Semantically Annotated Snapshot of the English Wikipedia

J. Atserias, H. Zaragoza, M. Ciaramita, G. Attardi
Yahoo! Research Barcelona
U. Pisa, on sabbatical at Yahoo! Research
LREC, 2008
Summary

- Introduction and Goals
  - Processing the wikipedia
  - Resulting Semanticaly Annotated Wikipedia
  - Conclusions and Future Work
Summary

- Introduction and Goals
- Processing the wikipedia
  - Resulting Semanticaly Annotated Wikipedia
  - Conclusions and Future Work
Introduction and Goals
Processing the wikipedia
Resulting Semanticaly Annotated Wikipedia
Conclusions and Future Work
Summary

- Introduction and Goals
- Processing the wikipedia
- Resulting Semanticaly Annotated Wikipedia
- Conclusions and Future Work
Pablo Ruiz Picasso (October 25, 1881 – April 8, 1973), often referred to simply as Picasso, was a Spanish painter and sculptor. His full name is Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y Picasso. One of the most recognized figures in 20th-century art, he is best known as the co-founder, along with Georges Braque, of cubism.

Biography

Pablo Picasso was born in Málaga, Spain, the first child of José Ruiz y Blasco and María Picasso y López. He was christened with the names Pablo, Diego, José, Francisco de Paula, Juan Nepomuceno, María de los Remedios, and Cipriano de la Santísima Trinidad. Picasso’s father was a painter whose specialty was the naturalistic depiction of birds and who for most of his life was also a professor of art at the School of Crafts and a curator of a local museum. The young Picasso showed a passion and a skill for drawing from an early age, according to his mother, his first word was “piz,” a shortening of lápiz, the Spanish word for pencil. It was from his father that Picasso had his first formal academic art training, such as figure drawing and painting in oil. Although Picasso attended art schools throughout his childhood, often those where his father taught, he never finished his college-level course of study at the Academy of Arts.
Processing the Wikipedia

Basic preprocessing
PoS tagging
Lemmatization
Dependency parsing
Semantic Tagging

Semantic Annotated Wikipedia
DeSR: open source statistical parser\textsuperscript{1} [Attardi et al., 2007] trained on the WSJ Penn Treebank was used to obtain syntactic dependencies, e.g. Subject, Object, Predicate, Modifier, etc. (85.85\% LAS, 86.99\% UAS in the CONLL 2007 English Multilingual shared task)

SuperSense Tagger\textsuperscript{2} [Ciaramita and Altun, 2006] open source, first-order Hidden Markov Model trained with a regularized average perceptron algorithm.

\textsuperscript{1}http://desr.sourceforge.net

\textsuperscript{2}Available at http://sourceforge.net/projects/supersensetag/
Tagsets

- **WordNet SuperSenses (WNSS)**: [Miller et al., 1993]. The accuracy of this tagger estimated by crossvalidation is about 80% F1.

- **Wall Street Journal (WSJ)**: BBN Pronoun Coreference and Entity Type Corpus, 105 categories, 87% F1.

- **WSJCONLL**: trained on BBN Pronoun Coreference and Entity Type Corpus where the WSJ labels were converted into the CONLL 2003 NER tagset using a manually created map. 91% F1.
Why different Tagsets?

