Una metodología semiautomática de anotación de entidades nombradas para la creación de un *gold standard*

Susana Sotelo[†] Pablo Gamallo[†] Álvaro Iriarte[‡]

[†] CiTIUS – Universidade de Santiago

[‡] CEHUM – Universidade do Minho

V Congreso de la Soc. Internacional de Humanidades Digitales Hispánicas

Santiago de Compostela, 4-8 de octubre de 2021

Índice

1 Contexto y objetivos

2 Corpus

Descripción

Proceso de codificación

Proceso de anotación de entidades nombradas

3 Explotación

Contexto

Proyectos de investigación:

- *Discursos, imágenes y prácticas culturales de Santiago de Compostela como meta del Camino* (FFI2012-35521)
- *Narrativas, usos y consumos de visitantes como aliados o amenazas para el bienestar de la comunidad local: el caso de Santiago de Compostela* (FFI2017-88196-R)

Objetivo:

- Identificar las prácticas efectivas de los visitantes de Santiago de Compostela y cruzarlas con aquellas presentes en productos culturales y literarios
- Analizar el impacto de dichas prácticas sobre la comunidad local

Recursos desarrollados:

1. Corpus literario: catálogo de 546 libros con discursos y prácticas culturales relativas a Santiago de Compostela o al Camino de Santiago
2. Redes sociales: en torno a 350.000 *tweets* mencionando el Camino de Santiago
3. Base de datos de encuestas a visitantes de Santiago de Compostela y transcripción de las entrevistas telefónicas a parte de los encuestados

Contexto

Proyectos de investigación:

- *Discursos, imágenes y prácticas culturales de Santiago de Compostela como meta del Camino* (FFI2012-35521)
- *Narrativas, usos y consumos de visitantes como aliados o amenazas para el bienestar de la comunidad local: el caso de Santiago de Compostela* (FFI2017-88196-R)

Objetivo:

- Identificar las prácticas efectivas de los visitantes de Santiago de Compostela y cruzarlas con aquellas presentes en productos culturales y literarios
- Analizar el impacto de dichas prácticas sobre la comunidad local

Recursos desarrollados:

1. Corpus literario: catálogo de 546 libros con discursos y prácticas culturales relativas a Santiago de Compostela o al Camino de Santiago
2. Redes sociales: en torno a 350.000 *tweets* mencionando el Camino de Santiago
3. Base de datos de encuestas a visitantes de Santiago de Compostela y transcripción de las entrevistas telefónicas a parte de los encuestados

Contexto

Proyectos de investigación:

- *Discursos, imágenes y prácticas culturales de Santiago de Compostela como meta del Camino* (FFI2012-35521)
- *Narrativas, usos y consumos de visitantes como aliados o amenazas para el bienestar de la comunidad local: el caso de Santiago de Compostela* (FFI2017-88196-R)

Objetivo:

- Identificar las prácticas efectivas de los visitantes de Santiago de Compostela y cruzarlas con aquellas presentes en productos culturales y literarios
- Analizar el impacto de dichas prácticas sobre la comunidad local

Recursos desarrollados:

1. Corpus literario: catálogo de 546 libros con discursos y prácticas culturales relativas a Santiago de Compostela o al Camino de Santiago
2. Redes sociales: en torno a 350.000 *tweets* mencionando el Camino de Santiago
3. Base de datos de encuestas a visitantes de Santiago de Compostela y transcripción de las entrevistas telefónicas a parte de los encuestados

Descripción del corpus

1. Base de datos con 2080 encuestas a visitantes de la ciudad de Santiago
   - Muestreo no probabilístico accidental (por lugar de residencia)
   - Encuestas a pie de calle en puntos de interés turístico, zonas de esparcimiento y terminales de transporte, entre 2013 y 2014
2. Corpus de 252 entrevistas telefónicas
   - Semiestructuradas
   - En tres idiomas: español, gallego y portugués

Composición:

Por origen del informante:

	Encuestas	Porcentaje	Entrevistas	Porcentaje
Galicia	398	19,13 %	47	18,65 %
España	878	42,21 %	92	36,51 %
Portugal	408	19,62 %	54	21,43 %
Brasil	396	19,04 %	59	23,41 %
Total	**2080**	**100,00 %**	**252**	**100,00 %**

Por lengua y número de palabras:

Lengua	Palabras	Porcentaje
gallego	177.999	12,01 %
español	655.112	44,22 %
portugués	648.404	43,77 %

Descripción del corpus

1. Base de datos con 2080 encuestas a visitantes de la ciudad de Santiago
   - Muestreo no probabilístico accidental (por lugar de residencia)
   - Encuestas a pie de calle en puntos de interés turístico, zonas de esparcimiento y terminales de transporte, entre 2013 y 2014
2. Corpus de 252 entrevistas telefónicas
   - Semiestructuradas
   - En tres idiomas: español, gallego y portugués

Composición:

Por origen del informante:

	Encuestas	Porcentaje	Entrevistas	Porcentaje
Galicia	398	19,13 %	47	18,65 %
España	878	42,21 %	92	36,51 %
Portugal	408	19,62 %	54	21,43 %
Brasil	396	19,04 %	59	23,41 %
Total	**2080**	**100,00 %**	**252**	**100,00 %**

Por lengua y número de palabras:

Lengua	Palabras	Porcentaje
gallego	177.999	12,01 %
español	655.112	44,22 %
portugués	648.404	43,77 %

Descripción del corpus

1. Base de datos con 2080 encuestas a visitantes de la ciudad de Santiago
   - Muestreo no probabilístico accidental (por lugar de residencia)
   - Encuestas a pie de calle en puntos de interés turístico, zonas de esparcimiento y terminales de transporte, entre 2013 y 2014
2. Corpus de 252 entrevistas telefónicas
   - Semiestructuradas
   - En tres idiomas: español, gallego y portugués

Composición:

Por origen del informante:

	Encuestas	Porcentaje	Entrevistas	Porcentaje
Galicia	398	19,13 %	47	18,65 %
España	878	42,21 %	92	36,51 %
Portugal	408	19,62 %	54	21,43 %
Brasil	396	19,04 %	59	23,41 %
Total	**2080**	**100,00 %**	**252**	**100,00 %**

Por lengua y número de palabras:

Lengua	Palabras	Porcentaje
gallego	177.999	12,01 %
español	655.112	44,22 %
portugués	648.404	43,77 %

Proceso de codificación

Fases de codificación

1. Transcripción a partir de grabaciones
   - Sin representación de mecanismos propios de la oralidad (vacilaciones, interrupciones, etc.)
   - Realizada por un equipo heterogéneo de personas
2. Corrección y armonización de los textos
   - Corrección y normalización ortográfica
   - Fijación de omisiones y errores por distancia cultural del transcriptor

     Algunos ejemplos: *fevereiro* en lugar de *Cebreiro*, *fui no ferry* por *Finisterre* o *farol* por *Ferrol*
3. Codificación en XML con esquema propio:

   `https://galabra.linguarum.net/corpus/schemadoc/`
   - Sección con metadatos
   - Identificación de turnos (entrevistador/entrevistado) y lengua de cada turno

Proceso de codificación

Fases de codificación

1. Transcripción a partir de grabaciones
   - Sin representación de mecanismos propios de la oralidad (vacilaciones, interrupciones, etc.)
   - Realizada por un equipo heterogéneo de personas
2. Corrección y armonización de los textos
   - Corrección y normalización ortográfica
   - Fijación de omisiones y errores por distancia cultural del transcriptor

     Algunos ejemplos: *fevereiro* en lugar de *Cebreiro*, *fui no ferry* por *Finisterre* o *farol* por *Ferrol*
3. Codificación en XML con esquema propio:

   `https://galabra.linguarum.net/corpus/schemadoc/`
   - Sección con metadatos
   - Identificación de turnos (entrevistador/entrevistado) y lengua de cada turno

