission	Harvard University	Endesa
Spanish Ministry of Health	Universidad de Barcelona	Facebook
Spanish Ministry of Environment	Spanish Statistics Institute	Samsung
US Food & Drugs Administration	Universidad Autónoma de Madrid	Nokia
World Bank	Health Institute Carlos III	Twitter
Spanish Council of Nuclear Safety	Univerity of Washington	GlaxoSmithkline

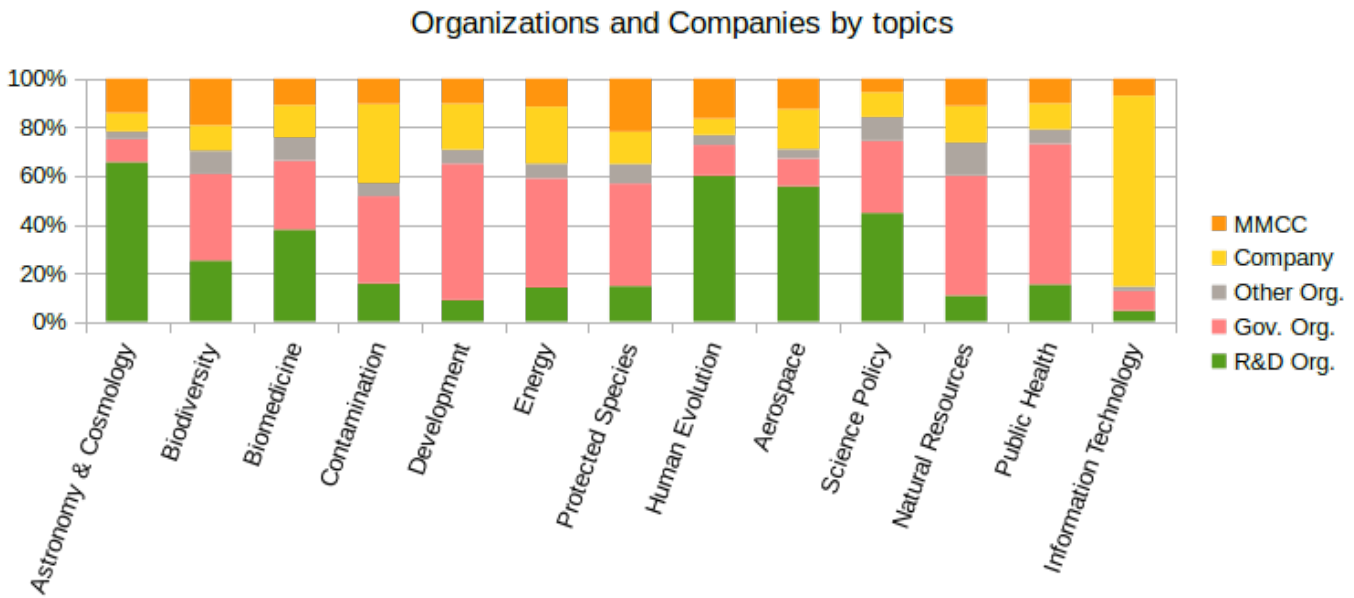
Companies, activity fields



Companies, activity fields

- foreign companies appear in many more news
- differential elements between spanish and foreign companies
  - Energy field is the most significant in Spanish companies
  - Information Technology companies are the most relevant in foreign comps.
  - Financial companies are more relevant among spanish companies than foreign ones
  - Biochemical ones are more relevant among foreign companies than spanish

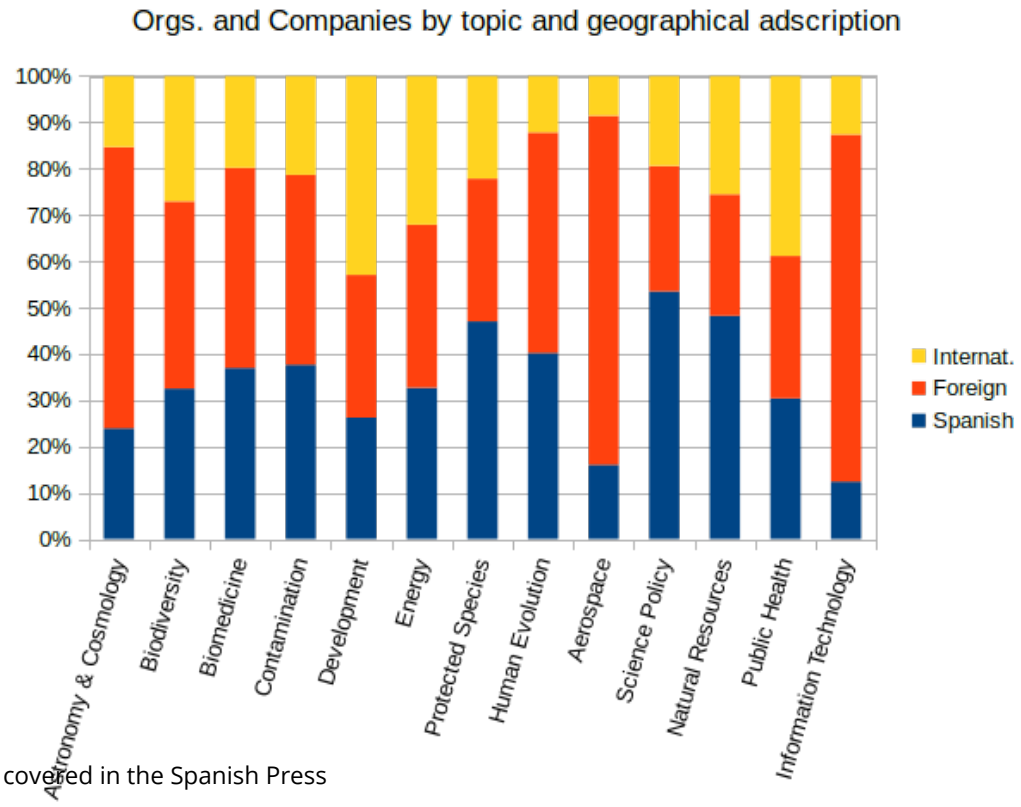
Topics, Organizations and Companies



Topics, Organizations and Companies

- companies or enterprises are more relevant in Information Technology
- companies are relevant in Contamination topics (maybe treated with criticism)
- R & D organizations are relevant in Astronomy, Human Evolution and Aerospace Research
- Governmental organizations are significant in Development, Public Health, Natural Resources and Energy
- Governmental and Research organizatios follow opposite paths
  - except in Science Policy and Biomedicine, where they are balanced

Nationality of organizations and Companies and Topics



Nationality of organizations and Companies and Topics

- Spanish organizations and companies have bigger presence in:
  - Science Policy, Natural Resources, Protected species
- Foreign organizations and companies are most relevant in:
  - Information technology, Aerospace Research, Astronomy
- International Organizations in:
  - Development, Public Health
  - also in Energy, Biodiversity, Natural Resources

Conclusions

- The data is not conclusive, but point in the same direction
  - for example, IT is important inside Companies, but is also important in foreign companies
  - topics of news can explain results given by Organizations and Companies analysis
  - for example, majority of R & D orgs. in topics related with 'intrinsec science' (Astronomy, Human Evolution ...)
- NER software is a good tool for high quantities of text, but requires manual revision and correction
- however, such manual revision is not so hard as one could think, as the most frequent entities are tractable quantity we did only a first look to obtained entities, but a deep analysis is still work in progress

Thank You!

Contact Information:

e-mail figue@usal.es

www ecyt.usal.es