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

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

Pablo Picasso Wikipedia Entry

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

search

Go Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

languages

- Afrikaans
- العربية
- বাংলা
- Bân-lâm-gú
- Bosanski
- Brezhoneg
- Български
- Català
- Česky
- Cymraeg
- Dansk

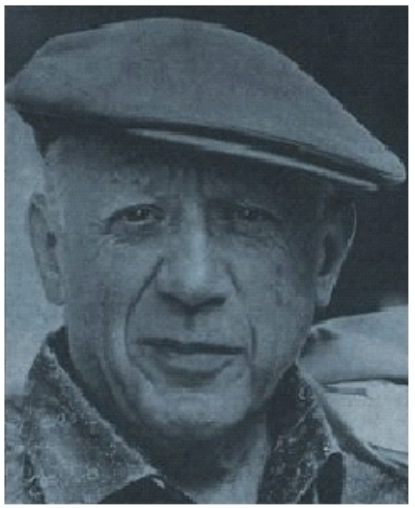
Pablo Ruiz Picasso (October 25, 1881 – April 8, 1973), often referred to simply as **Picasso**, was a Spanish painter and sculptor. His full name is **Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Clito Ruiz y Picasso**.[1] One of the most recognized figures in 20th century art, he is best known as the co-founder, along with Georges Braque, of cubism.

Contents [show]

Biography [edit]

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.[2] 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,[3] his first word was "piz," a shortening of *lápiz*, the Spanish word for pencil.[4] 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

Pablo Picasso



Picasso (January 1962)

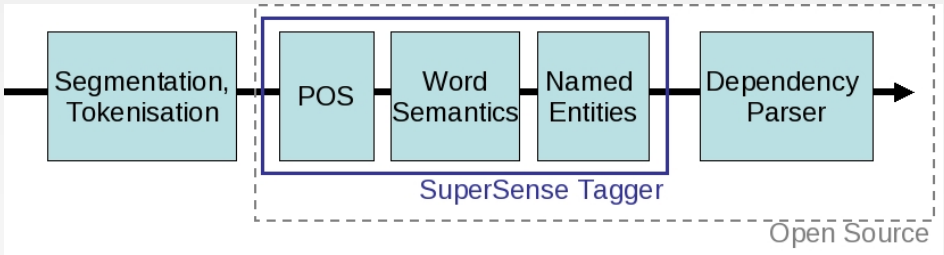
Birth name Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Martyr Patricio Clito Ruiz y Picasso

Born October 25, 1881
Málaga, Spain

Died April 8, 1973 (aged 91)
Mougins, France

Processing the Wikipedia

Basic preprocessing
PoS tagging
Lemmatization
Dependency parsing
Semantic Tagging



Semantic Annotated Wikipedia

The Dependency Parser and the Semantic Tagger

- **DeSR**: open source statistical parser[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**[2] [Ciaramita and Altun, 2006] open source, first-order Hidden Markov Model trained with a regularized average perceptron algorithm.

