Inflection Generation for Spanish Verbs using Supervised Learning

Cristina Barros
Department of Software
and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
cbarros@dlsi.ua.es

Dimitra Gkatzia
School of Computing
Edinburgh Napier University
Edinburgh, EH10 5DT, UK
d.gkatzia@napier.ac.uk

Elena Lloret
Department of Software
and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
elloret@dlsi.ua.es

Abstract

We present a novel supervised approach to inflection generation for verbs in Spanish. Our system takes as input the verb's lemma form and the desired features such as person, number, tense, and is able to predict the appropriate grammatical conjugation. Even though our approach learns from fewer examples comparing to previous work, it is able to deal with all the Spanish moods (indicative, subjunctive and imperative) in contrast to previous work which only focuses on indicative and subjunctive moods. We show that in an intrinsic evaluation, our system achieves 99% accuracy, outperforming (although not significantly) two competitive state-of-art systems. The successful results obtained clearly indicate that our approach could be integrated into wider approaches related to text generation in Spanish.

1 Introduction

Existing Natural Language Generation (NLG) approaches are usually applied to non morphological rich languages, such as English, where the morphological inflection of the word during the generation process can be addressed using simple handwritten rules or existing libraries such as SimpleNLG (Gatt and Reiter, 2009). In contrast, when it comes to morphological rich languages, such as Spanish, the use of rules can lead to incorrect inflection of a word, thus generating ungrammatical or meaningless texts. Our ultimate goal is to implement a morphological inflection approach for Spanish sentences within an NLG system based on the use of lexicons. However, lexicons lack some verbs' information, specifically, regarding grammatical moods (i.e., the grammatical features of verbs used for denoting modality - statement of facts, desires, commands, etc.). To create lexicons for all the verb inflections and moods would be a very time-consuming and costly task, so in this context the use of machine learning approaches can benefit the inflection of unseen verb forms. Based on this, the research challenge we tackle is defined as follows: given a Spanish verb in its base form (i.e., its lemma), we want to automatically generate all the inflections for that verb. This is very useful for tasks involving natural language generation (e.g., text generation, machine translation), since the generated texts would sound more natural and grammatically correct.

Our contributions to the field are as follows: we present a novel and efficient method for tackling the challenge of inflection generation for Spanish verbs using an ensemble of algorithms; we provide a high-quality dataset which includes inflection rules of Spanish verbs for all the grammatical moods (i.e. indicative, subjunctive and imperative, being this last one do not tackled by the current approaches); our models are trained with fewer resources than the state-of-art methods; and finally, our method outperforms the state-of-the-art methods achieving a 2% higher accuracy.

The rest of the paper is shaped as follows: In the next section (Section 2) we refer to the related work on inflection generation. In Section 3, we describe the overall methodology and the dataset used to train our model. In Section 4, we present a comparison to the state-of-art inflection generation approaches and in Section 5, we discuss the results. Finally, in Section 6, directions for future work are discussed.

2 Related Work

Morphological inflection has been addressed from different perspectives within the area of Compu-

tational Linguistics, commonly for morphological rich languages, such as German, Spanish, Finnish or Arabic, as well as less morphological rich languages such as English.

Previous work has used supervised or semi-supervised learning (Durrett and DeNero, 2013; Ahlberg et al., 2014; Nicolai et al., 2015; Faruqui et al., 2016) to learn from large datasets of morphological rules on word forms in order to apply them to inflect the desired words. Other approaches have relied on linguistic information, such as morphemes and phonology (Cotterell et al., 2016); morphosyntactic disambiguation rules (Suárez et al., 2005); and, graphical models (Dreyer and Eisner, 2009).

Recently, the morphological inflection has been also addressed at SIGMORPHON 2016 Shared Task (Cotterell et al., 2016) where, given a lemma with its part-of-speech, a target inflected form had to be generated (task 1). This task was addressed through several approaches, including align and transduce (Alegria and Etxeberria, 2016; Nicolai et al., 2016; Liu and Mao, 2016); recurrent neural networks (Kann and Schütze, 2016; Aharoni et al., 2016; Östling, 2016); and, linguistic-inspired heuristics approaches (Taji et al., 2016; Sorokin, 2016). Overall, recurrent neural networks approaches performed better, being (Kann and Schütze, 2016) the best performing system in the shared task, obtaining around 98%.

