Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 169-173 recibido 30-04-10 revisado 18-05-10 aceptado 20-05-10

Automatic Annotation of the Catalan Wikipedia: Exploring the Semantic Space via multiple NERC systems

Anotación automática de la Wikipedia catalana: Exploración del espacio semántico mediante múltiples sistemas de reconocimiento de entidades

Jordi Atserias
Barcelona Media – Centre d'Innovació
Av. Diagonal, 177, planta 9, 08018 Barcelona
jordi.atserias@barcelonamedia.org

Judith Domingo
Barcelona Media – Centre d'Innovació
Av. Diagonal, 177, planta 9, 08018 Barcelona
judith.domingo@barcelonamedia.org

Carlos Rodríguez-Penagos
Barcelona Media – Centre d'Innovació
Av. Diagonal, 177, planta 9, 08018 Barcelona
carlos.rodriguez@barcelonamedia.org

Teresa Suñol
Barcelona Media – Centre d'Innovació
Av. Diagonal, 177, planta 9, 08018 Barcelona
teresa.sunol@barcelonamedia.org

Resumen: Este artículo presenta WikiNer, una versión de la Wikipedia catalana procesada mediante diferentes herramientas de PLN (etiquetadores de POS, NERC, parsers de dependencias). El artículo se centra en el análisis de las diferentes anotaciones de NERC realizadas con 3 etiquetadores: JNET, Yamcha y SST. A pesar de que el texto de la Wikipedia (especialmente las tablas, listas y referencias) difiere significativamente, en sus propiedades distribucionales, del corpus empleado para entrenar los etiquetadores, se han obtenido resultados satisfactorios que apuntan a la posibilidad de una rápida disponibilidad de un recurso de gran masa textual anotada con un grado de fiabilidad suficiente tanto para algunas de las tareas de investigación como para ciertas aplicaciones: Q&A, enriquecimiento de ontologías y búsqueda semántica.
Palabras clave: Wikipedia, Reconocimiento de Entidades, Recuperación de Información.

Abstract: This paper presents WikiNer, a snapshot of the Catalan Wikipedia processed with different NLP tools (POS tagger, NERC, dependency parsers). The article focuses on the analysis of different NERC annotations using 3 taggers: JNET, YamCha and SST. Although Wikipedia text (specially in tables, lists, references) differs significantly in distributional properties from the corpora used to train the taggers, we believe that results of automatically annotating the semantic space of the Catalan Wikipedia point to the quick availability of a resource containing massive text annotated with a degree of reliability that is enough for some research tasks as well as for applications, such as simple Q&A, ontology enrichment and semantic search.
Keywords: Wikipedia, Named Entity Recognition, Information Retrieval.

1 Introduction

Unsupervised annotation of linguistic resources always involves a certain trade-off in reliability when compared against resource-intensive manually-crafted resources, although this compromise can be acceptable when the intended analyses require massive amounts of data. Minority languages frequently face this dilemma, since funding for high quality research and development resources is limited. We present here our experiences in automatically creating a semantically annotated corpus for Catalan (WikiNer), by processing its Wikipedia with an ensemble of Named-Entity Recognition and Classification systems, in order to obtain and explore a massive semantic space many times the order of magnitude of the *de-facto*, multi-level standard corpus in that language: AnCora. We evaluate the interannotator agreement of the taggers based

ISSN 1135-5948

on Conditional Random Fields (CRF), Support Vector Machine (SVM) an Hidden Markov Models (HMM) with an average perceptron as well as the overall reliability against a manually-reviewed sample. In recent years Wikipedia has become a valuable resource for both the Natural Language Processing (NLP) and the Information Retrieval (IR) communities. Although NLP technology for processing Wikipedia has been applied for extracting semantics (Weld, 2007), not all researchers and developers have the computational resources to process such a volumes of data. The aim of this work is also to provide easy access to syntactic and semantic annotations for researchers of both NLP and IR communities by building a reference corpus to homogenize experiments and make results comparable. WikiNer is licensed under the GNU Free Documentation License.

2 *Data description and annotation methods*

WikiNer is a snapshot of the Catalan Wikipedia enriched with syntactic and semantic automatic annotations. WikiNer contains 155,276 Wikipedia entries with 1,848.246 sentences and about 44 million tokens. In order to build this resource, we started XML-ing a snapshot of the Catalan Wikipedia dated 23/01/2009 using the tools distributed by the University of Amsterdam[1].

After basic sentence and word segmentation, the text was annotated with Part-Of-Speech (POS) and lemma using TreeTagger (Schmidt, 1994) with a Parole tagset variation developed at GLICOM[2], using a model trained on the 6 million-word LEXESP corpus (Carmona, Cervell et. al., 1998) and a lexicon of approximately 970k entries with lemma-POS pairs. This output was fed to three different Named Entity recognizers (NERC) systems that used tokens, lemmas and POS information as features to identify and classify candidate Named-Entities:

1- **SuperSense Tagger (SST)** (http://sourceforge.net/projects/supersensetag/) (Ciaramita, 2006): The basic tagger is a first-order Hidden Markov Model trained with a regularized average perceptron algorithm. Features used are lowercased word, PoS, shape (regular expression simplification), bi/tri-grams of characters from the beginning and ending of each word.