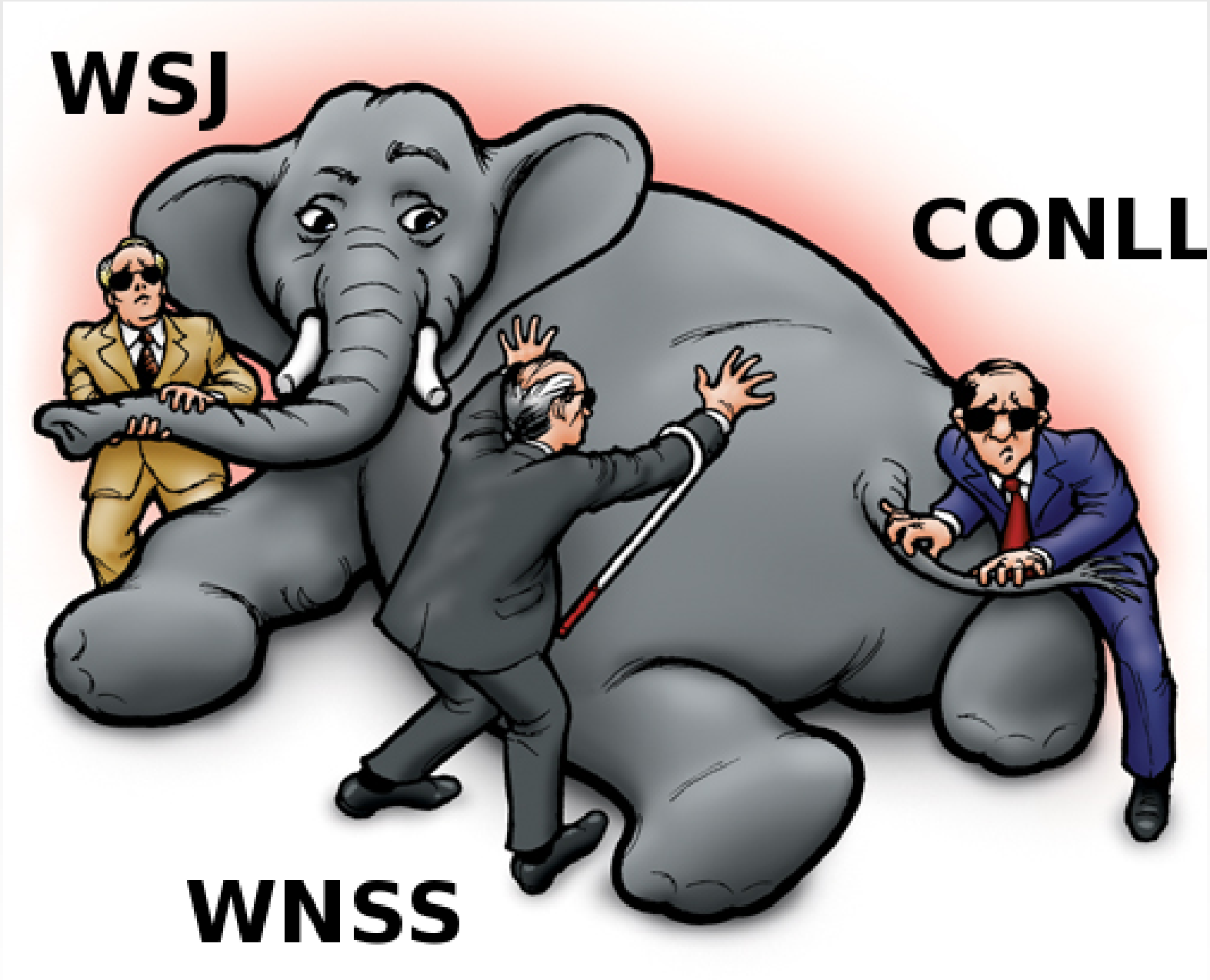
[1] http://desr.sourceforge.net

[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?



%%#DOC wiki816.24176								
%%#PAGE Pablo_Picasso								
.....								
%%#SEN 22476 wx10								
Pablo	NNP	pablo	B-PER	B-noun.person	B-E:PERSON	2	NMOD	0
Picasso	NNP	picasso	I-PER	I-noun.person	I-E:PERSON	14	SBJ	0
(	(	(	0	0	0	4	P	0
October	NNP	october	0	B-noun.time	B-T:DATE:DATE	2	PRN	B-/wiki/October_25
25	CD	25	0	B-adj.all	I-T:DATE:DATE	4	NMOD	I-/wiki/October_25
,	,	,	0	0	I-T:DATE:DATE	4	P	0
1881	CD	1881	0	0	I-T:DATE:DATE	9	NMOD	B-/wiki/1881
	NNP		0	0	I-T:DATE:DATE	9	NMOD	0
April	NNP	april	0	B-noun.time	I-T:DATE:DATE	4	NMOD	B-/wiki/April_8
8	CD	8	0	0	I-T:DATE:DATE	9	NMOD	I-/wiki/April_8
,	,	,	0	0	I-T:DATE:DATE	4	P	0
1973	CD	1973	0	0	I-T:DATE:DATE	4	NMOD	B-/wiki/1973
)	)	)	0	0	0	4	P	0
was	VBD	be	0	B-verb.stative	0	0	ROOT	0
a	DT	a	0	0	0	18	NMOD	0
Spanish	JJ	spanish	B-MISC	B-adj.pert	B-E:NORP:NATIONALITY	18	NMOD	B-/wiki/Spain
painter	NN	painter	0	B-noun.person	B-E:PER_DESC	18	COORD	I-/wiki/Painter
and	CC	and	0	0	0	14	VMOD	0
sculptor	NN	sculptor	0	B-noun.person	B-E:PER_DESC	18	COORD	B-/wiki/Sculpture
.	.	.	0	0	0	14	P	0
%%#SEN 22477 wx11								
One	CD	one	0	0	B-N:CARDINAL	13	ADV	0
of	IN	of	0	0	0	1	NMOD	0
the	DT	the	0	0	0	6	NMOD	0
most	RBS	most	0	B-adv.all	0	5	AMOD	0

