Inflection detection:
Finite-state morphological analysis for French verbs using Pynini

Martine Harrison
aharrison1@gradcenter.cuny.edu

Abstract

This paper seeks to put forward motivation for finite-state morphological analysis, as well as outline an approach to this knowledge-based series of tasks developed for French verbs in Python using Pynini, an open-source library intended for streamlined grammar-building using WFSTs (weighted finite-state transducers). Unimorph, a series of hand-engineered and annotated corpora designed to deliver underlying morphological representations for many world languages, is used by the morphological analyzer: its overarching annotation schema for deriving morphological labels, as well as its `fra` dataset toward morphological analysis and later, evaluation. The analyzer is evaluated using word error rate, a commonly-used performance metric.

1 Introduction and background

Work in the area of morphological analysis using finite-state techniques not only has substantial historical precedent, but is enduringly influential for computational linguists working with lemmatization and morphological tagging, as well as morphological generation and analysis Gorman and Sproat (2021). Until the new millennium, computational morphology was, in fact, "completely dominated by finite-state approaches" Roark and Sproat (2007). More recently, this has shifted, with sequence-to-sequence architectures increasingly being tested for their utility and deployed towards capturing the inflectional morphology of various languages. Earlier attempts within the area of sequence-to-sequence prediction were themselves, at least in part, knowledge-based, with exploratory work incorporating transformational rules learned from inflection charts Durrett and DeNero (2013), as well as transduction Nicolai et al. (2015) coming to the fore Elsner et al. (2019).

As work within this vein progressed, a new interest emerged, directing itself toward encoder-decoder models, which use recurrent neural networks (RNNs) for sequence-to-sequence prediction. Originally developed for machine translation, encoder-decoders have also found applications in text summarization, caption generation, and generative chatbots, where LSTM or GRU layers are typically applied. The widespread utility and "representational flexibility" Elsner et al. (2019) of these models were not lost on inflection generation, with their "capab[ility] of learning non-concatenative morphological processes" and of "stem-affix relationships, both morphological and phonological" Elsner et al. (2019) being especial qualities—capturing phenomena such as reduplication and vowel harmony with an impressive degree of generalizability—which marked them as valuable addition to a long-held knowledge-based paradigm.

A robust preliminary comparison of the performance of finite-state grammars and sequence-to-sequence neural models in the area of computational morphology was brought forward by Beemer et al. in their paper, 'Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars'. The performance of finite-state grammars, which are themselves comprised of cascades of morphological transformations embedded in transducers, were compared across over 25 languages with that of state-of-the-art neural models. Though handwritten grammars were found to be competitive with neural technology, in certain cases reliably surpassing the performance of their counterparts, the authors explicate this success in terms of human labor time, as well as in terms of flaws in the data. Success for any given finite-state grammar must occur across data of adequate complexity in order for its competitivity with neural models to be considered non-trivial. Conversely, in cases where high-complexity grammars are being used upon very morphologically rich data, human labor time expended must be taken into account, with grammars of this variety extending into the hundreds of hours to construct Beemer et al. (2020).

Using the insights brought by Beemer et al. as to the meaningfulness of the predictive power of FST-driven grammars given complexity of data and time spent in development, this paper will attempt to situate morphological analysis for regular French verbs using WFSTs within a broader scope of how technology of this variety is understood and implemented.

2 Unimorph

Unimorph's French dataset (`fra`), like many of the sets available via the database, is not grammatically exhaustive. At the time of this paper's writing, it is comprised only of verbs (around 7,500 lemmas), with the compound tense-aspect forms (made up of an auxiliary verb, *avoir* or *être*, and participle) such as the *passé composé* (compound past), *plus-que-parfait* (pluperfect), and *conditionnel passé* (past conditional) being notably absent. The sole compound verb form found within the dataset is the *gérondif* (gerund), labeled by Unimorph as a converb (`V.CVB`). The remaining available verb forms are not compounded, such as the present-tense forms of the language's four finite moods, as well as the *passé simple* (simple past), *futur simple* (simple future), and *imparfait* (past imperfective). Past and present verb participles are also included. Each row of the dataset is comprised of a lemma gloss, an inflected form, and a morphological feature vector Sylak-Glassman (2016):

```
noter notez V;IND;PRS;2;PL
```

According to Unimorph's annotation schema, French verbs are necessarily annotated for mood, and optionally annotated for polarity, tense, aspect, number, and person. There are some minor inconsistencies in the set's analytical labels. For example, verb with inflects annotated as carrying the conditional mood (`COND`), despite being morphologically marked as present tense, are not annotated as such in the dataset. This leads to a less-than-ideal analysis on the part of the morphological analyzer—in the interest of retaining consistency with the dataset's labeling scheme, present conditional verbs are not annotated for tense, despite carrying morphological tense-markers.