WSJ

WNSS

CONLL
<p>| | | | | | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pablo</td>
<td>NNP</td>
<td>pablo</td>
<td>B-PER</td>
<td>B-noun.person</td>
<td>B-E:PERSON</td>
<td>2</td>
<td>NMOD</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Picasso</td>
<td>NNP</td>
<td>picasso</td>
<td>I-PER</td>
<td>I-noun.person</td>
<td>I-E:PERSON</td>
<td>14</td>
<td>SBJ</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>NNP</td>
<td></td>
<td></td>
<td>B-noun.time</td>
<td>B-T:DATE:DATE</td>
<td>2</td>
<td>PRN</td>
<td>B-/wiki/October_25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>October</td>
<td>CD</td>
<td>25</td>
<td></td>
<td>B-adj.all</td>
<td>I-T:DATE:DATE</td>
<td>4</td>
<td>NMOD</td>
<td>I-/wiki/October_25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>CD</td>
<td>,</td>
<td></td>
<td></td>
<td>I-T:DATE:DATE</td>
<td>4</td>
<td>P</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1881</td>
<td>NNP</td>
<td>1881</td>
<td></td>
<td></td>
<td>I-T:DATE:DATE</td>
<td>9</td>
<td>NMOD</td>
<td>B-/wiki/1881</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April</td>
<td>NNP</td>
<td>april</td>
<td></td>
<td></td>
<td>I-T:DATE:DATE</td>
<td>4</td>
<td>NMOD</td>
<td>B-/wiki/April_8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CD</td>
<td>8</td>
<td></td>
<td></td>
<td>I-T:DATE:DATE</td>
<td>9</td>
<td>NMOD</td>
<td>I-/wiki/April_8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>CD</td>
<td>1973</td>
<td></td>
<td></td>
<td>I-T:DATE:DATE</td>
<td>4</td>
<td>P</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>was</td>
<td>VBD</td>
<td>be</td>
<td></td>
<td>B-verb.stative</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
<td>a</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>JJ</td>
<td>spanish</td>
<td>B-MISC</td>
<td>B-adj.pert</td>
<td>B-E:NORP:NATIONALITY</td>
<td>18</td>
<td>NMOD</td>
<td>B-/wiki/Spain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>painter</td>
<td>NN</td>
<td>painter</td>
<td></td>
<td>B-noun.person</td>
<td>B-E:PER_DESC</td>
<td>18</td>
<td>COORD</td>
<td>I-/wiki/Painter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>CC</td>
<td>and</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sculptor</td>
<td>NN</td>
<td>sculptor</td>
<td></td>
<td>B-noun.person</td>
<td>B-E:PER_DESC</td>
<td>18</td>
<td>COORD</td>
<td>B-/wiki/Sculpture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure: Multitag Format Example
Figure: Detailed Graph, Live of Pablo Picasso
### Entity Containment Graph

**Figure:** Format of the Entity Containment Graph

<table>
<thead>
<tr>
<th>DocId:SenId</th>
<th>Named Entity</th>
<th>WSJ Tag</th>
<th>FileId.WikiId.SenId</th>
</tr>
</thead>
<tbody>
<tr>
<td>405750:0</td>
<td>Pablo Picasso</td>
<td>E:PERSON</td>
<td>wiki816.24176.0</td>
</tr>
<tr>
<td>405750:1</td>
<td>Picasso</td>
<td>E:WORK_OF_ART:PAINTING</td>
<td>wiki816.24176.1</td>
</tr>
<tr>
<td>405750:2</td>
<td>Picasso</td>
<td>E:PERSON</td>
<td>wiki816.24176.2</td>
</tr>
<tr>
<td>405750:3</td>
<td>Young Pablo Picasso</td>
<td>E:PERSON</td>
<td>wiki816.24176.3</td>
</tr>
<tr>
<td>405750:4</td>
<td>Pablo Picasso</td>
<td>E:PERSON</td>
<td>wiki816.24176.4</td>
</tr>
<tr>
<td>405750:4</td>
<td>October 25, 1881 April 8, 1973</td>
<td>T:DATE:DATE</td>
<td>wiki816.24176.4</td>
</tr>
<tr>
<td>405750:4</td>
<td>Spanish</td>
<td>E:PER_DESC</td>
<td>wiki816.24176.4</td>
</tr>
<tr>
<td>405750:4</td>
<td>painter</td>
<td>E:PER_DESC</td>
<td>wiki816.24176.4</td>
</tr>
<tr>
<td>405750:4</td>
<td>sculptor</td>
<td>E:PER_DESC</td>
<td>wiki816.24176.4</td>
</tr>
<tr>
<td>405750:5</td>
<td>One</td>
<td>N:CARDINAL</td>
<td>wiki816.24176.5</td>
</tr>
<tr>
<td>405750:5</td>
<td>20th century</td>
<td>T:DATE:DATE</td>
<td>wiki816.24176.5</td>
</tr>
<tr>
<td>405750:5</td>
<td>co-founder</td>
<td>E:PER_DESC</td>
<td>wiki816.24176.5</td>
</tr>
<tr>
<td>405750:5</td>
<td>Georges Braque</td>
<td>E:PERSON</td>
<td>wiki816.24176.5</td>
</tr>
<tr>
<td>405750:6</td>
<td>Picasso</td>
<td>E:PERSON</td>
<td>wiki816.24176.6</td>
</tr>
</tbody>
</table>
Figure: Full Entity Containment Graph
Names related to "Pablo Picasso"