Proceso de codificación

Fases de codificación

1. Transcripción a partir de grabaciones
   - Sin representación de mecanismos propios de la oralidad (vacilaciones, interrupciones, etc.)
   - Realizada por un equipo heterogéneo de personas
2. Corrección y armonización de los textos
   - Corrección y normalización ortográfica
   - Fijación de omisiones y errores por distancia cultural del transcriptor

     Algunos ejemplos: *fevereiro* en lugar de *Cebreiro*, *fui no ferry* por *Finisterre* o *farol* por *Ferrol*
3. Codificación en XML con esquema propio:

   `https://galabra.linguarum.net/corpus/schemadoc/`
   - Sección con metadatos
   - Identificación de turnos (entrevistador/entrevistado) y lengua de cada turno

Proceso de codificación

```
1  <?xml version="1.0" encoding="utf-8" standalone="no" ?>
2  <transcription xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="transcription.xsd">
3  <metadata>
4      <participants>
5          <speaker id="119">
6              <age value="2"/>
7              <gender value="H"/>
8              <grade value="7"/>
9              <pilgrim value="1"/>
10             <origin value="PT"/>
11         </speaker>
12         <researcher id="5" name="Ana Rita"/>
13     </participants>
14     <source phase="1">
15         <text>
16             <file name="Fator09_FASE1_PT_119_CristianoRodrigues_20-02-2014.doc" type="doc"/>
17         </text>
18         <audio duration="2701">
19             <file name="00_PT_fase1_119.mp3" type="mp3"/>
20         </audio>
21     </source>
22 </metadata>
23 <data>
24 <q lang="pt">Bom dia. O meu nome é Ana Rita. Ligo-lhe desde a Universidade de Santiago. Esteve a trocar uns emails comigo.</q>
25 <a lang="pt">Sim, estou recordado. Por causa de um estudo?</a>
26 <q lang="pt">Exatamente, sobre Santiago. Gostaríamos de saber um pouco mais de aquilo que disse ao meu colega.</q>
27 [...]
```

Proceso de anotación

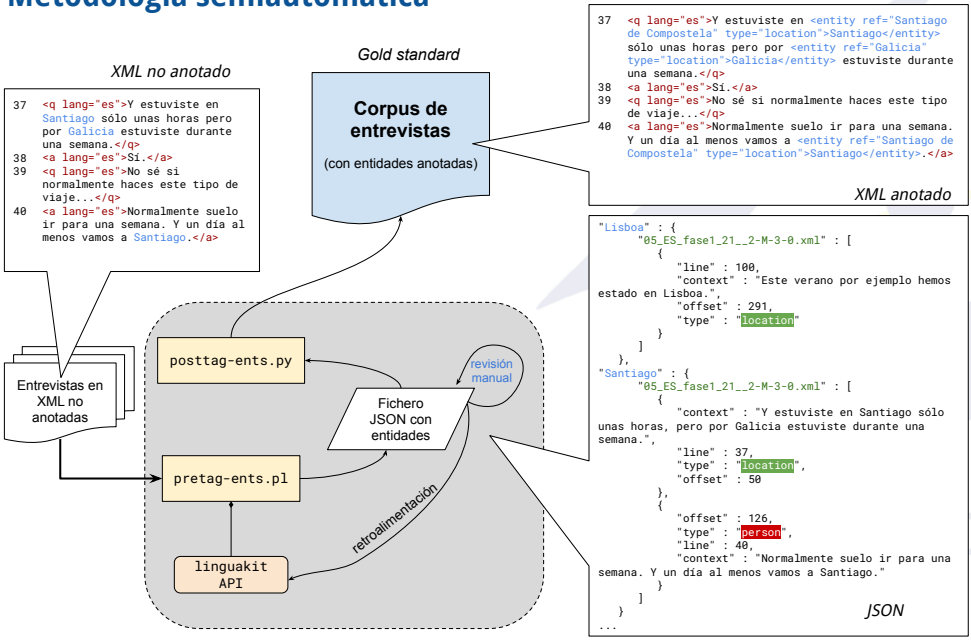
Anotación NER

Tarea del Procesamiento del Lenguaje Natural que tiene por objeto la detección y clasificación de entidades nombradas en un corpus textual.

