Developing New Linguistic Resources and Tools for the Galician Language

Rodrigo Agerri [*], **Xavier Gómez Guinovart** [**], **German Rigau** [*], **Miguel Anxo Solla Portela** [**]
[*] IXA NLP Group, University of the Basque Country UPV/EHU | [**] TALG Research Group, University of Vigo
rodrigo.agerri@ehu.eus, xgg@uvigo.es, german.rigau@ehu.eus, miguelsolla@uvigo.es

Abstract

In this paper we describe the work towards developing new resources and Natural Language Processing (NLP) tools for the Galician language. First, a new corpus, manually revised, for POS tagging and lemmatization is described. Second, we present a new manually annotated corpus for Named Entity tagging for Galician. Third, we train and develop new NLP tools for Galician, including the first publicly available Galician statistical modules for lemmatization and Named Entity Recognition, and new modules for POS tagging, Wikification and Named Entity Disambiguation. Finally, we also present two new Web demo applications to easily test the new set of tools online.

Keywords: Galician, Less-resourced languages, Language Resources, Linguistic Tools

1. Introduction

We present new developments on linguistic resources and Natural Language Processing tools for the Galician language. There are previous works addressing the creation of resources to allow the automatic processing of the Galician language, including the Galician WordNet (Gómez Guinovart and Solla Portela, 2017), the Galician SemCor (Solla Portela and Gómez Guinovart, 2017), works on terminology (Solla Portela and Gómez Guinovart, 2015) or annotation of large corpora (Gómez Guinovart and López Fernández, 2009). Furthermore, there are also publicly available Natural Language Processing tools for the Galician language, notably Freeling (Padró and Stanilovsky, 2012) and Linguakit (Gamallo and Garcia, 2017). However, it is remarkable the lack of linguistic processors that are available for other less-resourced languages, such as statistical tools for lemmatization or Named Entity Recognition.

In this work we aim to contribute to the development of NLP resources for Galician by providing a new manually revised corpus for POS tagging and lemmatization, and a new manually annotated corpus for Named Entity Recognition. This would allow us to train previous unavailable statistical tools for the processing of Galician texts. As a result of our effort, seven new linguistic processors for Galician are presented: a rule-based tokenizer and statistical tools for POS tagging, lemmatization, Named Entity Recognition, Named Entity Disambiguation, Wikification and graph-based Word Sense Disambiguation. In particular, this paper presents the first open source statistical lemmatizer and Named Entity tagger for the Galician language.

2. Corpora and Lexical Resources

In this paper the main work on corpora and lexical resources was undertaken in order to create new resources to train a statistical POS tagger and lemmatizer, and a new Named Entity Recognition and Classification tagger. The basis for the corpora required is the CTG Galician Technical Corpus[1] (TALG Research Group, 2016). The CTG corpus contains around 18 million words from various domains:

- GALEX Galician legal texts (∼3M words).
- XIGA Galician texts on computing and telecommunications (∼3M words).
- AUGA Galician texts on ecology and environmental sciences (∼3M words).
- ACHEGA Galician economy texts (∼2M words).
- SOGAL Galician texts on sociology and social sciences (∼3M words).
- MEDIGAL Galician texts on medicine (∼4M words).

In order to provide the required resources to train a POS tagger and lemmatizer for Galician, we manually revised a subset of the CTG corpus containing 2,852,472 tokens, 105,986 sentences and 938 texts (from scientific-technical communications, academic works and news from ecology and environmental sciences, and legal texts). Apart from the annotated corpora, two dictionaries to perform dictionary-based lemmatization and multiword recognition were built from several sources:

- Dicionario da Real Academia Galega[2].
- Vocabulario ortográfico da lingua galega (VOLGa)[3].
- Hunspell Spellchecker for Galician[4].
- Galician dictionary distributed by Apertium[5].
- Galician dictionary distributed by Freeling[6].

[1] http://sli.uvigo.gal/CTG/

[2] http://academia.gal/dicionario/

[3] http://www.realacademiagalega.org/recursos-volg

[4] https://github.com/meixome/hunspell-gl

[5] http://sourceforge.net/projects/apertium/

[6] http://nlp.lsi.upc.edu/freeling/

- Other textual and lexical resources developed by the TALG research group.

The collection of part-of-speech tags used in the CTG subset corpus and dictionaries are based on the CTAG tagset developed by the TALG Group (Gómez Guinovart and López Fernández, 2009). The subset used for training the POS tagger contains around 200 morphological tags and 23K different lemmas. For training and testing, we created three splits following the Penn Treebank as model. Thus, we took the first 950K tokens for training, the following 150K for development and the next 150K for test.

Finally, a new corpus for Named Entity Recognition in Galician (TALG Research Group, 2018) was manually annotated on a subsection of the CTG corpus consisting of 202,334 tokens in 8,137 sentences (from the news and ecology and environmental sciences domains). The CoNLL guidelines for annotation were followed (Tjong Kim Sang and De Meulder, 2003). This resulted in an inventory of 4 named entity classes distributed as follows: 1,293 persons (PER), 3,183 organizations (ORG), 2,616 locations (LOC) and 1,375 miscellaneous entities (MISC). From this corpus 162K tokens are used for training and 41K for test.

3. NLP Tools

In order to develop new linguistic processors using the resources described in the previous section, we decided to try the IXA pipes tools[7] (Agerri et al., 2014). The aim of IXA pipes is to provide a modular set of ready to use Natural Language Processing (NLP) tools. Apart from being easy to train and deploy, they are also a good fit for our current aim of providing new tools for Galician because every module but the tokenizer is machine learning based. In fact, IXA pipes tries to use the same approach across NLP tasks in order to create robust processors both across domains and languages. This strategy has proven to be very successful for Named Entity Recognition and Classification (NER) (Agerri and Rigau, 2016) and Opinion Target Extraction (OTE) (San Vicente et al., 2015) benchmarks, both in out-of-domain and in-domain evaluations.

3.1. Semi-supervised approach

IXA pipes learns supervised models based on the Perceptron algorithm (Collins, 2002). To avoid duplication of efforts, IXA pipes uses the Apache OpenNLP project implementation[8] customized with its own features. By design, IXA pipes tools aim to establish a simple and shallow feature set, avoiding any linguistic motivated features, with the objective of removing any reliance on costly extra gold annotations apart from the target task (POS, lemmas, NER) and/or cascading errors if automatic language processors are used. IXA pipes modules consist of: (i) Local, shallow features based mostly on orthographic, word shape and n-gram features plus their context; and (ii) three types of simple clustering features, based on unigram matching. Specifically, IXA pipes implements, on top of the local features, a combination of three word representation features: (i) Brown (Brown et al., 1992) clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark (Clark, 2003) clusters and, (iii) Word2vec (Mikolov et al., 2013) clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm. The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then we add the class as a feature. The Brown clusters only apply to the token related features, which are duplicated. The word representation features are *combined* and *stacked* from features induced over different data sources. For this work the new Galician POS tagger, lemmatizer and NER tagger are based on this design.

Furthermore, IXA pipes are extended with third-party tools for those type of annotations not developed within its toolchain. Most notably, it includes integration of wikification and Named Entity Disambiguation (NED) via DBPedia Spotlight (Mendes et al., 2011) as well as graph-based Word Sense Disambiguation (Agirre et al., 2014).

Summarizing, the new set of NLP tools for Galician implemented within IXA pipes consists of the following modules:

- ixa-pipe-tok: A rule-based tokenizer and sentence segmenter.
- ixa-pipe-pos: A statistical lemmatizer and POS tagger (ixa-pipe-pos); ixa-pipe-pos is complemented by the dictionaries described in the previous section for dictionary-based lemmatization and multiword detection. For efficiency, these dictionaries are deployed as finite state automata based on Morfologik[9].
- ixa-pipe-nerc: A state of the art NER tagger.
- ixa-pipe-wikify: Wikification tool based on DBpedia Spotlight.
- ixa-pipe-ned: A NED tool based on DBpedia Spotlight. The NED uses the entities spotted by ixa-pipe-nerc as input to perform the disambiguation.
- ukb-naf: UKB graph-based Word Sense Disambiguation.

3.2. Named Entity Disambiguation and Wikification

Galician language already had a previous version of the DBpedia Spotlight[10] that was implemented together with the official server of the Galician DBpedia (Solla Portela and Gómez Guinovart, 2016), but it used the Lucene version (Mendes et al., 2011). A new, better performing, generative model (Daiber et al., 2013) for DBpedia Spotlight has been created by configuring the Galician language in model-quickstarter[11] in order to handle Wikification and Named Entity Disambiguation with IXA pipes, using the modules already available for other languages.