Figure: Multitag Format Example

Entity Containment Graph

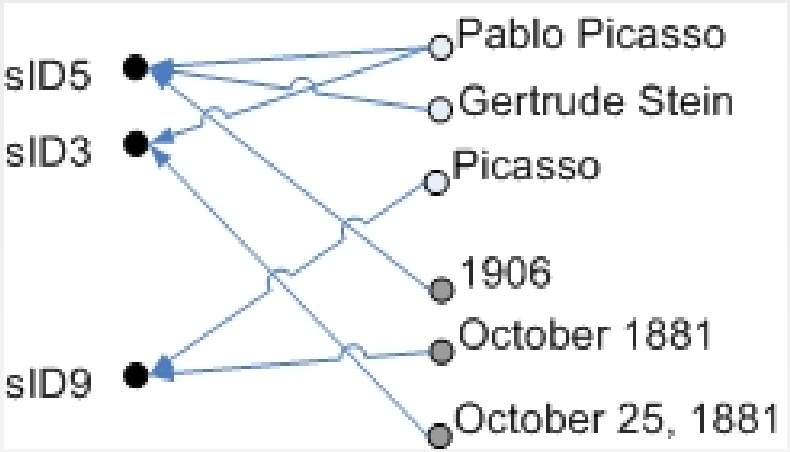


Figure: Detailed Graph, Live of Pablo Picasso

Entity Containment Graph

DocId:SenId	Named Entity	WSJ Tag	FileId.WikiId.SenId
405750:0	Pablo Picasso	E:PERSON	wiki816.24176.0
405750:1	Picasso	E:WORK_OF_ART:PAINTING	wiki816.24176.1
405750:2	Picasso	E:PERSON	wiki816.24176.2
405750:3	Young Pablo Picasso	E:PERSON	wiki816.24176.3
405750:4	Pablo Picasso	E:PERSON	wiki816.24176.4
405750:4	October 25 , 1881 April 8 , 1973	T:DATE:DATE	wiki816.24176.4
405750:4	Spanish	E:PER_DESC	wiki816.24176.4
405750:4	painter	E:PER_DESC	wiki816.24176.4
405750:4	sculptor	E:PER_DESC	wiki816.24176.4
405750:5	One	N:CARDINAL	wiki816.24176.5
405750:5	20th century	T:DATE:DATE	wiki816.24176.5
405750:5	co-founder	E:PER_DESC	wiki816.24176.5
405750:5	Georges Braque	E:PERSON	wiki816.24176.5
405750:6	Picasso	E:PERSON	wiki816.24176.6

