MeSpEn: English-Spanish Machine Translation and Technologies

Marta Villegas, **Ander Intxaurrondo**, 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.intxaurrondo@bsc.es

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

cnio Centro Nacional de Investigaciones Oncológicas

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

LREC 2018 MIYAZAKI
Spring 2018
Phoenix Seagaia Resort
Miyazaki, Japan

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).

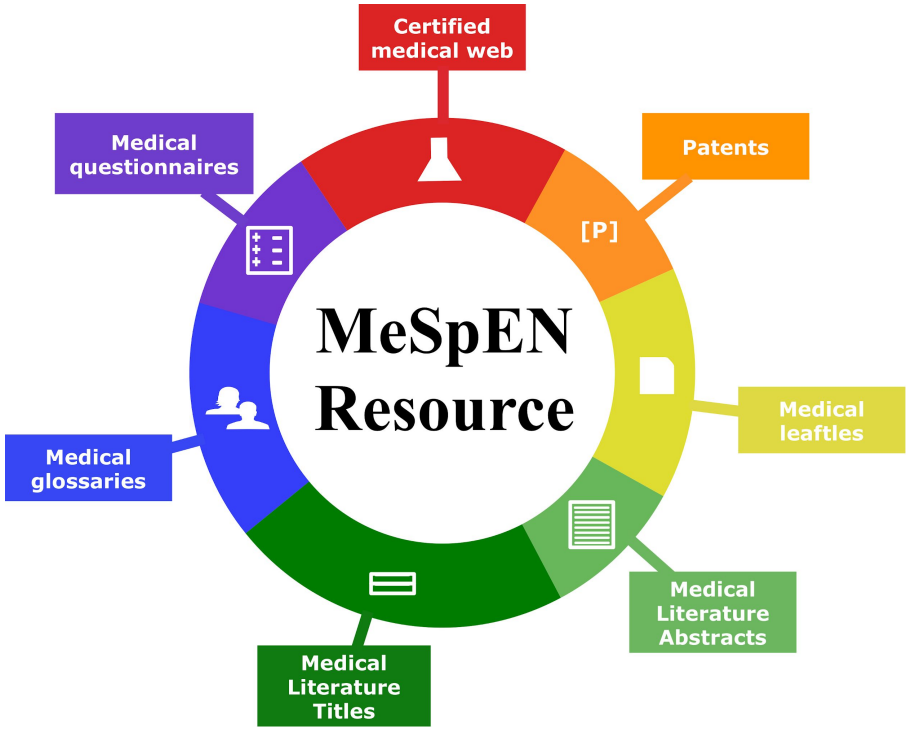
Related work

- 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>
  −<oai-dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:creator>Sánchez Hernández,Lara</dc:creator>
      <dc:creator>Rihuete Galve,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:identifier>
        http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1132-12962016000300003
      </dc:identifier>
      <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 6
        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 facto
        por la distorsión gustativa, la cual es frecuente (62,9%). Los resultados muestran que determinados hábitos higiénico-dietéticos podrían prevenir el desarrollo de distorsiones s
        alimentarias debidas a alteraciones del gusto parecen actuar indirectamente sobre el estado de desnutrición del paciente oncológico al ocasionar pérdida de apetito y disminu
      </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 o
        Methods: 62 patients from University Hospital of Salamanca, receiving treatment with chemotherapy, were included in our study. To investigate the nutritional state and secon
        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 re
        influenced by taste disorder, which is also frequent among patients (62.9%). The results show that hygiene-dietetic habits could prevent the development of sensorial disorders.
        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 techonology hackathon at the Mobile World Congress 2018: http://www.hackathonplantl.es/

SciELO Network statistics

Publications in the SciELO network:

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

Publications in MeSpEn:

Country	Publications in Spanish
Bolivia	4,034
Brazil	39
Colombia	45,658
Costa Rica	6,871
Cuba	25,694
Mexico	44,807
Paraguay	1,648
Portugal	1,648
South Africa	7
Spain	28,843
Uruguay	3,639
Total	**161,710**

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

Collection	IBECS	SciELO	PubMed
Total publications	168,198	161,710	330,928
Publications with titles in English and Spanish	114,798	119,820	300,690
Publications with abstracts in English and Spanish	100,824	119,722	4,147
Publications with titles and abstracts in English and Spanish	98,132	103,684	3,971
Number of titles in Spanish	155,094	158,600	300,690
Number of titles in English	166,469	120,913	330,928
Number of abstracts in Spanish	149,742	121,774	4,340
Number of abstracts in English	118,972	120,546	125,703

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.