2- **JNET** (http://www.julielab.de/): is a Named-Entity Recognizer using mallet's implementation of Conditional Random Fields. We used lemmas and POS, as well as orthotypographical features, in a bigram window before and after each tagged token.

3- **Yamcha** (http://chasen.org/~taku/software/YamCha/) (Kudo, 2003): uses Support Vector Machines, and was used originally as a chunker, but has shown good results in other NLP tasks such as NER. The features used are lemmas and POS, and a trigram window.

All classifiers where trained with the Catalan version of Ancora Corpus (Ancora), using its annotated Named Entities (with some exceptions, namely long document titles). The training corpus has a size of 484,779 tokens, once multiwords (among them NEs) were split from their original representation as single lexical items, and a tagset, different from the AnCora's original tagset, was mapped into the text. WikiNer was also processed (PoS, lemma, most frequent wordnet sense) using Freeling (Atserias, 2006) and then syntactic dependencies extracted using the DeSR parser (Attardi, 2007) that reported 86.86 LAS and 91.91 UAS for Catalan in the CoNLL 2007 Shared Task on Dependency.

Table 1 shows an example of the resulting resource (some fields like dependencies and wikipedia links are omitted for readability). We also preserved the entry title and the link information of the Wikipedia articles which will allow researchers to explore the structure of the Wikipedia (e.g. category structure or the hyperlinks) as well as to explore the multilingual component due to the parallel nature of the Wikipedia (that is, the Wikipedia entries can be aligned with the entries of other annotated Wikipedia snapshots, for instance English (Atserias, 2008)).

[1] http://ilps.science.uva.nl/WikiXML/

[2] http://www.glicom.upf.edu/

Forma	FreeLing	POS	TT POS	Lema TT	Lema Free Ling	SST	YamCha	JNET	GS
Luci	NP00000	N5	N	luci	luci_vinici	B-PER	B-PER	B-PER	B-PER
Vinici	NP00000	N4	N	vinici	luci_vinici	I-PER	I-PER	I-PER	I-PER
(	Fpa	Fpa	F	(	(	O	O	O	O
Lucius	NP00000	N5	N	Lucius	lucius_vinicius	B-MISC	B-PER	B-LOC	B-LOC
Vinicius	NP00000	N4	N	Vinicius	lucius_vinicius	I-MISC	I-PER	I-LOC	I-LOC
o	CC	CC	C	o	o	O	O	O	O
Vinucius	NP00000	N4	N	Vinucius	vinucius	B-LOC	B-PER	B-LOC	B-LOC
)	Fpt	Fpt	F	)	)	O	O	O	O
fou	VSIS3S0	VD	V	ser	ser	O	O	O	O
un	DI0MS0	E6	E	un	un	O	O	O	O
magistrat	NCMS000	N5	N	magistrat	magistrat	O	O	O	O
romà	AQ0MS0	JQ	J	Romà	romà	O	O	O	O
.	Fp	Fp	F	.	.	O	O	O	O

Table 1: Example of a corpus fragment

3 Evaluations

Table 2 shows the different number of entities tagged by each of the NERC systems. We observe that all NERC tag a similar number of persons and locations although regarding MISC category SST seems to tag more MISC than the others, while JNET seems to be more conservative and tags less sub-specified. Moreover, JNET also tags as ORG significantly less entities than SST and YAMCHA.

	SST	JNET	YAMCHA
LOC	1.187.835	1,004,426	1,301,302
MISC	791.945	208,530	402,075
ORG	829.483	383,552	810,066
PER	1.349.188	1,321,006	1,421,245

Table 2: Number of Named Entities tagged

Tested against a manually-created Gold Standard of a random sample of 19 articles, an accuracy of 0.88 was measured for the Part-of-Speech tagging. With regard to Named-Entity performance, we measured accuracies using both full labels indicating *Begin, Inside, Outside* positions of the word with regard to NEs, as well as a less fine-grained measure involving just the category labels (ORG, PER, LOC, MISC), as shown in Table 3.

NER accuracy	positional (IOB)	non-positional (IO)
SuperSense	0.840	0.846
Yamcha	0.843	0.848
JNET	0.875	0.879

Table 3: Individual Tagger accuracies.

Table 4 shows the pairwise results. The inter system agreement for the 3 algorithms was measured using weighted kappa (Cohen, 1968) of 0.741, an average interannotator agreement of 0.894 and an average overall disagreement of 0.105.

	SST	YAMCHA	JNET
SST	-	0.686	0.796
YAMCHA	0.686	-	0.742
JNET	0.796	0.742	-

Table 4: Pairwise results.

In order to better understand the results we observed the different kind of errors made and we determine that each NER tagger has its own weaknesses: SST tags wrongly some expressions as name entities because of a wrong sentence tokenization. JNET usually tags isolated words in full capital letters as an entity even if they do not represent true entities. Usually this kind of error occurs when it finds headlines.

Although the methods are not much independent statistically (as shown in table 4) and they use similar features, the qualitative analysis of the errors shows they have different strengths and weaknesses. Thus we decide to explore whether the combination of the annotation would improve the results. In order to test that, we used a weighted voting scheme to test the hypothesis that an ensemble system would outperform each separate tagger's accuracy. A simple voting strategy (as there are only three methods, ties are resolved using JNET's prediction) shows a similar performance (accuracy 87.15 positional and 87.53 non-positional) to JNET's performance. Since the experiment was not conclusive, we plan to use more test data and to add per-category performance weights to each classifier (via a confusion matrix) so as to favor the one that showed more consistency in each category. All of the NERC taggers used similar features to make comparison of the performance of the different ML methods possible; being all classifiers different in nature (CRF, SVM, HMM) we also hope that WikiNer can also help to explore co-learning techniques (Dasgupta, 2001) which can be extremely useful specially for languages with few resources.

Doing a more fine-grained analysis, the structures that get less accurated results are long entities such as book titles, magazines and award names. These structures usually contain sub-structures with name entities that makes it difficult to recognize the complete structure and as a consequence the entity is not properly tagged.

4 *Applications and conclusions*

While preparing the WikiNer resource, we came across some interesting issues, both from a technical and theoretical points of view: as evidenced in Table 1, location and person entity counts across the 3 taggers seem to be point to less ambiguous and regular contexts for the appearance of these kinds of entities, while organizations and miscellaneous present greater divergences. The small amount of preliminary test data, unfortunately, does not allow us to confirm these first impressions, but construction of resources such as WikiNer can allow investigation of questions such as these.

From a technical perspective, extracting the text to be processed from the XML version not only implies which xml tags contain the text to be processed, but also involves segmentation decisions (that is which xml tag should split words or sentences). In the current version we have tried to process all the tags that contain text, including metatext, tables, captions, etc.

The nature of Wikipedia text (especially in the tables, lists, references) differs significantly in distributional properties from the corpora used to train the taggers and the parser. Therefore the quality of these NLP processors is considerably lower than what results from the evaluation in-domain. For example, articles about antiquity contain person names that include their birthplace (e.g. *Diògenes de Sinope*), a common practice then, that can confuse the classifiers. As a matter of fact we hope that this resource will help to understand the domain adaptation problem for this kind of semantic processing.

Lately, there has also been an increasing interest in building new browsing interfaces and IR applications that make an extensible use of Named Entities automatically detected e.g. *Correlator*[3] on wikipedia or *Silobreaker*[4] on news. These tools provide novel ways to explore the Semantic Space represented by the persons, organizations, locations and other diverse entities of interest mentioned in texts, and discover relations and facts about them emerging from a sufficiently big volume of documents.

We believe that the results shown here for automatically annotating the semantic space of the Catalan Wikipedia point to the quick availability of a resource containing massive text annotated with a degree of reliability enough for some research tasks (though not all) as well as applications, such as simple Q&A, ontology enrichment, semantic search, etc. Minority languages with less developed resources, in particular, can benefit from having access to coarse-grained data on a scale needed for using techniques such as Semantic Vector Spaces, for which data availability might outweight a few percentage points of lost precision. We hope the WikiNer presented here will be a useful resource for NLP researchers,

[3] http://correlator.sandbox.yahoo.com

[4] http://www.silobreaker.com/

as well as for linguists in general. We have shown the soundness of the automatic annotation methodology that can be extended to other minority languages that have at least a minimal amount of manually annotated resources to train the tools, but lack massive textual corpus.

Bibliografía

AnCora http://clic.ub.edu/ancora/

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

Atserias, J., B. Casas, E. Comelles, M. González, Ll.. Padró and M. Padró. 2006. *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library.* Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genova, Italy. May, 2006.

Atserias, J., H. Zaragoza, M. Ciaramita and G. Attardi. 2008. *Semantically Annotated Snapshot of the English Wikipedia* Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), ELRA, Marrakesh, Morroco,2008

Carmona, J, Cervell, S., Màrquez, L., Martí, M. A., et al. 1998 *An environment for r morphosyntactic processing of unrestricted spanish text*. In Proceedings of LREC'98 (pp. 915–922).

Ciaramita, M. and Y. Altun. 2006. *Broad-coverage sense disambiguation and information extraction with super-sense sequence tagger.* In Proceedings of the EMNLP'06.

Cohen, J. 1968. *Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit.* Psychol Bull 1968; 70: 213-20.

Dasgupta, S., M. Littman, D. McAllester. 2001. *PAC generalization bounds for co-training*. In Proceedings of Neural Information Processing Systems, 2001

Kudo, T. and Y. Matsumoto. 2003. *Fast Methods for Kernel-Based Text Analysis,* ACL 2003

Schmidt, H. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Proceedings of International Conference on New Methods in Language Processing.

Weld, D. and F. Wu. 2007. *Autonomously Semantifying Wikipedia,* the Sixteenth ACM Conference on Information and Knowledge Management (CIKM-07)*, Lisbon, Portugal*, 2007.