Figure: Format of the Entity Containment Graph

Entity Containment Graph

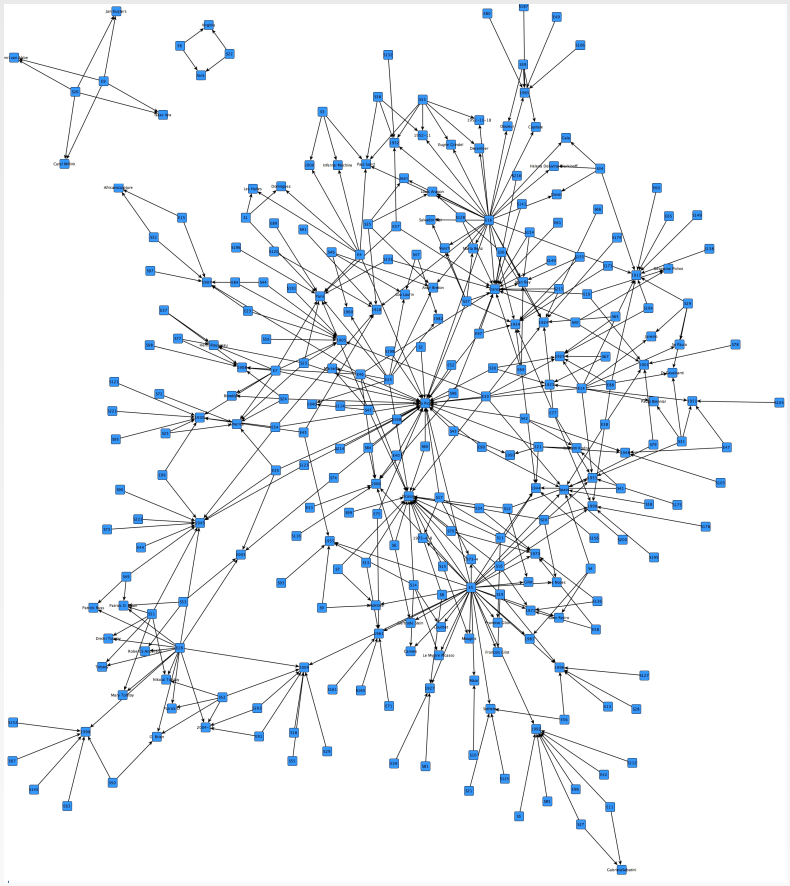


Figure: Full Entity Containment Graph

Entity Containment Graph

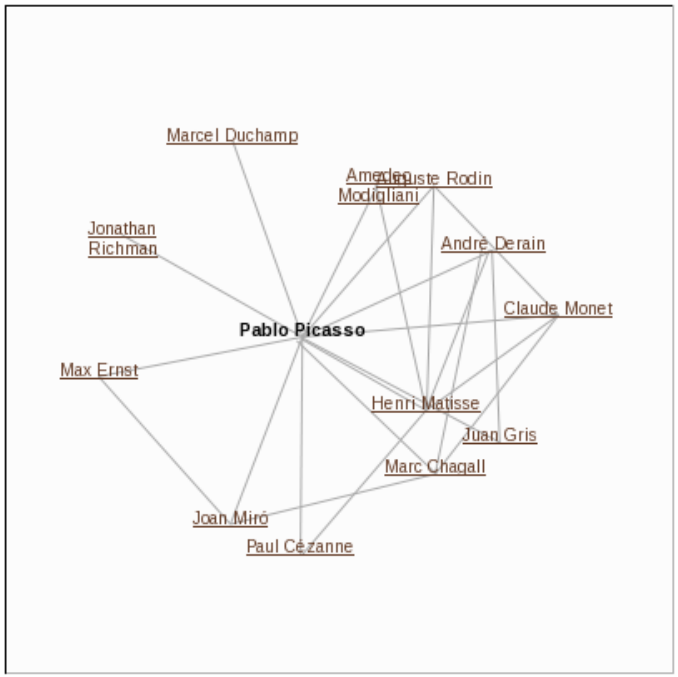
Pablo Picasso

Overview | Names | Places | Concepts | Events | Photos | Queries | News | Answers | Sites | All related

Names related to "Pablo Picasso"