Given the observations of Beemer et al., paucities within the data used toward grammatical modeling should be taken into account when assessing the meaningfulness of the results received when the grammar is tested.

3 The analyzer

Constructed using Pynini, the analyzer defines three paradigmatic French verb forms: verbs ending in *-er*, *-ir*, and *-re*. Given that all French verbs end in one of these three ways, this does extend the grammar's coverage to any verb handed it. However, a small minority of French verbs (e.g. *aller*, *avoir*, *être*) receive irregular conjugations, putting the grammar at a disadvantage in its chances of generating a correct inflection for any given irregular verb form. For example, the analyzer will correctly generate an inflection and feature vector for irregular verb *aller* in its present indicative, first person plural form—*allons*—by reshaping the stem to *all-* and attaching *-ons*. However, irregularly conjugated verb forms, like the present indicative, third person singular form (*va*) will be incorrectly inflected (in this case as '*alle*').

The analyzer also follows Unimorph's annotation schema for French verbs, meaning most compound verb forms are not incorporated. Given some lemma, verbs are analyzed and inflected along the six features present in the `fra` dataset: polarity (A), mood (B), aspect (C), tense (D), number (E), and person (F). A line of the analyzer's raw output looks like the following:

```
dégage[A=none][B=V;SBJV]
[C=PRS][D=3][E=SG][F=none]
```

Providing inflection schemata for semi-regular, stem-changing, and irregular French verbs is more than possible in Pynini. Semi-regular verbs could potentially be correctly inflected and analyzed by implementing weights, which were not used here. Highly irregular verbs, like *être*, could be handled on a case-by-case basis by defining individual transducers which map from from word forms to their corresponding grammatical features Gorman and Sproat (2021). Following the constraints on human labor time brought up by Beemer et al., this may mean that said approach has disadvantages: "the list of forms would be need to be...large to obtain broad coverage" Gorman and Sproat (2021).

4 Evaluation

In order to evaluate the morphological analyzer's performance, 100 percent of the lemmas present in `fra` were passed to the analyzer. Hypothesis labels generated were then compared with ground truth labels (i.e. the inflects/feature vectors found in `fra`), and all hypothesis labels which did not appear in the gold set were added to an error count. Rounded to the next whole number, a WER of 20 was obtained. Though error in this case was not negligible, the analyzer was not built to handle, nor was it evaluated against a large chunk of French verb forms. To have a more thorough idea of its performance, in the future, compound verb forms would hopefully be incorporated.

5 Conclusion

Though approaches to computational morphology have diversified in recent decades, finite-state technologies remain a viable way to construct knowledge-based solutions to many morphological problems. Because they differ architecturally from their neural counterparts, their successes and failures must be assessed using alternate considerations, particularly when being compared with said counterparts. The success of any handwritten grammar must be weighed against the morphological richness and complexity of the data it was meant to express: in cases where the data is lacking in morphological complexity, success may be considered trivial; if the data is rich in morphology, one must consider the drawback presented by the dozens of hours that may be necessary for a trained linguist to develop such a grammar.

Future applications may benefit from equipping the morphological analyzer outlined here to handle semi-regular and irregular French verbs, as well as incorporating compound verb forms, in the interest of enhancing its ability to reflect the realistic complexity of French verbal inflectional morphology. This would not only increase the grammar's coverage, but would add dimension to the interpretation of evaluative results. Additionally, potential future iterations of this work could include testing performance on novel data, and comparison with sequence-to-sequence neural models.

References

Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, Piyush Mishra, Saliha Muradoglu, Lan Sang, Tyler Short, Sagarika Shreevastava, Elizabeth Spaulding, Testumichi Umada, Beilei Xiang, Changbing Yang, and Mans Hulden. 2020. Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 162–170. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Inflection Generation as Discriminative String Transduction. In *Proceedings of NAACL-HLT 2013*, pages 1185–1195, Atlanta, GA. Association for Computational Linguistics.

Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Karim Shuan, David L. King, Luana Lamberti Nunes, Byung-Doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V. Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of LanguageModeling*, 7(1):53–98.

Kyle Gorman and Richard Sproat. 2021. *Finite-State Text Processing*. Morgan and Claypool, San Rafael, CA.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection Generation as Discriminative String Transduction. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 922–931, Denver, CO. Association for Computational Linguistics.

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press, New York, NY.

John Sylak-Glassman. 2016. The composition and Use of the Universal Morphological Feature Schema (UniMorph Schema). Technical report, Center for Language and Speech Processing Johns Hopkins University.