1. Herramienta: Linguakit → `https://github.com/citiususc/Linguakit`
   - Entidades *numex* (de base numérica) y *enamex* (nombres propios)
   - Método de supervisión distante para la clasificación en 4 tipos: *persona*, *organización*, *localización* y *miscelánea*
   - Soporte para todas las lenguas del proyecto: español, gallego y portugués
   - Software libre
2. Resultados

	español		gallego		portugués		Total
location	18323	85.79 %	5921	83.36 %	12183	85.71 %	36427
person	1207	5.63 %	370	5.21 %	1063	7.48 %	2640
org	436	2.05 %	200	2.81 %	399	2.81 %	1035
misc	1393	6.53 %	612	8.62 %	568	4.00 %	2573
Total	**21359**	**100 %**	**7103**	**100 %**	**14213**	**100 %**	**42675**

Metodología semiautomática



Proceso de anotación

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<transcription xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="transcription.xsd">
<metadata>
    <participants>
        <speaker id="119">
            <age value="2"/>
            <gender value="H"/>
            <grade value="7"/>
            <pilgrim value="1"/>
            <origin value="PT"/>
        </speaker>
        <researcher id="5" name="Ana Rita"/>
    </participants>
    <source phase="1">
        <text>
            <file name="Fator09_FASE1_PT_119_CristianoRodrigues_20-02-2014.doc" type="doc"/>
        </text>
        <audio duration="2701">
            <file name="00_PT_fase1_119.mp3" type="mp3"/>
        </audio>
    </source>
</metadata>
<data>
<a lang="pt">Sim, estou recordado. Por causa de um estudo?</a>
<q lang="pt">Bom dia. O meu nome é <entity ref="Ana Rita" type="person">Ana Rita</entity>. Ligo-lhe desde a <entity ref="
    Universidade de Santiago" type="org">Universidade de Santiago</entity>. Esteve a trocar uns emails comigo.</q>
<q lang="pt">Exatamente, sobre <entity ref="Santiago de Compostela" type="location">Santiago</entity>. Gostaríamos de saber um pouco
    mais de aquilo que disse ao meu colega.</q>
[...]
```

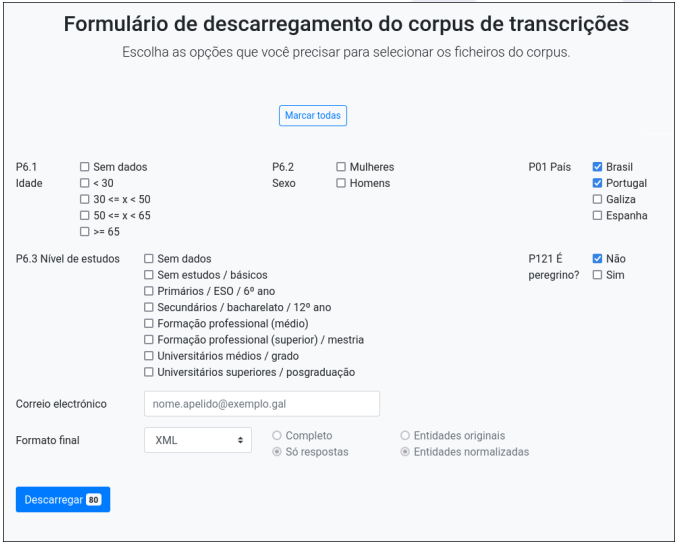
Explotación y líneas de trabajo

Explotación

1. Análisis cuantitativos y cualitativos
   - Corpus disponible para los investigadores del proyecto
   - Formatos HTML, XML y TXT
2. Entrenamiento de un NER basado en aprendizaje automático

Líneas de trabajo futuro

- Anonimización de datos personales
- Publicación bajo licencia libre
- Construcción de mapas de densidad de consumos y prácticas de visitantes



Gracias

Contacto: <susana.sotelo.docio@usc.gal>