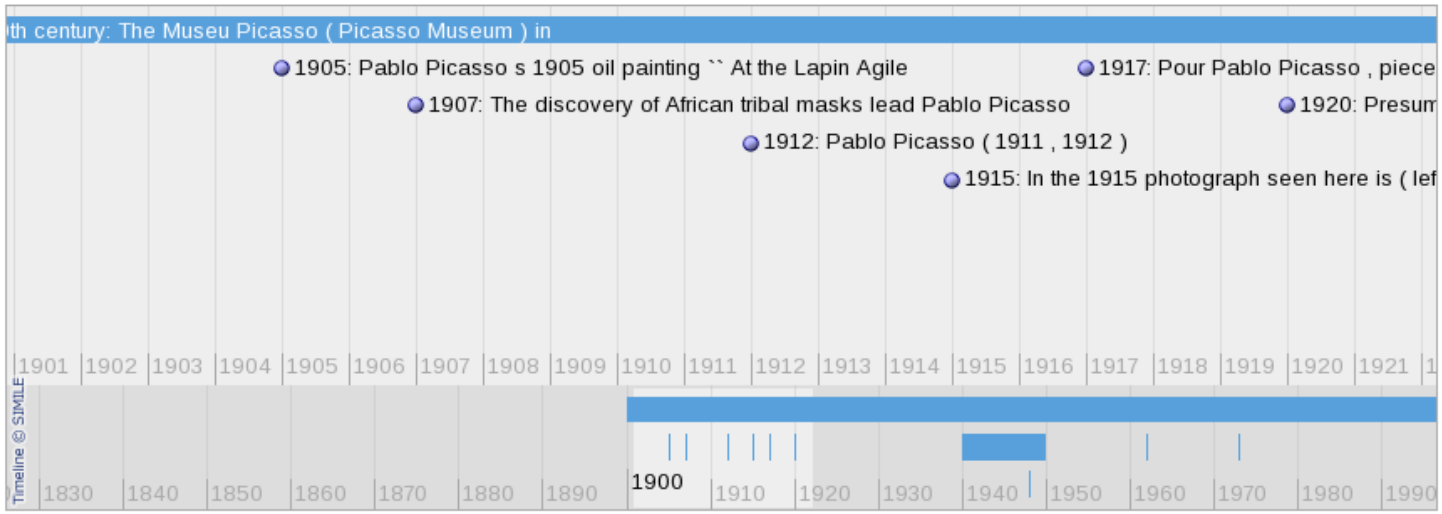
- Pablo Picasso
- Henri Matisse
- Georges Braque
- Salvador Dali
- Joan Miró
- Max Ernst
- Pablo Ruiz Picasso — Picasso
- Amedeo Modigliani
- Les Demoiselles d'Avignon
- Guernica
- Dora Maar
- Juan Gris
- Claude Monet
- Jonathan Richman
- Marc Chagall
- Marcel Duchamp
- Paul Cézanne
- Auguste Rodin
-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

Mouse over an element to see details >>



Entity Containment Graph

Timeline



Events in the timeline

the 20th century

- (From Museu Picasso) "The Museu Picasso (Picasso Museum) in Barcelona , Spain , has one of the most extensive collections of artworks by **the 20th century** artist Pablo Picasso . "
- (From PabloDraw) "PabloDraw was named after **the 20th century** artist Pablo Picasso . "
- (From Farley 's yard) "Farley 's became meeting place for many of the most influential artists of **the 20th century** including Pablo Picasso , Joan Miró , Max Ernst , Man Ray , Paul Éluard and Henry Moore . "

1905

- (From Lapin Agile) "Pablo Picasso 's **1905** oil painting `` At the Lapin Agile '' helped to make this cabaret world famous . "
- (From Staatsgalerie Stuttgart) "Pablo Picasso 's `` Tumblers ( Mother and Son ) `` - **1905** , `` Laufende Frauen am Strand '' - 1922 , `` the breakfast in the free one '' 1961 "

SW1 Snapshot

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.

	#Tags	Average Length
WNSS	360,499,446	1,27
WSJ	189,655,435	1,70
WSJCONLL	96,905,672	2,01

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

Future Work

Open issues:

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

[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!


Attardi, G., Dell'Orletta, F., Simi, M., Chanev, A., and Ciaramita, M. (2007).
Multilingual dependency parsing and domain adaptation using desr.
In *Proceedings the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.


Ciaramita, M. and Altun, Y. (2006).
Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger.
In *Proceedings of the EMNLP*.


Miller, G., Leacock, C., Tengi, R., and R.Bunker (1993).
A semantic concordance.
In San Mateo, C. M. K.-m. P., editor, *Proceedings of the ARPA Human Language Technology Workshop.*, Princeton, NJ.


Sang, E. F. T. K. and Muelder, F. D. (2003).
Introduction to the CoNLL-2003 shared task:
Language-independent named entity recognition.
In *CoNLL 2003 Shared Task*, pages 142–147.


Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., and Attardi, G. (2007).
Ranking very many typed entities on wikipedia.
In *CIKM*, pages 1015–1018.