Furthermore, the work described here differs from existing statistical surface realisation methods which use phrase-based learning (e.g., (Konstas and Lapata, 2012)) since they do not usually include morphological inflection. In this respect, our work is more similar to (Dušek and Jurčíček, 2013), where the inflected word forms are learnt through multi-class logistic regression by predicting edit scripts. The aforementioned data-driven methods achieve high accuracy in predicting the appropriate inflection by learning from huge datasets. For example, Durret and DeNero (2013) use 11400 amount of data (i.e. the total number of instances or rules used to predict the inflections of a verb). In contrast, we use almost half to train our system (4556 instances), and we achieve comparable or better results for Spanish. Finally, the work presented here relies on ensembles of classifiers which has been proved successful for content selection in data-to-text systems (Gkatzia et al., 2014).

3 Methodology

In order to perform the inflection task, we first created a dataset to be used for training machine learning algorithms to inflect verbs in Spanish. As part of this submission we will make our dataset freely available[1]. Then, we trained a model capable of predicting the appropriate inflection of a verb automatically, given a verb base form. Next, each of the stages of the proposed approach are described in more detail.

3.1 Dataset Creation

For the purposes of this research, we created a parallel dataset of Spanish base forms and their corresponding inflected form. The Spanish verbs can be divided into regular and irregular verbs, where all the regular verbs share the same inflection patterns whereas, the irregular ones do not and can completely vary from one verb to another, as it is shown in Figure 1.

Regular			Irregular		
Base form		Inf. Form	Base form		Inf. Form
partir		parta	ir		vaya
añadir	=	añada	decir	≠	diga
compartir		comparta	argüir		arguya

Figure 1: Differences between regular and irregular verbs in Spanish, for the first singular person of the present tense and in the subjunctive mood.

Therefore, we constructed a dataset, containing the necessary examples of inflection for all the tenses in the Spanish language, by consulting the *Real Academia Española*[2] and the *Enciclopedia Libre Universal en Español*[3]. We further considered that a verb can be divided in three parts: (1) *ending*, (2) *ending stem*, and (3) *penSyl*. An example is shown in Figure 2. This information will be later used as features within the dataset. In Spanish, the verbs can be classified depending on their *ending*, specifically, the verbs ended by "-ar", "-er" and "-ir" belong to the first, second an third conjugation, respectively. Moreover, for the feature *penSyl*, the previous syllable of the ending, formed by the whole syllable, or its dominant vowel is extracted. Finally, the *ending stem* is the closest consonant to the ending.

[1] Our dataset for the Spanish verbs inflection is available here: https://github.com/cbarrosua/infDataset

[2] http://www.rae.es/diccionario-panhispanico-de-dudas/apendices/modelos-de-conjugacion-verbal

[3] http://enciclopedia.us.es/index.php/Categoría:Verbos

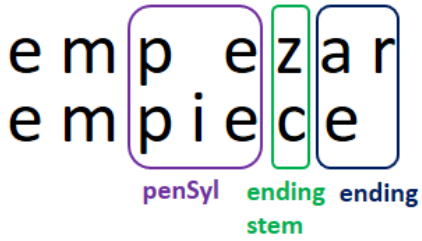

Figure 2: Division of the Spanish verb *to begin* and its inflection for the first singular person of the present tense and in the subjunctive mood.

Besides the previous features obtained from the verb, other features, such as *suff1*, *suff2* or *stemC1* were included because in Spanish some verbs have several variations of an inflection for the same tense, person and number. Therefore, our dataset is finally composed of the following features: (1) *ending*, (2) *ending stem*, (3) *penSyl*, (4) *person*, (5) *number*, (6) *tense*, (7) *mood*, (8) *suff1*, (9) *suff2*, (10) *stemC1*, (11) *stemC2*, (12) *stemC3*. In particular, *suff1* and *suff2* are the inflection predicted for the suffix of the verb form; and *stemC1*, *stemC2* and *stemC3*, refer to the inflection predicted for the penSyl of the verb form. An example of an entry of the dataset is shown in Table 1. Overall, there are 4556 possible inflections. An example of a verb and several of its inflections is shown in Table 2.

3.2 Obtaining the Model and Reconstructing the Verb

As mentioned earlier, our learning task is formed as follows: given a set of 7 features, select the inflection which is most appropriate for the verb. The set of 7 features are as follows: (1) *ending*, (2) *ending stem*, (3) *penSyl*, (4) *person*, (5) *number*, (6) *tense*, (7) *mood*. Using these features, we trained a group of individual models for each of the features described in Section 3.1, which represents a potential inflection value. We used the WEKA (Frank et al., 2016) implementation of the Random Forest algorithm to train the models for the *stemC3* and *stemC2* features, and the Random Tree algorithm to train the models for the *suff1*, *suff2* and *stemC1* features.

