WixNLP: Probabilistic Finite-State morphological analyzer for Wixarika language

Mager Hois, Jesus Manuel
Universidad Autónoma Metropolitana, Unidad Azcapotzalco
jmmh@correo.azc.uam.mx

Carrillo González, Diónico
dionico94@gmail.com

Meza Ruíz, Ivan Vladimir
Universidad Nacional Autónoma de México, IIMAS
ivanvladimir@turing.iimas.unam.mx

Abstract

We present the first morphology analyzer for the Mexican indigenous language Wixarika also known as Huichol. Indigenous languages in Mexico are seldom studied with an NLP focus. However, in recent years there has been a growing interest to include them into the research. Most of those languages share a complex agglutinative morphology of their verbs and have few written digital resources. A morphological analyzer is fundamental for other NLP tasks, such as MT. We further explain that a probabilistic finite state approach that exploits Yutonahua language agglutinative pattern that requires low linguistic knowledge, and then we show that our approach outperforms unsupervised methods in a low-resource context. The dataset used in this work was released for future work.

1 Introduction

In this paper, we present a probabilistic finite-state open-source morphological analyzer for the indigenous language Wixarika [1]. This language is spoken in the Mexican states of Jalisco, Nayarit, Durango and Zacatecas by approximately 50 thousand people. As most of South and North American indigenous languages, Wixarika has a complex verb morphology too (Campbell and Grondona, 2012), e.g. the word *nep+ka'ukats+k+* is segmented as *ne-p+-ka-'u-ka-ts+k+* and translating it into English means "I don't have a dog". Wixarika's alphabet is a set of 18 symbols: $\Sigma = \{a, e, h, i, +, k, m, n, p, r, t, t, s, u, w, x, y, \}$. As the symbol + is part of the alphabet, we use - to delimit morphemes.

[1]The software is available from https://github.com/pywirrarika/smtwixes/tree/master/wixnlp

Morphological segmentation is an important task that helps to improve other areas in natural language processing, especially for morphologically rich languages. To achieve this, a word w needs to be segmented into a tuple of substrings called morphs. The research has focused on unsupervised methods, but they can only be applied to languages for which there exists a sufficiently large corpus of words (Ruokolainen et al., 2016). For indigenous languages with scarce available resources, this is a limitation that bounds the quality of these methods. Efforts to gather large collections of digital texts for Yutonahua languages exists only for Nahuatl (Gutierrez-Vasques et al., 2016).

On the other hand, rule-based automatic morphological analyzers require a deep knowledge of the language and the expensive support of linguists (Creutz and Lagus, 2005). Rule-based morphological analyzers were developed for Quechua, Toba (Porta, 2010) and Aymara (Homola, 2011). But this again is a limitation for languages that has not been sufficiently studied. Our approach to handling scarce linguistic knowledge and digital corpus for morphological segmentation of Wixarika is a hybrid system, that combines language knowledge and a probabilistic model learned from previous seen segmented words.

Our contributions are as follows: the presentation of the first morphological analyzer for Wixarika, using a set morphemes and stems together with and a n-gram model. This hybrid method can achieve good performance for a morphologically rich language with scarce resources and low grammatical knowledge.

2 Method

Wixarika belongs to the family of Yutonahua languages, like Nahua, Nayeri, Raramuri, etc., and therefore shares agglutinative morphology using prefixation as well as suffixation, mainly around

the verb stem. The agglutination is very regular, and each morpheme has a specific position around the stem. The same morpheme in a different position has other meanings e.g., the prefix *ne-* in position 17 acts as a pronominal mark, and on position 4 is a possessive morpheme (Gómez, 1999). There are 18 such prefix positions and 23 suffixes identified by Iturrio and Gómez López (1999), where each position contains a given set of morphemes, plus an empty symbol. This information is not a complete grammar but can be used to model a Finite-State Machine. Although there are more complex morphological rules, we will assume that the only condition is the order of the sets of morphemes. The infractions to the morphological rules will be corrected later by the n-grams model.

