MeSpEn: English-Spanish Machine Translation and Technologies

Marta Villegas, Ander Intxaurondo, Aitor Gonzalez-Agirre, Montserrat Marimon, Martin Krallinger

Spanish National Cancer Research Center (CNIO) / Barcelona Supercomputing Center (BSC)
2018/05/08 - MultilingualBIO Workshop - Miyazaki, Japan

ander.intxaurondo@bsc.es
Motivation

● Medical-related information resources for English:
  ○ Terminologies, literature, corpora, components and tools.
● All other languages are "underresourced".
● Medical MT:
  ○ Important for generating term translations, assist medical translators/interpreters.
  ○ Improvements in patient safety, diagnostic aid, integration of multilingual clinical information sources...
● MT systems trained on general-domain data perform poorly in the biomedical domain (Zeng-Treitler et al., 2010).
Related work

- Biomedical Translation Task at EMNLP WMT'18.
- SCIELO corpus by Neves et al.
- EMEA corpus.
- COOPA corpus.
MeSpEn (Medical Spanish English) Resource

Compilation of:

- Biomedical and clinical literature.
- Patient oriented web content.
- Bilingual glossaries.
Biomedical and clinical literature
Biomedical and clinical literature

Titles and abstracts of biomedical publications written in Spanish.

Sources:
- IBECS.
- SciELO.
- PubMed.

Harmonization of data in Dublin Core Schema (DC).
- Used to describe digital and physical resources
- Standard metadata in XML format.
Dublin Core Metadata Example

- `<metadata>`
    - `<dc:creator>`Sánchez Hernández, Lara</dc:creator>
    - `<dc:creator>`Rihuete Gálve, María Isabel</dc:creator>
    - `<dc:subject>`Desnutrición</dc:subject>
    - `<dc:subject>`Cáncer</dc:subject>
    - `<dc:subject>`Distorsión Sensorial</dc:subject>
    - `<dc:subject>`Gusto</dc:subject>
    - `<dc:subject>`Olfato</dc:subject>
    - `<dc:publisher>`Fundación Index</dc:publisher>
    - `<dc:source>`Index de Enfermería v.25 n.4 2016</dc:source>
    - `<dc:date>`2016-12-01</dc:date>
    - `<dc:type>`journal article</dc:type>
    - `<dc:format>`text/html</dc:format>
    - `<dc:language>`es</dc:language>
    - `<dc:title xml:lang="es">`Influencia de las distorsiones sensoriales sobre el estado nutricional del paciente oncológico</dc:title>`
    - `<dc:title xml:lang="en">`Influence of taste and smell dysfunction on oncology patients' nutrition</dc:title>`
    - `<dc:description xml:lang="es">`Objetivo: Determinar la influencia de la alteración del sabor y olor de los alimentos, secundaria al tratamiento antineoplásico, sobre la malnutrición, una de las complicaciones Metodología: Mediante una entrevista y el método de Valoración Subjetiva Global Generada por el Paciente (VSG-GP), se determinó la sintomatología y estado nutricional de 66 en el Hospital Universitario de Salamanca. Resultados: La desnutrición es frecuente (67,8%), y está relacionada con la pérdida de apetito y disminución de ingesta. Estos factores por la distorsión gustativa, la cual es frecuente (62,9%). Los resultados muestran que determinados hábitos higiénico-dietéticos podrían prevenir el desarrollo de distorsiones: alimentarias debidas a alteraciones del gusto parecen actuar indirectamente sobre el estado de desnutrición del paciente oncológico al ocasionar pérdida de apetito y disminución de ingesta.</dc:description>`
    - `<dc:description xml:lang="en">`Objective: To determine the influence of taste and smell disorders, a side effect of antineoplastic treatments, on nutrition of oncologic patients, amongst whom malnutrition is common. Methods: 62 patients from University Hospital of Salamanca, receiving treatment with chemotherapy, were included in our study. To investigate the nutritional state and secondary interview and used the Patient-Generated Subjective Global Assessment (PG-SGA) method. Results: Malnutrition is frequent (67.8%), and is related to the loss of appetite and reduced food intake. Disorders seem to indirectly affect the malnutrition status of oncologic patients by causing loss of appetite and reduced intake.</dc:description>`
  </oai_dc:dc>
</metadata>
IBECS