Once the models were trained, we predicted all the possible inflections given a verb in its base form, i.e., all the tenses for each mood in Spanish. For accomplishing this task, we first analysed the base form to extract the necessary features for the inflection. In this manner, the base form was divided into syllables, taking the penultimate one to obtain the *penSyl* feature. Since all verbs in Spanish always end with "-ar", "-er" and "-ir", as described in the previous section, we split the last syllable into the *ending* and *ending stem* features. Then, for each model we predicted its potential inflection using these extracted features combined with the ones related to the verb tense, i.e., the number, person, etc. Finally, the predicted inflections were employed to replace the features previously identified in the base form, leading to the reconstruction of the base form into the desired inflection, as it can be seen in Figure 3.

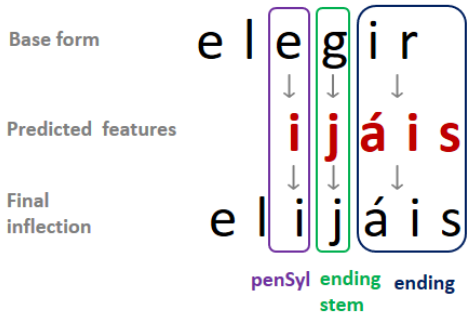

Figure 3: Reconstruction of the verb *elegir* (to choose) with the features predicted by the models.

4 Experiments

We compared our system (RandFT) with two very competitive baselines described below by measuring the accuracy of their output for Spanish verb inflections. The baselines are as follows:

- **Durret13**: This system automatically extracts the orthographic transformation rules of the morphology from labeled examples, and then learns which of those transformations to apply in different contexts by using a semi-Markov conditional random field (CRF) model.
- **Ahlberg14**: This system uses a semi-supervised approach to generalise inflection paradigms from inflection tables by using a finite-state construction.

We reproduced the experiments presented in Durrett and DeNero (2013) and in Ahlberg et al. (2014). In order to compare our system with both baselines, we employed the test set of examples (200 different verbs) which was made available

verb pattern	ending	endingstem	penSyl	person	number	tense	mood	suff1	suff2	stemC1	stemC2	stemC3
	ar	ANY	ANY	1	0	1	1	ara	ase	ANY	ANY	ANY
amar	ar	ANY	ANY	2	0	1	1	aras	ases	ANY	ANY	ANY
	ar	ANY	ANY	3	0	1	1	ara	ase	ANY	ANY	ANY
	er	ANY	yac	1	0	0	0	o	ANY	yazc	yazg	yag
yacer	er	ANY	yac	2	0	0	0	es	ANY	yac	ANY	ANY
	er	ANY	yac	3	0	0	0	e	ANY	yac	ANY	ANY

Table 1: Example of the 1st, 2nd and 3rd singular person of the subjunctive past tense of "amar" (*to love*); and the 1st, 2nd and 3rd singular person of present tense in indicative mood of "yacer" (*to lie*). We assigned the term *ANY* to indicate that the value of a feature does not need to change during the inflection with respect to its value in the base form.

Verb: regar (*to water*)	
Features	**Inflection**
ar, g, e, 1P, Sing, Pres., Ind	riego
ar, g, e, 2P, Sing, Pres., Sub	riegues
ar, g, e, 2P, Plural, Cond., Ind	regaríais
ar, g, e, 3P, Sing, Past I., Sub	regara, regase

Table 2: Example of some possible inflections for the verb "regar" (*to water*) (Pres. = present; Cond. = conditional; Past I. = imperfect past; Ind = Indicative; Sub = subjunctive).

Approach	**Correctly predicted verb tables**	**Correctly predicted verb forms**
RandFT	**99**%	**99.98**%
Durret13	97%	99.76%
Ahlberg14	96%	99.52%

Table 3: Accuracy of predicting inflection of verb tables and individual verb forms given only the base form, evaluated with an unseen test set of 200 verbs. For the imperative mood, our system achieves 100% accuracy, however the baselines do not predict the imperative form.

by Durrett and DeNero (2013), since this test set included verbs with both uncommon and regular forms. This test set does not included any entry that appeared in the training data. For the experiments, we generated all the verb inflections for the 200 base forms. Furthermore, the aforementioned baselines do not predict all the grammatical moods that exist in the Spanish language (both baselines are only able to predict the indicative and subjunctive mood, but not the imperative one, which is not easy, especially for irregular forms). Therefore, we used an additional test-set to evaluate this grammatical mood. We created the additional test-set by employing information from the Freeling's lexicon for the imperative forms of these 200 verbs (Padró and Stanilovsky, 2012).

5 Results and Discussion

The results obtained, together with the results of Durrett and DeNero (2013) and Ahlberg et al. (2014), are shown in Table 3, where we compared the inflection of the same verb tenses as Durret and Ahlberg using the test set described in the previous section. Our group of classifiers (RandFT), trained with our generalised dataset for Spanish, obtained a higher overall accuracy (but not significantly) regarding the state-of-the-art baselines systems.