As the stem is not defined by any rule and can even import words from other languages. For the present study, we limited the possible stems to a tuple of 374 strings learned from examples. A finite-state transducer can recognize any string w searching to find a path that matches the search space, returning a set of tuples for each path found in this space. The transducer can be modeled as a forward graph as shown in figure 1 as a general model for Yutonauha languages.

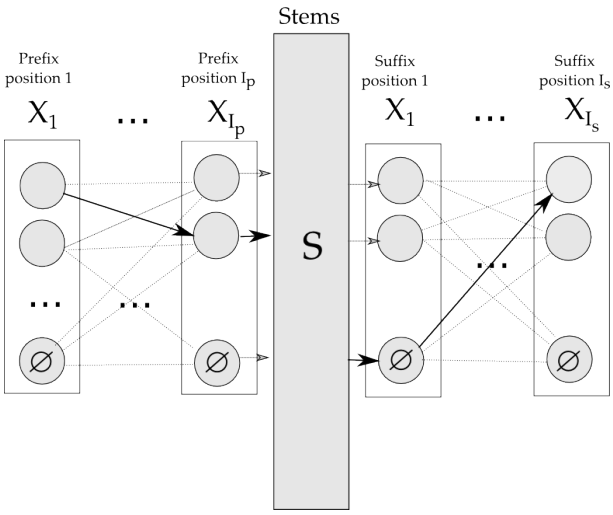

Figure 1: Generalized Final State Transducer for segmentation of verbs of Yutonahua languages

This approach contains evident limitations: the transducer shouldn't stop at the first match because the path is not guaranteed to be the best segmentation, therefore we need to explore the entire space, and decide the *best* segmentation; and if a word contains an unknown stem, the transducer cannot recognize the string and will fail. For the first issue, a simple n-gram model, where each gram is a morpheme, can be trained from segmented corpus. As the transducer already found all possible tuples, we only need to evaluate the probabilities for each n-gram and return the best-ranked tuple. The second issue is more difficult to solve. Irregular agglutinations and unknown stems can mislead the transducer and it will fail to recognize the word. If this happens, we can use an unsupervised method to analyze this word. Usually an unsupervised analyzer under-performs with scarce resources, but can improve in practice the final segmentation.

3 Results

For our experiment we have collected two corpora: the first is a high quality segmented text taken from Gómez (1999) containing 1079 unique words, which we will use as our golden standard. We randomly extracted 400 words from this collection, to be used as a test set, and the rest were used for the training of a semi-supervised morfessor model and our n-grams model. The second text is a translation of classical Anderson's fairly tales[2] to Wixrika containing an estimation of 47131 segmentable words, used for the training of the unsupervised morfessor model.

As the baseline, we used unsupervised Morfessor (Virpioja, 2013) trained with the Viterbi algorithm as well as its semi-supervised method. As unsupervised methods are language independent, the greatest limitation for a good performance is the number of words available for training. Morfessor was also used for the hybrid model.

Evaluating morphological segmentations is difficult since for a single word there are several valid segmentations. There are two categories of metrics for morphologies: those that directly compare the hypotheses against the golden standard and those that perform the comparison indirectly "y by measuring the strength of an isomorphic like relationship between the proposed and answer morphemes" (Spiegler and Monson, 2010). Among the various metrics that have been proposed for morphological evaluation, Virpioja et al. (2011) did a comparison between them.

We will use both of them. For direct comparison we do the evaluation as Kann et al. (2017) us-

[2]The dataset is available from https://github.com/pywirrarika/wixarikacorpora

Method	ED	1-best
Morfessor	64.95	0.213
Morfessor SS	49.93	0.355
WixNLP	41.77	0.477
WixNLP 2-grams	39.16	0.485
WixNLP 3-grams	32.48	0.579
Hybrid 2-grams	31.48	0.562
Hybrid 3-grams	27.85	0.599

Table 1: Results for the morphological segmentation task on Wixarika using direct comparison.

ing the error rate (the proportion of analyses that are completely correct) subtracted to 1, referred as 1-best, and the Edit Distance between the hypothesis and the golden standard. For the indirect evaluation we will use EMMA (Spiegler and Monson, 2010), which takes into account: precision, recall and the F-measure.

