WixNLP: Probabilistic Finite-State morphological analyzer for Wixarika language

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

Diónico Carrillo
dionico94@gmail.com

Ivan Meza
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México,
ivanvladimir@turing.iimas.unam.mx

Abstract

We present the first morphological analyzer for the Mexican indigenous language Wixarika, also known as Huichol. Indigenous languages in Mexico have seldom been the focus of NLP. However, with the recent reach of computing technology into indigenous communities, this has started to change and has generated a growing interest in the topic. An important aspect that these languages share is complex agglutinative verbal morphology. In this work, we present a morphological analyzer for Wixarika which also holds this property. On top of the agglutinative nature of the language, the low amount of resources and the lack of a orthographic standard among dialects add to the challenge. Our proposal is based on a probabilistic finite-state approach that exploits regular agglutinative patterns and requires little linguistic knowledge. 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 morphological analyzer for the Wixarika indigenous language.[1] Wixarika is a language spoken in the Mexican states of Jalisco, Nayarit, Durango and Zacatecas (in central west Mexico) by approximately fifty thousand people. Like most South and North American indigenous languages, Wixarika has complex verbal morphology (Campbell and Grondona, 2012). For instance, the word *nep+ka'ukats+k+*, which can be translated into English as "I don't have a dog," is segmented into the morphs *ne|p+ |ka|'u|ka|ts+k+*. The symbol + denotes one of the vowels in the language; for this reason, we use | to delimit morphemes. Notice that although this word is a verb form, its agglutinative nature makes it a full sentence. In this example *ts+k+* is the stem and means "dog".

Different linguistic studies have recorded Wixarika in written form, but its spelling is still not standardized. The most common spelling in practice by native speakers is an alphabet of 18 symbols: $\Sigma = \{a, e, h, i, +, k, m, n, p, r, t, s, u, w, x, y, \}$, as proposed in Gómez (1999) and Iturrio and Gómez López (1999). Our dataset is written with this convention.

Morphological segmentation is an important task that helps to improve other areas of natural language processing, especially for morphologically rich languages. Each word w needs to be segmented into a tuple of substrings called morphs. Past 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 exist only for Nahuatl (Gutierrez-Vasques et al., 2016). For Wixarika NLP, some prior work on SMT has been done (Mager Hois et al., 2016). Our proposal sprang from this effort since morphology had a notable impact on the translation performance.

On the other hand, rule-based automatic morphological analyzers require deep knowledge of the language or the expensive support of linguists (Creutz and Lagus, 2005). Rule-based morphological analyzers have been developed for Quechua, Toba (Porta, 2010) and Aymara (Homola, 2011). However, it is difficult to tackle poorly studied languages. Our approach to the

[1] The analyzer is open source and can be download it from: https://github.com/pywirrarika/smtwixes/tree/master/wixnlp

morphological segmentation of Wixarika deals with the scarcity of linguistic knowledge and digital corpora, since it is a hybrid system that combines language knowledge with a probabilistic model learned from supervised data (previously seen segmented words).

Our contribution is the construction of the first morphological analyzer for Wixarika, using hand-specified lists of legal stems and affixes together with an n-gram model that describes sequences. 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, such as Nahua, Nayeri, Raramuri, etc.. These languages have agglutinative morphology, using prefixation as well as suffixation around the verb stem. The agglutination is almost strictly concatenative, and each morpheme must be realized at a specific position in the word. The same string in a different position conveys a different meaning: e.g., the prefix *ne-* in position 17 acts as a pronominal morpheme, but in position 4 it is a possessive morpheme (Gómez, 1999). There are 18 such prefix positions and 23 suffix positions identified by Iturrio and Gómez López (1999), where each position allows a certain set of morphemes (or can be left empty).

This description of the language can be used to construct a finite-state machine (FSM) from a list of legal morphemes at each position. Although there are more complex rules that govern sequences of morphemes, we will assume that the only condition is that each position allows only morphemes from its list. The errors introduced by this assumption will be corrected later by the n-gram model.

The stem is not defined by any rule and it can be based on words from other languages (e.g., Spanish). For the present study, however, we limited the possible stems to a tuple of 374 strings learned from examples. The list of sets used for affixes was taken from the linguistic work of Iturrio and Gómez López (1999), which is a revision of an earlier study (Grimes, 1964).

A finite-state automaton can accept any string w in the language that it defines, and returns a set of accepting paths. The automaton for Wixarika verbs is shown in Figure 1; its different accepting paths for a word w correspond to different morphological analyses of w. In practice, there are few enough analyses that we can enumerate all of them. To choose the most probable analysis from among these, we used a simple n-gram model with Kneser-Ney (1995) smoothing, where each gram is a morph (a surface string associated with some morpheme). This model scores the sequence of non-empty strings (morphs) without considering their absolute positions. As a result, it can be trained simply from a segmented corpus.