TEI example

```
      −<title>
          Myelodysplastic syndrome: MedlinePlus Medical Encyclopedia
        </title>
        <author>MedlinePlus</author>
      </titleStmt>
    </fileDesc>
  </teiHeader>
−<text>
  −<body>
    −<div n="1" type="section">
      −<p>
          Myelodysplastic syndrome is a group of disorders when the blood cells produced in the bone marrow do not mature
          matured may not function properly.
        </p>
      −<p>
          Myelodysplastic syndrome (MDS) is a form of cancer. In about a third of people, MDS may develop into acute myeloi
        </p>
      </div>
    −<div n="2" type="section">
        <head>Causes</head>
      −<p>
          Stem cells in bone marrow form different types of blood cells. With MDS, the DNA in stem cells becomes damaged. B
        </p>
      −<p>
          The exact cause of MDS is not known. For most cases, there is no known cause.
        </p>
        <p>Risk factors for MDS include:</p>
      −<list type="bulleted">
          <item>Certain genetic disorders</item>
        −<item>
            Exposure to environmental or industrial chemicals, fertilizers, pesticides, solvents, or heavy metals
          </item>
          <item>Smoking</item>
        </list>
      −<p>
          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

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

Explorative use of MeSpEn

Explorative use of MeSpEn (1)

Automatic enrichment of UMLS.

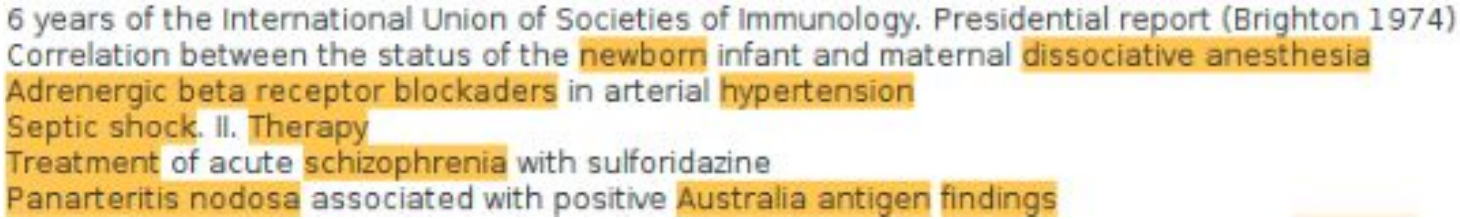
- Suggesting new Spanish term translation candidates.
- Suggesting synonyms of already existing terms in Spanish.

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.

Explorative use of MeSpEn (2)

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++

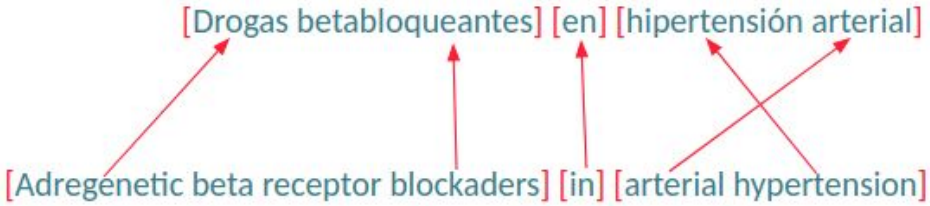
“Adregenetic beta receptor blockaders in arterial hypertension” → “Drogas betabloqueantes en hipertensión arterial” →

```
# 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.



- Initial sample set evaluation:
    - 200 candidate terms manually validated.
    - 47% correct.
    - 22% general terms (hypernym / hyponym).
    - 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

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje