MorphoGen: Full Inflection Generation Using Recurrent Neural Networks

Octavia-Maria Şulea[1,2,3], Steve Young[4], Liviu P. Dinu[1,2]

[1] University of Bucharest, Faculty of Mathematics and Computer Science, Romania
[2]Human Language Technologies Research Center, University of Bucharest
[3]GumGum, Santa Monica, California, USA
[4]Pronto.AI, San Francisco, California, USA
mary.octavia@gmail.com, steve.young@pronto.ai, liviu.p.dinu@gmail.com

Abstract. Sub-word level alternations during inflection (apophonies) are an common linguistic phenomenon present in morphologically-rich languages, like Romanian. Inflection learning, or predicting the inflection class of a partially regular or fully irregular verb or noun in such a language has been a widely studied task in NLP, but generative models are limited to capturing the most common ending patterns and apophonies. In this paper, we show how to train a character-level Recurrent Neural Network language model to be able to accurately generate the full inflection of verbs in Romanian, Finish, and Spanish and model stem-level phonological alternations triggered by inflection in an unsupervised way. We also introduce a method to evaluate the accuracy of the generated inflections.

1 Introduction

Predicting the form a concept will take once it is placed in context (e.g. the selling point of a house, given its features) is one of the major goals of Statistical Machine Learning. In Natural Language Processing, this goal can surface into many different tasks, depending at which level of the language the predictive models are applied. Unlike English, where content words (nouns, verbs, adjectives) rarely shift form in context, languages with rich inflectional morphology, where the base form of a word (stem) can go through many phonological transformations (apophonies) when contextual markers denoting gender, person, number, case, definitiness or tense are added (affixation), require more special attention before higher level NLP tasks, like Machine Translation or Question Answering, can be succesfully carried out. While great progress in text classification, information extraction, or language understanding tasks has been made through language modelling with deep neural networks and transfer learning [3], [4], these models still suffer when the background information needed for the task is out of vocabulary or rarely occuring and require fine-tuning on labeled datasets [16].

Table 1: Alternations in the verbal domain for Romanian and Spanish

	Regular		**Partially Irregular**		**Irregular**	
	Romanian	Spanish	Romanian	Spanish	Romaian	Spanish
	a colora	tensar	a purta	pensar	a fi	ser
Tag	(to color)	(to tauten)	(to wear)	(to think)	(to be)	(to be)
1st sg.	color-ez	tens-o	p***o***rt-	p***ie***ns-o	sunt	soy
2nd sg.	color-ezi	tens-as	p***o***r***ț***-i	p***ie***ns-as	ești	eres
3rd sg.	color-ează	tens-a	p***oa***rt-ă	p***ie***ns-a	este	es
1st pl.	color-ăm	tens-amos	p***u***rt-ăm	p***e***ns-amos	suntem	somos
2nd pl.	color-ați	tens-áis	p***u***rt-ați	p***e***ns-áis	sunteți	sois
3rd pl.	color-ează	tens-an	p***oa***rt-ă	p***ie***ns-an	sunt	son

Within Computational Morphology, much research has been carried out in order to deal with this high variability at the stem but also affix level for morphologically rich languages, beyond stemming algorithms. The majority of the work until recently has been focused on on inflection learning or inflectional class prediction using supervised models: [10], [11], [9], [12], [2], [8]. On the other had, the task of generating all inflected forms of a word from its base (uninflected) form in a fully unsupervised manner, given no information of the existing paradigms of a language, has not been studied so far to our kowledge. Our interest was sparked by the advent of neural generative models being applied to the related task of morphological reinflection ([13], [1], [18]) and availability of morphological datasets coming from Wikitionary [12] and the SIGMORPHON shared tasks [7].

In this paper, we investigate the extent to which the attention-based recurrent neural language model implemented in [17], which has become a popular plug-and-play text generation tool outside of the academic community [1], can be used to generate the paradigm of a word given only its uninflected form and being trained in a fully unsupervised fashion. To this end we propose a new task, that of *unsupervised inflection generation*, and show that the textgenrnn architecture can reach state-of-the-art scores when trained in this novel setting, requiring no pre-training on large corpora and no fine-tuning or supervision.

2 Datasets

For our experiments, we used a subset of the wikitionary corpus [12] and a subset of the corpus introduced in [6] and used in [11] and we focused on verbal inflection generation for Finnish, Spanish, and Romanian. We chose these languages because they are known to have rich inflectional morphology, but display it in different ways: Romanian and Spanish words go through the process of apophony, or stem alternation, during affixation, whereas Finnish is an agglutinative language meaning that words and affixes never shift. The three

[1] airweirdness.com has used it extensively

languages also pertain to different language families: Romanian and Spanish are both Latin-based, whereas Finnish is a Fino-Ugric language, highly unrelated to the other two.

While the Spanish and Finnish datasets contained no direct (labeled) information about conjugational class, for Romanian we were able to attain the labeled dataset introduced in [11] which associates the infinitive of a verb (uninflected form) with a number from 0 to 29 representing the conjugational class label which they identified based on various linguistic works ([5], [15]) and which was succesfully predicted by their proposed model using only character ngrams of the infinitive. This label is supposed to encode the pattern of conjugational endings the verb receives as well as the pattern of alternations the stem goes through during conjugation. In our experiments on the Romanian dataset, we tested whether this conjugational class would lead to better inflection generation if the model was conditioned on it when trained. Interestingly, our results show that conditioning the model on the conjugational class decreases performance slightly. Table 2 shows the size of the three train and test datasets. As you can see, the wikitionary datasets are considerably smaller.

Table 2: Dataset statistics

Stage	Romanian	Spanish	Finnish
Train	25,914	3,855	7,049
Test	2,366	200	200

3 MorphoGen Architecture

We adapt the character-level language model `textgenrnn` developed in [17, 14], whose architecture is reviewed here for completeness.

Input sequences to the model are strings of up to T characters. Each character in an input sequence is first translated into a 100-D embedding-vector. These are then fed through two bi-directional 128-unit LSTM layers. Next, the outputs of the embedding and both LSTM layers are concatenated and fed into an attention layer which weights the most important temporal features and averages them together. Note that this skip-connects the embedding and first LSTM layer to the attention layer, which helps alleviate vanishing gradients. Finally, the output of the attention layer is routed through a fully-connected MLP layer with output dimension equal to the number of possible characters.

4 Generation Experiments

The first set of experiments investigated the generation of full inflection given the infinitive. Specifically, we wanted to see if the model could generate all 6 forms of

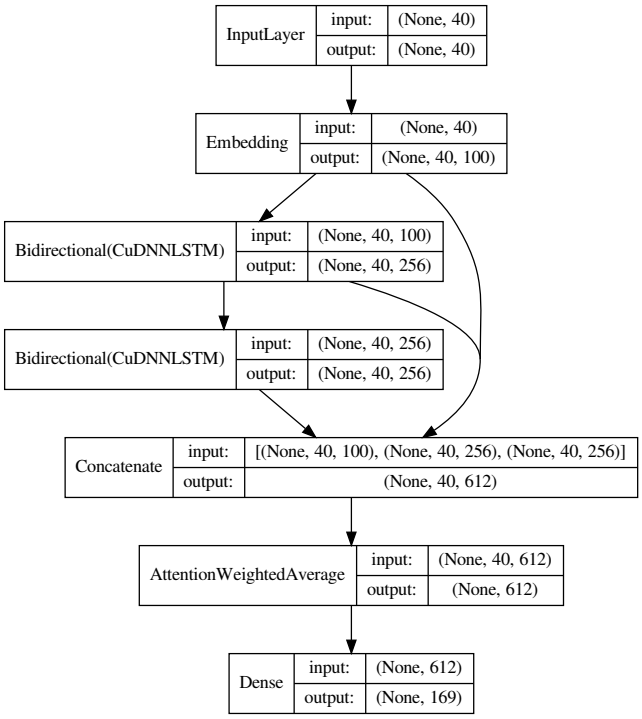

Fig. 1: Structure of the model. We use a max-length $T = 40$, a 100-D embedding matrix, 2 128-unit LSTMs, and an attention weighting with 169 outputs.

a verb in the present tense in one go, given only the uninflected (dictionary) form. To this end, the training datasets were arranged such that each row contained the infinitive of a verb followed by all the 6 inflected forms resulted from combining person (i.e. 1st, 2nd, 3rd) and number (i.e. singular, plural) in the indicative present tense.

Each form in the training set was separated by a comma as in example. The test sets were formatted the same. For Spanish and Finnish, we coupled the training and development sets together and used the test set for generation. For the Romanian set, since we also had access to the conjugational class for each infinitive, we split the 7k examples in train and test sets, making sure that they both maintained the conjugational class distribution.

During the generation phase, the trained MorphoGen model is given the uninflected form from the test set as the value to the *prefix* parameter in the *generate* function of the textgenrnn model. If the forms generated by the model

match exactly with the test set forms, then the corresponding entry in the test set is counted as having been correctly generated. Interestingly, we noticed poorer generation results when we omitted the comma character after the infinitive form in the prefix parameter. Since the model is character based, it must've learned that the comma character is a separator. We report the results when the comma is part of the prefix.

The experiments were carried out with different training parameters. We trained the three language models for 14 and 28 epochs to see if generation performance improves with longer training. We also tested the influence of the number of tokens to consider before predicting the next one by changing *max_len* from 40 characters (default value) to 70 for Romanian and to 60 for Spanish. These decisions were made based on the distribution of lengths in the training corpora 2, 3, 4.

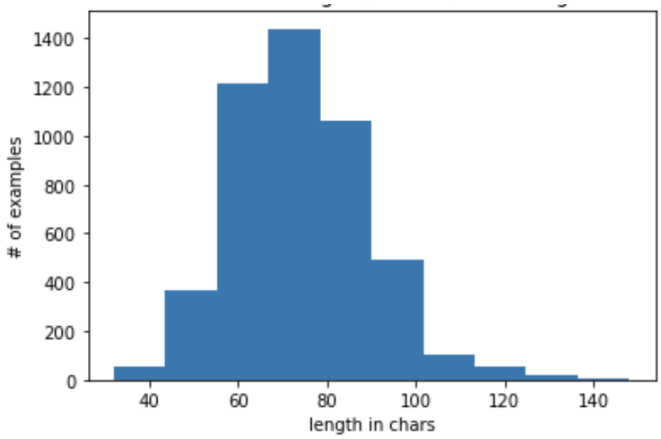

Fig. 2: Distribution of character lengths for the Romanian training corpus

5 Results

For Finish and Spanish, we see that our training scheme of the textgenrnn model reaches performance close to the state-of-the-art [1] only after 14 epochs. We note that this is with a much simpler model requiring less data and no supervision compared to previous models. For Romanian, we report the accuracy of generating the full sequence of inflected forms for regular and partially irregular verbs in the indicative present tense both when the model is (MorphoGenCond) and is not conditioned (MorphoGen) on the conjugational class. As we can see, conditioning on the class decreases performance slightly from 84.86% to 83.26%,

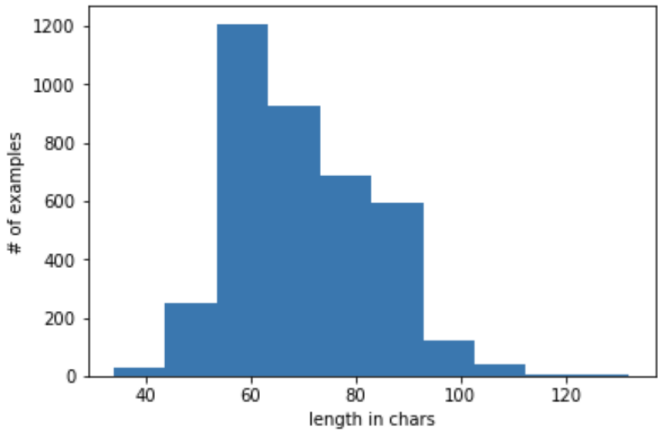

Fig. 3: Distribution of character lengths for the Spanish training corpus

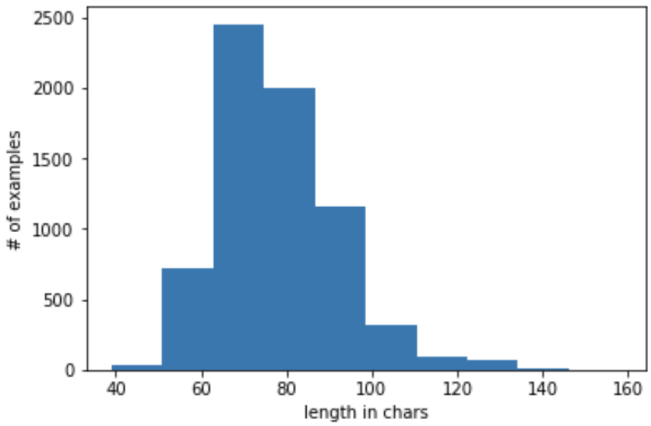

Fig. 4: Distribution of character lengths for the Finnish training corpus

but in either case our generation performance still beats state-of-the-art for this language [18]. More epochs also seem to decrease accuracy of the conditional model.

We also include the results from previous work for reference, although these models are either fully supervised or semi-supervised and use quite a lot of extensive pretraining. In contrast, our non conditional models use no supervision such as tags related to the context in which each generated form needs to ap-

pear (person, number). Our conditioning is also done on a higher level concept (conjugational class) than the inflectional context used in privious work and we show it is not necessary as it leads to slightly lower generation performance. We also saw that increasing the *max_length* parameter value from 40 to 70 leads the Romanian model to reach the same generation accuracy in half the time (only 14 epochs instead of 28). For Spanish, we saw 4% improvements in generation accuracy for Spanish when the max_length parameter is set to 60.