[7] http://ixa2.si.ehu.es/ixa-pipes
[8] http://opennlp.apache.org/
[9] https://github.com/morfologik/
[10] http://sli.uvigo.gal/dbpedia/spotlight/
[11] https://github.com/dbpedia-spotlight/model-quickstarter

3.3. Web demo

The full new set NLP tools provided by the IXA pipes for the Galician language can be easily tested through the *Lingaliza* Web application[12] developed by the TALG Research Group of the University of Vigo. Furthermore, *DContado* [13], the user-oriented version of Lingaliza, is designed to be used in the field of Digital Humanities for Research in Humanities and Social Sciences under the auspices of the European infrastructure CLARIN[14] (Bel et al., 2016).

4. Experimental Results

In this section we will report on results for the POS tagger and Named Entity Recognition taggers trained on the CTG corpora as described in section 2. The aim of this section is to give us a first idea of the performance of these tools for Galician. In order to train our systems with the semi-supervised approach described in section 3.1., first the three types of word representations needed to be induced from large unlabelled data. The first obvious candidate was to use the Galician Wikipedia. However, as shown by previous experiments using this approach (Agerri and Rigau, 2016), it is convenient for best performance to provide word representation features induced from different data sources. In order to do that we compiled a *large corpus* from various domains by crawling data from the following Web data sources:

- Asociación para a Defensa Ecolóxica de Galiza (ADEGA) (http://adega.gal).
- The Galician Political Party, *Bloque Nacionalista Galego* (BNG) (http://www.bng.gal/).
- Galician newspaper *Praza Pública* (http://praza.gal/).
- Galician weekly and newspaper *Sermos Galiza* (http://www.sermosgaliza.gal/).
- Galician government official website: *Xunta de Galiza* (http://www.xunta.gal).

Roughly speaking, the text used from the Galician wikipedia to train the clusters contained 31M words, whereas the *large corpus* compiled contained around 20M words (see (Agerri and Rigau, 2016) for details on the clusters training process). For POS tagging we choose our best feature configurations on the development set whereas for NERC we did our development via 5-fold cross validation. The NER results reported in Table 1 confirm the behaviour of *ixa-pipe-nerc* previously observed in (Agerri and Rigau, 2016) for other languages. The local features are improved substantially by the clustering features. At the same time, the combination of those features provide the best results.

Table 2 reports on the results of POS tagging. As it is known from previous approaches using distributional semantic features for POS tagging, the gains obtained from using word representations for this task as not as large as those obtained for NER. At least if we look at word accuracy (WA) only. However, by looking at the sentence accuracy (SA) and unknown accuracy (UA) scores it can be seen the clear improvements in performance from the local features to the models using clustering features on top of the local ones. The UA scores in particular are quite interesting as it show that, despite the negligible improvements in terms of WA, the models containing clustering features are much more robust to tag unseen words.

Features	P	R	F1
Local	79.95	81.93	80.93
Local + BL2000	82.34	83.65	82.99
Local + CW300	82.32	83.21	82.76
Local + W2VW100	81.83	84.22	83.01
Local + BL2000 + CW300	83.85	84.54	84.19

Table 1: NER results. BL2000: Brown 2000 classes from Large corpus; CW300: Clark 300 classes from Wikipedia, and W2V100: Word2vec 100 classes from Wikipedia.

Features	SA	UA	WA
Local	72.08	80.09	98.31
Local + BW1000	70.83	81.14	98.24
Local + CW300	74.41	83.94	98.50
Local + W2VW100	71.93	81.45	98.30
Local + CW300 + CL400	75.22	85.35	98.54

Table 2: POS results. SA: Sentence Accuracy; UA: unknown accuracy; WA: word accuracy.

5. Related Work

Previous work on NLP for Galician have been centred around Freeling (Padró and Stanilovsky, 2012) and Linguakit (Gamallo and Garcia, 2017). Freeling provides wide support for several languages, however, it does not yet include machine learning based lemmatization or Named Entity Recognition and Classification for Galician. Current support includes tokenization, POS tagging, and rule-based lemmatization and detection (no classification) of named entities.