Our WixNLP system was evaluated with both 2-gram and 3-gram models. Irregular agglutinations and unknown stems can mislead the automaton, so it sometimes fails to recognize an input word. If this happens, we can fall back to an unsupervised method to analyze this word. Usually an unsupervised analyzer under-performs with scarce resources, but it can improve the final segmentation in practice.

3 Results

For our experiment we collected two corpora. The first is a high-quality segmented text taken from a grammar (Gómez, 1999) containing 1079 unique words, which we used as our gold 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-gram model. The second text is a translation of Hans Christian Andersen's classic fairy tales[2] to Wixarika containing an estimation of 47,131 segmentable words, used for the training of the unsupervised Morfessor model.

Evaluating morphological segmentations is difficult since for a single word there are several valid segmentations. There are two types of metrics for morphologies: those that directly compare the hypotheses against the gold standard and those that perform the comparison indirectly "by measuring the strength of an isomorphic like relationship between the proposed and answer morphemes" (Spiegler and Monson, 2010).[3]

In this work, we used both types of metrics. For direct comparison we follow Kann et al. (2017) using the error rate (the proportion of analyses that are completely correct) subtracted from 1, referred to as 1-best, and the edit distance of morphs between the hypothesis and the golden standard. For the indirect evaluation we used EMMA (Spiegler

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

[3]For a comparison among the various metrics see Virpioja et al. (2011).

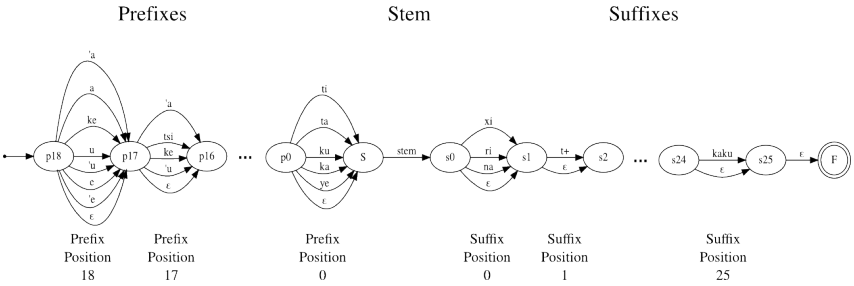

Figure 1: Extract of the FSA for Wixarika verbs. The "stem" arc stands for a collection of 374 arcs representing different stems.

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 to the gold segmentation: Edit distance (ED) and error rate (1-best).

and Monson, 2010), which produces precision, recall and F-measure scores.

The **WixNLP** system 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.

	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.

We compared against two baselines: **Unsupervised Morfessor** and **Semi-Supervised Morfessor** (Virpioja et al., 2013). Finally, the **Hybrid** methods use WixNLP when possible but fall back to Unsupervised Morfessor for words that have no valid paths.

Tables 1 and 2 show that the experimental results using Morfessor suffer from the lack of resources. Our first model, WixNLP, improves Morfessor without even using probabilities for disambiguation. WixNLP with 2-grams and 3-grams improve the results notably. The hybrid approach deals with the problem of unseen roots and suffixes, and thus achieves the best results in all metrics, particularly with a 3-gram model.

4 Conclusion

Morphological segmentation is an important task for language processing of indigenous languages. In this work we presented the first Wixarika morphology analyzer, a finite-state transducer that exploits the agglutinative pattern of Yutonahua languages, with lists of stems and affixes, together with a n-gram model to estimate the best segmentation among multiple matches. We showed that for Wixarika our method improves on the Morfessor baselines. We also created and publicly released a parallel Wixarika-Spanish dataset to encourage the community to study this language further.

For future work, we would apply this methodology to other Yutonahua languages. We also want to feed the morphological segmentation to a MT system. It is also desirable to find improved methodologies to combine unsupervised with supervised methods to address the scarce resource problem for agglutinative languages, including tagging each morph as in (Spoustová et al., 2007) and (Smith et al., 2005).

Acknowledgement

We thank the guidance, feedback and support of Professor Jason Eisner as a mentor of this work.

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.

Joseph Grimes. 1964. *Huichol Syntax*. Mouton, The Hague.

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.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *ICASSP*.

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* .

Noah A. Smith, David A. Smith, and Roy W. Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '05, pages 475–482. https://doi.org/10.3115/1220575.1220635.

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.

Drahomíra "johanka" Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 67–74. http://dl.acm.org/citation.cfm?id=1567545.1567558.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 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.