- Pablo Picasso
- Henri Matisse
- Georges Braque
- Salvador Dali
- Joan Miro
- Max Ernst
- Pablo Ruiz Picasso — Picasso
- Amedeo Modigliani
- Les Demoiselles d'Avignon
- Guernica
- Dora Maar
- Juan Gris
- Claude Monet
- Jonathan Richman
- Marc Chagall
- Marcel Duchamp
- Paul Cezanne
- Auguste Rodin
- Andre Derain

Mouse over an element to see details >>
Entity Containment Graph

Timeline

In the 20th century: The Museu Picasso (Picasso Museum) in Barcelona, Spain, has one of the most extensive collections of artworks by the 20th century artist Pablo Picasso.

- 1905: Pablo Picasso’s 1905 oil painting “At the Lapin Agile” helped to make this cabaret world famous.
- 1905: “Tumblers (Mother and Son)” - 1905, “Laufende Frauen am Strand” - 1922, the breakfast in the free one “1961”
The SW1 snapshot of the Wikipedia contains 1,490,688 entries from which we extract 843,199,595 tokens in 74,924,392 sentences. Table 1 shows the number of semantics tags for each tagset and the average length in the number of tokens.

<table>
<thead>
<tr>
<th></th>
<th>#Tags</th>
<th>Average Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNSS</td>
<td>360,499,446</td>
<td>1,27</td>
</tr>
<tr>
<td>WSJ</td>
<td>189,655,435</td>
<td>1,70</td>
</tr>
<tr>
<td>WSJCONLL</td>
<td>96,905,672</td>
<td>2,01</td>
</tr>
</tbody>
</table>

**Table:** Semantic Tag figures
Conclusions

- First version of a semantically annotated snapshot of the English Wikipedia (SW1)
- Valuable resource for both the NLP and the IR community.
  - Used in [Zaragoza et al., 2007]
  - Tag visualiser\(^3\) by Bestiario\(^4\).
  - Up to you to find new uses!
  - ...

\(^3\)http://www.6pli.org/jProjects/yawibe/
\(^4\)http://www.bestiario.org/web/bestiario.php
Open issues:

- Preprocessing Wikipedia
  - Using new-cleaner-stable wikipedia dumps, maybe Wikipedia Extraction (WEX\textsuperscript{5}).
  - Which text is relevant? metatext, tables, captions?
- Processing Wikipedia
  - Adaptation: The nature of Wikipedia text (tables, lists, references) differs from training corpora. "Learning to tag and tagging to learn: A case study on Wikipedia" to appear in IEEE Intelligent Systems

\textsuperscript{5}http://download.freebase.com/wex/
The future versions, Why:

- Wikipedia is growing constantly
- Improved the processing, include new tagsets
- Multilingual (e.g. Italian, Catalan, Spanish)
SW1 at http://www.yr-bcn.es/semanticWikipedia

Thank you!