With respect to Linguakit, it provides applications for Galician such as a verb conjugator, a POS tagger, dependency parser, Named Entity tagger, sentiment analyzer, a keyword extractor and a summarization module. Lemmatization, Named Entity tagging and parsing are performed by language independent rule-based modules.

6. Conclusions and future work

In this paper we presented a new set of linguistic resources and NLP tools for the Galician language. In particular, we would like to highlight the contribution of two new manually revised and annotated corpora for Galician POS tagging, lemmatization and Named Entity Recognition (NER). Furthermore, this paper has presented a number of novel statistical NLP tools for the Galician language, including a lemmatizer, a NER tagger, and a wikification and NED module. Additionally, the reported results indicate that the

[12] http://sli.uvigo.gal/lingaliza/
[13] http://sli.uvigo.gal/dcontado/
[14] http://clarin.eu

performance obtained is similar to performance of those tasks for other languages which is promising start (Agerri and Rigau, 2016), although more experiments remain to be done.

With respect to future work, we believe that establishing a clear benchmarking between the different tools for Galician NLP would be very interesting. In this way, and apart from providing an objective comparison, it would be possible to learn what are the strengths and weaknesses of each toolchain currently available for Galician. Additionally, we would like to keep extending the coverage of NLP tools for Galician. For example, IXA pipes provides modules for statistical chunking and constituent parsing, which have been already deployed for Basque, English and Spanish.

The code for every tool presented in this paper is already available for public use through the IXA pipes website. In the same way, every trained model[15] and associated linguistic resources are freely available[16]. Furthermore, these tools and models can be tested using the two Web implementations developed by the TALG Research Group of the University of Vigo.

7. Acknowledgements

This research has been carried out thanks to the project TUNER (TIN2015-65308-C5-1-R) supported by the Ministry of Economy and Competitiveness of the Spanish Government and the European Fund for Regional Development (MINECO/FEDER).

8. Bibliographical References

Agerri, R. and Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*.

Agirre, E., López de Lacalle, O., and Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Comput. Linguist.*, 40(1):57–84, March.

Bel, N., González-Blanco, E., and Iruskieta, M. (2016). CLARIN centro-k-español. *Procesamiento del Lenguaje Natural*, 57:151–154.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66. Association for Computational Linguistics.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8.

[15] http://ixa2.si.ehu.es/ixa-pipes/

[16] http://sli.uvigo.gal/download/SLI_Galician_Corpora/

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Gamallo, P. and Garcia, M. (2017). Linguakit: a multilingual tool for linguistic analysis and information extraction. *Linguamática*, 9(1):19–28.

Gómez Guinovart, X. and López Fernández, S. (2009). Anotación morfosintáctica do Corpus Técnico do Galego. *Linguamática*, 1(1):61–71.

Gómez Guinovart, X. and Solla Portela, M. A. (2017). Building the galician wordnet: methods and applications. *Language Resources and Evaluation*.

Mendes, P. N., Jakob, M., Garcia-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2473–2479, Istanbul, Turkey, May. European Language Resources Association (ELRA).

San Vicente, I., Saralegi, X., and Agerri, R. (2015). Elixa: A modular and flexible absa platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752. Association for Computational Linguistics.

Solla Portela, M. A. and Gómez Guinovart, X. (2015). Termonet: Construcción de terminologías a partir de WordNet y corpus especializados. *Procesamiento del Lenguaje Natural*, 55:165–168.

Solla Portela, M. A. and Gómez Guinovart, X. (2016). Dbpedia del gallego: recursos y aplicaciones en procesamiento del lenguaje. *Procesamiento del Lenguaje Natural*, 57:139–142.

Solla Portela, M. A. and Gómez Guinovart, X. (2017). Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con wordnet 3.0. *Procesamiento del Lenguaje Natural*, 59:137–140.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147.

9. Language Resource References

TALG Research Group. (2016). *CTG Corpus [Galician Technical Corpus]*. University of Vigo, SLI resources, 1.0, ISLRN 437-045-879-366-6.

TALG Research Group. (2018). *SLI NERC Galician Gold CoNLL*. University of Vigo, SLI resources, 1.0, ISLRN 435-026-256-395-4.