- Spanish Bibliographical Index in Health Sciences.
- Maintained by the Spanish National Health Sciences Library (Carlos III Health Institute).
- 168,198 biomedical records written in Spanish.
- Files encoded following the LILACS model.
  - Widely used for document indexing in Latin American and Caribbean countries.
- Straightforward mapping from LILACS into DC.
SciELO (1)

- Scientific Electronic Library Online.
- Supported by the Sao Paulo Research Foundation (FAPESP) and the Brazilian National Council for Scientific and Technological Development (BIREME).
- Complete free full text articles from Scientific Journals of Latin America, South Africa, Portugal and Spain.
- Free Open Access through OAI-PMH servers.
  - Standard metadata harvesting methods for record collection.

Spanish SciELO used for Biomedical track (10 teams) at the second language technology hackathon at the Mobile World Congress 2018:
http://www.hackathonplantl.es/
## SciELO Network statistics

### Publications in the SciELO network:

<table>
<thead>
<tr>
<th>Country</th>
<th>Journals</th>
<th>Publications</th>
<th>Publications in Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolivia</td>
<td>23</td>
<td>4,728</td>
<td>4,568</td>
</tr>
<tr>
<td>Brazil</td>
<td>396</td>
<td>2,554</td>
<td>73</td>
</tr>
<tr>
<td>Colombia</td>
<td>228</td>
<td>53,987</td>
<td>50,783</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>37</td>
<td>8,510</td>
<td>6,894</td>
</tr>
<tr>
<td>Cuba</td>
<td>70</td>
<td>32,702</td>
<td>31,745</td>
</tr>
<tr>
<td>Mexico</td>
<td>200</td>
<td>54,728</td>
<td>45,074</td>
</tr>
<tr>
<td>Paraguay</td>
<td>14</td>
<td>1,723</td>
<td>1,657</td>
</tr>
<tr>
<td>Portugal</td>
<td>70</td>
<td>10,704</td>
<td>482</td>
</tr>
<tr>
<td>South Africa</td>
<td>69</td>
<td>23,565</td>
<td>7</td>
</tr>
<tr>
<td>Spain</td>
<td>60</td>
<td>34,820</td>
<td>28,919</td>
</tr>
<tr>
<td>Uruguay</td>
<td>27</td>
<td>4,016</td>
<td>3,754</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,194</strong></td>
<td><strong>210,205</strong></td>
<td><strong>173,956</strong></td>
</tr>
</tbody>
</table>

### Publications in MeSpEn:

<table>
<thead>
<tr>
<th>Country</th>
<th>Publications in Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolivia</td>
<td>4,034</td>
</tr>
<tr>
<td>Brazil</td>
<td>39</td>
</tr>
<tr>
<td>Colombia</td>
<td>45,658</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>6,871</td>
</tr>
<tr>
<td>Cuba</td>
<td>25,694</td>
</tr>
<tr>
<td>Mexico</td>
<td>44,807</td>
</tr>
<tr>
<td>Paraguay</td>
<td>1,648</td>
</tr>
<tr>
<td>Portugal</td>
<td>1,648</td>
</tr>
<tr>
<td>South Africa</td>
<td>7</td>
</tr>
<tr>
<td>Spain</td>
<td>28,843</td>
</tr>
<tr>
<td>Uruguay</td>
<td>3,639</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>161,710</strong></td>
</tr>
</tbody>
</table>
PubMed

- Free search engine to access Medline.
- Maintained by the U.S. National Library of Medicine (NLM).
- 28 million publications in different languages.
  - 330,000 records written originally in Spanish.
  - Possible to retrieve the abstract from 127,619.
- Possibility to retrieve search results in XML format.
- Easy conversion to Dublin Core.
### Clinical literature statistics