We compare against two baselines: Semi-supervised morfessor and Unsupervised Morfessor. WixNLP looks for all possible paths in the forward graph and chooses the shortest valid path. WixNLP with n-grams estimates the most probable segmentation among the valid paths. the hybrid segmentations additionay use unspervised segmentation for non segmented words by WixNLP.

	P	R	F
Morfessor	0.508	0.480	0.493
Morfessor SS	0.648	0.626	0.637
WixNLP	0.666	0.724	0.694
WixNLP 2-grams	0.697	0.733	0.710
WixNLP 3-grams	0.726	0.757	0.742
Hybrid 2-grams	0.739	0.773	0.756
Hybrid 3-grams	0.780	0.805	0.792

Table 2: Results for the morphological segmentation task on Wixarika using EMMA metric. P stands for precision, R for recall and F for the F-measure.

Table 1 and 2 shows the experimental results using Morfessor suffers from the lack of resources. Our first model called WixNLP without maximizing the probabilities of paths improves Morfessor in all metrics. Also, WixNLP with 2 grams and 3 grams improve the results notably. The hybrid approach, when dealing with the problem of unseen roots and suffixes, achieves the best results in all metrics, with a 3-grams model.

4 Conclusion

Morphology segmentation is an important task for language processing of American indigenous languages. In this work we presented the first Wixarika morphology analyzer, a finite-state transducer that exploits the agglutinative pattern of Yutonahua languages, using a morpheme table and a stem tuple, together with a n-gram model to estimate the best segmentation among multiple matches. We showed that our method improves the non-language specific baseline Morfessor, for the specific case Wixarika. We also created and publicly released a parallel data set Wixarika-Spanish to encourage the community to research it.

For future, work we would like to apply this methodology to other Yutonahua languages. Morfological segmentation also opens the possibility to improve SMT for American indigenous languages. For Wixarika some prior work has been already done by Mager Hois et al. (2016) combining SMT and morphological analysis.

References

Lyle Campbell and Verónica Grondona. 2012. *The indigenous languages of South America: a comprehensive guide*, volume 2. Walter de Gruyter.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Paula Gómez. 1999. *Huichol de San Andrs Cohamiata, Jalisco*. Archivo de lenguas indígenas de Mxico. Colegio de México.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Petr Homola. 2011. Parsing a polysynthetic language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. RANLP 2011 Organising Committee, Hissar, Bulgaria, pages 562–567. http://www.aclweb.org/anthology/R11-1079.

José Luis Iturrio and Paula Gómez López. 1999. *Gramática Wixarika I*. Archivo de lenguas indígenas de México. Lincom Europa.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. Neural multi-source morphological reinflection. In *Proceedings of the 2017 Conference European Chapter of the Association for Computational Linguistics*. Valencia, Spain.

Jesús Manuel Mager Hois, Carlos Barron Romero, and Ivan Vladimir Meza Ruíz. 2016. Traductor estadístico wixarika - español usando descomposición morfológica. *COMTEL* (6).

Andres Osvaldo Porta. 2010. The use of formal language models in the typology of the morphology of amerindian languages. In *Proceedings of the ACL 2010 Student Research Workshop*. Association for Computational Linguistics, Uppsala, Sweden, pages 109–114. http://www.aclweb.org/anthology/P10-3019.

Teemu Ruokolainen, Oskar Kohonen, Kairit Sirts, Stig-Arne Grönroos, Mikko Kurimo, and Sami Virpioja. 2016. A comparative study of minimally supervised morphological segmentation. *Computational Linguistics* .

Sebastian Spiegler and Christian Monson. 2010. Emma: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '10, pages 1029–1037.

Sami; Smit Peter; Grönroos Stig-Arne; Kurimo Mikko Virpioja. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. D4 julkaistu kehittmis- tai tutkimusraportti tai -selvitys. http://urn.fi/URN:ISBN:978-952-60-5501-5.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *TAL* 52(2):45–90.