In addition, our model can correctly perform the inflection of the imperative mood, which was not included in the baseline systems. This grammatical mood, which forms commands or requests, contains unique imperative forms among the irregular Spanish verbs, as shown in Table 4. For this experiment, our system achieves 100% accuracy when evaluated on the additional test set. Furthermore, our model contributes to the improvement of naturalness and expressivity of NLG (Barros et al., 2017).

Base form–*Inflected form*
contar–*cuenta*; **errar**–*yerra*; **haber**–*he*; **hacer**–*haz*; **oler**–*huele*; **ir**–*ve*; **oír**–*oye*; **decir**–*di*

Table 4: Variability of inflection in the imperative mood for the 2nd person singular of the present.

Error Analysis: Although our system obtains almost 100% accuracy, it fails on the inflection of the participles of extremely rare irregular verbs (e.g., **verb**: ejabrir → **generated**: ejabrido → **correct**: ejabierto). These errors could be corrected by adding specific rules for these cases.

6 Conclusion and Future Work

This paper presented a robust light-weight supervised approach to obtain the inflected forms of any Spanish verb for any of its moods (indicative, subjunctive and imperative). This approach uses an ensemble of supervised learning algorithms to learn how the verbs are composed in order to obtain the inflection of a verb given its base form. Our method obtained accuracy close to 100%, outperforming existing state-of-the-art approaches. In addition, our method is able to further predict the inflection of the imperative mood, which was not tackled by previous work. In future, we plan to test our inflection approach for other languages, as well as other types of words (not only verbs). Furthermore, we also plan to compare this approach with the ones obtaining the best results (i.e. the ones employing recurrent neural networks) in the reinflection task of the SIGMORPHON 2016 Shared Task. Our short-term goal would be to integrate it within a surface realisation method, which will allow us to inflect whole sentences in different ways and tenses, thus improving the generation capabilities of current NLG systems.

Acknowledgments

This research work has been partially funded by the Generalitat Valenciana through the projects "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0" (PROMETEOII/2014/001); and partially funded by the Spanish Government through projects TIN2015-65100-R, TIN2015-65136-C2-2-R, as well as by the project "Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales (ASAP)" funded by Ayudas Fundación BBVA a equipos de investigación científica.

References

Roee Aharoni, Yoav Goldberg, and Yonatan Belinkov. 2016. Improving sequence to sequence learning for morphological inflection generation: The biu-mit systems for the sigmorphon 2016 shared task for morphological reinflection. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 41–48.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pages 569–578.

Iñaki Alegria and Izaskun Etxeberria. 2016. Ehu at the sigmorphon 2016 shared task. a simple proposal: Grapheme-to-phoneme for inflection. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 27–30.

Cristina Barros, Dimitra Gkatzia, and Elena Lloret. 2017. Improving the naturalness and expressivity of language generation for spanish. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The sigmorphon 2016 shared task—morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*. Association for Computational Linguistics.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Association for Computational Linguistics, pages 101–110.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. pages 1185–1195.

Ondřej Dušek and Filip Jurčíček. 2013. Robust multilingual statistical morphological generation models. In *51st Annual Meeting of the Association for Computational Linguistics: Proceedings of the Student Research Workshop*. Association of Computational Linguistics, pages 158–164.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 634–643.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 4 edition.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*.

Dimitra Gkatzia, Helen Hastie, and Oliver Lemon. 2014. Comparing multi-label classification with

reinforcement learning for summarisation of time-series data. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Katharina Kann and Hinrich Schütze. 2016. Med: The lmu system for the sigmorphon 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 62–70.

Ioannis Konstas and Mirella Lapata. 2012. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. pages 369–378.

Ling Liu and Lingshuang Jack Mao. 2016. Morphological reinflection with conditional random fields and unsupervised features. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 36–40.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 922–931.

Garrett Nicolai, Bradley Hauer, Adam St. Arnaud, and Grzegorz Kondrak. 2016. Morphological reinflection via discriminative string transduction. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 31–35.

Robert Östling. 2016. Morphological reinflection with convolutional neural networks. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 23–26.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Alexey Sorokin. 2016. Using longest common subsequence and character models to predict word forms. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 54–61.

Octavio Santana Suárez, José Rafael Pérez Aguiar, Luis Javier Losada García, and Francisco Javier Carreras Riudavets. 2005. Spanish morphosyntactic disambiguator. In *Proceedings of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*. pages 201–204.

Dima Taji, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. The columbia university - new york university abu dhabi sigmorphon 2016 morphological reinflection shared task submission. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, pages 71–75.