<table>
<thead>
<tr>
<th>Collection</th>
<th>IBECS</th>
<th>SciELO</th>
<th>PubMed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total publications</td>
<td>168,198</td>
<td>161,710</td>
<td>330,928</td>
</tr>
<tr>
<td>Publications with titles in English and Spanish</td>
<td>114,798</td>
<td>119,820</td>
<td>300,690</td>
</tr>
<tr>
<td>Publications with abstracts in English and Spanish</td>
<td>100,824</td>
<td>119,722</td>
<td>4,147</td>
</tr>
<tr>
<td>Publications with titles and abstracts in English and Spanish</td>
<td>98,132</td>
<td>103,684</td>
<td>3,971</td>
</tr>
<tr>
<td>Number of titles in Spanish</td>
<td>155,094</td>
<td>158,600</td>
<td>300,690</td>
</tr>
<tr>
<td>Number of titles in English</td>
<td>166,469</td>
<td>120,913</td>
<td>330,928</td>
</tr>
<tr>
<td>Number of abstracts in Spanish</td>
<td>149,742</td>
<td>121,774</td>
<td>4,340</td>
</tr>
<tr>
<td>Number of abstracts in English</td>
<td>118,972</td>
<td>120,546</td>
<td>125,703</td>
</tr>
</tbody>
</table>
Web-content for patient information
MedlinePlus (1)

- Online free information about health in English and Spanish.
- Provided by the U.S. National Library of Medicine.
- Information about:
  - Health topics (1063 articles).
  - Drugs and supplements (1494).
  - Laboratory test information (50).
  - Medical encyclopedia (4426).
- 7,033 articles.
MedlinePlus (2)

● Health topics into Dublin Core.
  ○ Metadata stored in source code.
● High amount of information in the rest of the topics.
● For each topic:
  ○ Clean raw text in English.
  ○ TEI file with text in English.
  ○ Clean raw text in Spanish.
  ○ TEI file with text in Spanish.
● Text Encoding Initiative (TEI).
  ○ Standard representation of texts in digital format.
  ○ Used in libraries, digital text collections, and for linguistic corpora creation.
Myelodysplastic syndrome is a group of disorders when the blood cells produced in the bone marrow do not mature properly. Myelodysplastic syndrome (MDS) is a form of cancer. In about a third of people, MDS may develop into acute myeloid leukemia.

Causes

Stem cells in bone marrow form different types of blood cells. With MDS, the DNA in stem cells becomes damaged. The exact cause of MDS is not known. For most cases, there is no known cause.

Risk factors for MDS include:

- Certain genetic disorders
- Exposure to environmental or industrial chemicals, fertilizers, pesticides, solvents, or heavy metals
- Smoking

Prior cancer treatment increases the risk for MDS. This is called secondary or treatment-related MDS.
Medical bilingual glossaries
Medical bilingual glossaries

- 46 bilingual glossaries for various language pairs (> 400 medical translators).
- Made from free online medical glossaries and dictionaries.
- Encoded in TSV format.
  - 26 files include English terms.
  - 8 files include Spanish terms.
  - 13 files include other languages.
# Medical bilingual glossaries, statistics