Table 3: Results for verb full inflection generation

Model	Supervision	Epochs	Language	max_len	Accuracy %
MorphoGen	Unsupervised	14	Romanian	40	74.68
MorphoGen	Unsupervised	28	Romanian	40	**84.86**
MorphoGen	Unsupervised	14	Romanian	70	84.65
MorphoGenCond	Conditional	14	Romanian	40	84.10
MorphoGenCond	Conditional	28	Romanian	40	83.26
[18]	Semi-Supervised	-	Romanian	-	78.6
MorphoGen	Unsupervised	14	Finnish	40	95.50
[1]	Seq2Seq	-	Finnish	-	98.07
MorphoGen	Unsupervised	14	Spanish	40	92.00
MorphoGen	Unsupervised	14	Spanish	60	96.00
[1]	Seq2Seq	-	Spanish	-	99.81

6 Conclusions

In this paper, we've shown how to train in a fully unsupervised manner an off-the-shelf artificial deep recurrent neural network architecture in order to generate full inflections for verbs in the morphologically-rich Romanian, Spanish, and Finnish, introducing the task of *unsupervised inflection generation*. We showed that even when the training dataset is small for deep learning standards and without any supervision, we can achieve accuracy close to the state of the art for Finish and Spanish and we surpass previous state-of-the-art for Romanian.

Acknowledgements

SY would like to thank Noisebridge Hackerspace in San Francisco for use of their computing facilities.

References

1. Aharoni, R., Goldberg, Y.: Morphological inflection generation with hard monotonic attention. In: Proceedings of the 55th Annual Meeting of the Association for

Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 2004–2015 (2017), https://doi.org/10.18653/v1/P17-1183

2. Ahlberg, M., Forsberg, M., Hulden, M.: Paradigm classification in supervised learning of morphology. In: Mihalcea, R., Chai, J.Y., Sarkar, A. (eds.) NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 1024–1029. The Association for Computational Linguistics (2015), http://aclweb.org/anthology/N/N15/N15-1107.pdf
3. Alec Radford, Karthik Narasimhan, T.S.I.S.: Improving language understanding by generative pre-training
4. Alec Radford, Karthik Narasimhan, T.S.I.S.: Language models are unsupervised multitask learners
5. Barbu, A.M.: Conjugarea verbelor româneşti. Dicţionar: 7500 de verbe româneşti grupate pe clase de conjugare. Bucharest: Coresi (2007), 4th edition, revised. (In Romanian.) (263 pp.)
6. Barbu, A.M.: Romanian lexical databases: Inflected and syllabic forms dictionaries (2008)
7. Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., Hulden, M.: The SIGMORPHON 2016 shared task—morphological reinflection. In: Proceedings of the 2016 Meeting of SIGMORPHON. Association for Computational Linguistics, Berlin, Germany (August 2016)
8. Şulea, O.M.: Semi-supervised approach to romanian noun declension. In: Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 20th International Conference KES-2016, York, UK, 5-7 September 2016. pp. 664–671 (2016)
9. Dinu, L.P., Şulea, O.M., Niculae, V.: Sequence tagging for verb conjugation in romanian. In: RANLP. pp. 215–220 (2013)
10. Dinu, L.P., Ionescu, E., Niculae, V., Şulea, O.M.: Can alternations be learned? A machine learning approach to verb alternations. In: Recent Advances in Natural Language Processing 2011. pp. 539–544 (September 2011)
11. Dinu, L.P., Niculae, V., Şulea, O.M.: Learning how to conjugate the romanian verb. Rules for regular and partially irregular verbs. In: European Chapter of the Association for Computational Linguistics 2012. pp. 524–528 (April 2012)
12. Durrett, G., DeNero, J.: Supervised learning of complete morphological paradigms. In: Vanderwende, L., III, H.D., Kirchhoff, K. (eds.) Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. pp. 1185–1195. The Association for Computational Linguistics (2013), http://aclweb.org/anthology/N/N13/N13-1138.pdf
13. Faruqui, M., Tsvetkov, Y., Neubig, G., Dyer, C.: Morphological inflection generation using character sequence to sequence learning. CoRR abs/1512.06110 (2015), http://arxiv.org/abs/1512.06110
14. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2017)
15. Guţu-Romalo, V.: Morfologie Structurală a limbii române. Editura Academiei Republicii Socialiste România (1968), in Romanian

16. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: ACL. Association for Computational Linguistics (2018), http://arxiv.org/abs/1801.06146
17. Woolf, M.: textgenrnn, http://github.com/minimaxir/textgenrnn
18. Zhou, C., Neubig, G.: Morphological inflection generation with multi-space variational encoder-decoders. In: Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, Vancouver, BC, Canada, August 3-4, 2017. pp. 58–65 (2017), https://doi.org/10.18653/v1/K17-2005