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Terms</th>
<th>Language pair</th>
<th>Terms</th>
<th>Language pair</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Spanish</td>
<td>123,788</td>
<td>Latin-Russian</td>
<td>2,486</td>
<td>German-Russian</td>
<td>225</td>
</tr>
<tr>
<td>English-Korean</td>
<td>69,368</td>
<td>German-Swedish</td>
<td>2,208</td>
<td>English-Romanian</td>
<td>205</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>66,939</td>
<td>German-Portuguese</td>
<td>2,028</td>
<td>Italian-Spanish</td>
<td>196</td>
</tr>
<tr>
<td>English-Italian</td>
<td>24,155</td>
<td>English-Swedish</td>
<td>1,067</td>
<td>Danish-English</td>
<td>193</td>
</tr>
<tr>
<td>English-German</td>
<td>18,534</td>
<td>German-Italian</td>
<td>976</td>
<td>French-Spanish</td>
<td>179</td>
</tr>
<tr>
<td>English-Japanese</td>
<td>18,320</td>
<td>English-Slovenian</td>
<td>945</td>
<td>Danish-Polish</td>
<td>166</td>
</tr>
<tr>
<td>Arabic-English</td>
<td>9,384</td>
<td>Bengali-English</td>
<td>841</td>
<td>Russian-Spanish</td>
<td>122</td>
</tr>
<tr>
<td>English-Turkish</td>
<td>7,675</td>
<td>English-Thai</td>
<td>835</td>
<td>English-Hindy</td>
<td>120</td>
</tr>
<tr>
<td>German-Spanish</td>
<td>7,004</td>
<td>Dutch-French</td>
<td>585</td>
<td>French-German</td>
<td>119</td>
</tr>
<tr>
<td>Dutch-English</td>
<td>6,878</td>
<td>English-Indonesian</td>
<td>491</td>
<td>French-Italian</td>
<td>117</td>
</tr>
<tr>
<td>English-French</td>
<td>6,571</td>
<td>Bulgarian-English</td>
<td>347</td>
<td>Croatian-German</td>
<td>115</td>
</tr>
<tr>
<td>English-Russian</td>
<td>4,346</td>
<td>Croatian-English</td>
<td>339</td>
<td>German-Romanian</td>
<td>109</td>
</tr>
<tr>
<td>English-Polish</td>
<td>3,727</td>
<td>Polish-Spanish</td>
<td>271</td>
<td>Dutch-Spanish</td>
<td>70</td>
</tr>
<tr>
<td>English-Hungarian</td>
<td>2,711</td>
<td>Dutch-Turkish</td>
<td>238</td>
<td>Portuguese-Spanish</td>
<td>61</td>
</tr>
<tr>
<td>English-Greek</td>
<td>2,626</td>
<td>Latin-Polish</td>
<td>237</td>
<td>English-Norwegian</td>
<td>44</td>
</tr>
<tr>
<td>English-Portuguese</td>
<td>2,517</td>
<td>Croatian-French</td>
<td>235</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>390,713</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Explorative use of MeSpEn
Explorative use of MeSpEn (1)

Automatic enrichment of UMLS.

- Suggesting new Spanish term translation candidates.
- Suggesting synonyms of already existing terms in Spanish.

Steps:

1. Identify automatically UMLS terms in English texts.
2. Align words in the English text with corresponding translations in the Spanish text.
3. Detect candidate UMLS terms for Spanish.
1. Identify automatically UMLS terms in English texts with cTakes:
   - Assign most-likely CUIs (UMLS concept identifiers).
Explorative use of MeSpEn (3)

2. Align words in the English text with corresponding translations in the Spanish text:

- GIZA++

  “Adregenetic beta receptor blockaders in arterial hypertension” →
  “Drogas betabloqueantes en hipertensión arterial”

# Sentence pair (1955197) source length 7 target length 5 alignment score : 3.89774e-08
Drogas betabloqueantes en hipertensión arterial
NULL ( { } ) Adrenergic ( { 1 } ) beta ( { } ) receptor ( { } ) blockaders ( { 2 } ) in ( { 3 } ) arterial ( { 5 } ) hypertension ( { 4 } )
Explorative use of MeSpEn (4)

3. Detect candidate UMLS terms for Spanish:
   - Predict candidate text span with potential translated terms in Spanish.
   - Assign English CUIs to terms.

   ![Diagram]

   - Initial sample set evaluation:
     - 200 candidate terms manually validated.
     - 47% correct.
     - 22% general terms (hyponym / hypernym).
     - Average validation time per term: 2.03 seconds (MyMiner).
Conclusion
Conclusion

- Complementary resources for medical MT in English and Spanish.
  - Biomedical and clinical literature.
  - Web-content for patient information.
  - Bilingual glossaries.
- Very useful for expansion of terminological resources for the medical domain.
  - Detection of UMLS concepts between languages.
- Limitation: does not cover co-official languages: Basque, Catalan and Galician.
  - Few publications in Catalan in SciELO (8) and PubMed (88).
  - Need to derive glossaries for Basque and Spanish.
- Analyze additional resources (patents, drug labels,...), improve sentence alignment, etc.

http://temu.bsc.es/mespen
Thanks!

Ander Intxaurrondo

Spanish National Cancer Research Center (CNIO) / Barcelona Supercomputing Center (BSC).

ander.intxaurrondo@bsc.es