Learning How to Conjugate the Romanian Verb

Rules for Regular and Partially Irregular Verbs

Liviu P. Dinu[1, 3]
ldinu@fmi.unibuc.ro

Vlad Niculae[1, 3]
vlad@vene.ro

Octavia-Maria Șulea[1, 2, 3]
mary.octavia@gmail.com

[1] Faculty of Mathematics and Computer Science
[2] Faculty of Foreign Languages and Literatures
[3] Center for Computational Linguistics
University of Bucharest

I. Stem alternations in Romanian verbs

Stem alternations, or **apophony**, is one of the reasons why the Romanian language is difficult to acquire.

For **partially irregular verbs** it is not enough to learn a generic suffix variation pattern, because there are simultaneous variations in the stem.

conjugation of "a purta":	nearly identical infinitives:	conjugation of "a curta":
eu port (I wear)		eu curtez (I court)
tu porți (you wear)	which one is simpler?	tu curtezi (you court)
el poartă (he wears)		el curtează (he courts)
noi purtăm (we wear)		noi curtăm (we court)
voi purtați (you wear)		voi curtați (you court)
ei poartă (they wear)		ei curtează (they court)

What is needed for automatic conjugation?

Romanian received a Latin-inspired classification of verbs into 4 conjugational classes, based on the ending of the infinitive form. This does not discriminate the two verbs shown above, so the standard model is **insufficient**.

The goal: given an infinitive form, know what letters change, and how they change.
The trick: craft a sufficient, near-exhaustive, disjoint set of conjugation classes.

II. Previous work

Moisil (1960): variable letters
purta = pu_0rt_0a where: u_0 = {u, oa, o}, t_0 = {t, ț}

Dinu, Ionescu (2011, unpublished): context-sensitive rules to decode variable letters for some verbs. Idea: alternations are identifiable by their context.

Dinu et al (2011): 7 conjugation classes for verbs ending in **-ta**. Knowing the class means knowing the alternations that occur. Idea: the classes can be learned.

A class corresponds to a **conjugation rule**: a set of 6 regular expressions matching the 6 conjugation forms of present tense verbs. Parts of the forms that are not accounted for must remain fixed, i.e. a rule **accounts for all the variation**.

Classification using **character n-gram** features + SVM: n-gram size chosen to be 3 for model simplicity (n ~= 5 is optimal)
Input: 'purta' => 'p', 'u', 'r', 't', 'a', 'pu', 'ur', 'rt', 'ta', 'pur', 'urt', 'rta'
Output: label in {1, 2, 3, 4, 5, 6, 7}, imbalanced classes
Results: 82.71% accuracy and 80% F-score

III. Crafting conjugation rules using regular expressions

Process

To manually expand a set of conjugation rules:
1. Select unmatched verb
2. Add rule to completely conjugate it
3. Match verbs against new rules

For example:

1. Say the verb 'a omorî' (to kill) is not matched
2. A conjugation rule matching this verb would be:

1sg:	^(.*)o(.*)$	omor
2sg:	^(.*)o(.*)i$	omori
3sg:	^(.*)oa(.*)ă$	omoară
1pl:	^(.*)o(.*)âm$	omorâm
2pl:	^(.*)o(.*)âți$	omorâți
3pl:	^(.*)oa(.*)ă$	omoară

3. This rule also matches, among others, the verb 'a doborî' (to defeat) so mark this one as matched too.

Results:

We threw out rules covering **<4** cases, leaving 30 rules covering 95% of the verbs

Rules overview:

rule #:	size:	rule #:	size:
1	547	16	13
2	8	17	6
3	18	18	4
4	5	19	14
5	8	20	124
6	16	21	25
7	3330	22	15
8	273	23	7
9	89	24	41
10	4	25	51
11	5	26	185
12	4	27	1554
13	106	28	486
14	13	29	5
15	5	30	27

Interaction between rules

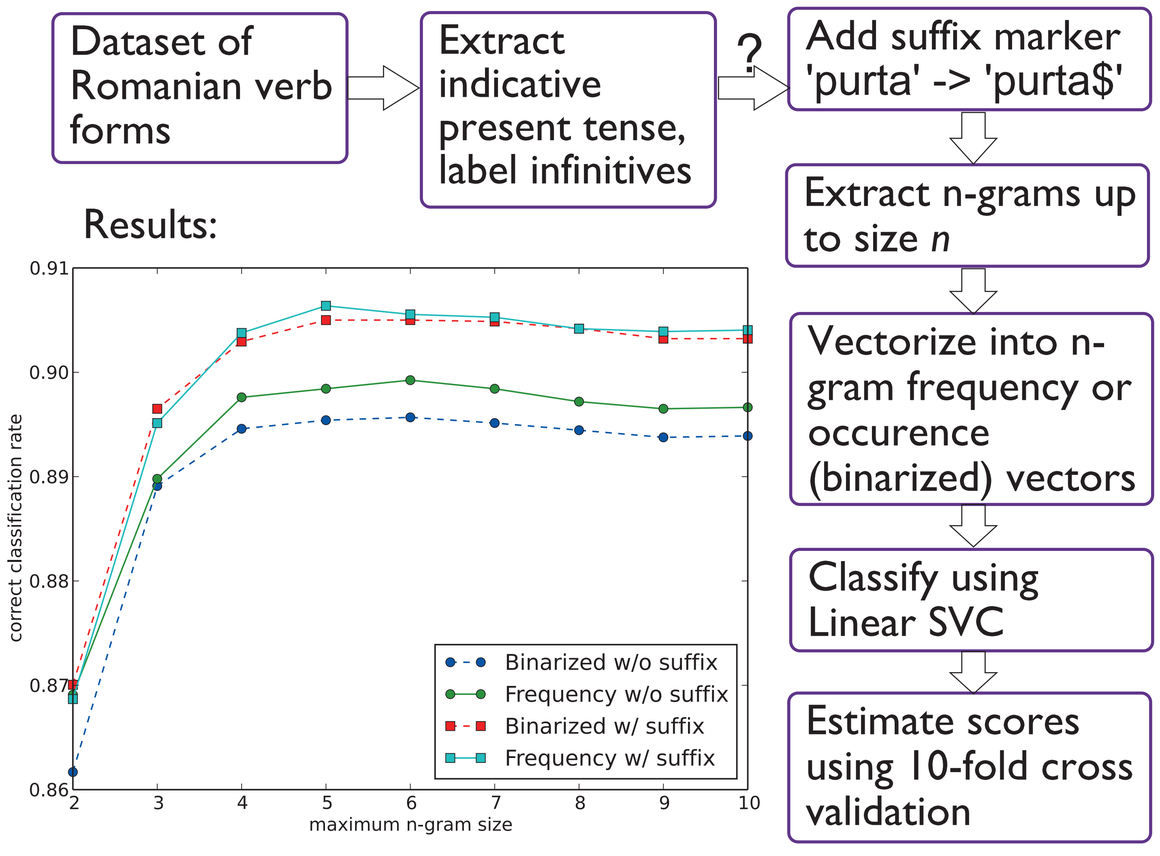
The largest covering rule has no alternations in the root, just the suffix. Other rules model 0-2 apophonys. Some rules correspond to the same variable letter, but it varies differently. For example:

Some rules overlap:

	rule 10 (a cânta)	rule 12 (a deștepta)	rule 13 (a deșerta)	rule 15 (a desfăta)
1sg	^(.*)t$	^(.*)e(.*)t$	^(.*)e(.*)t$	^(.*)ăt$
2sg	^(.*)ți$	^(.*)e(.*)ți$	^(.*)e(.*)ți$	^(.*)eți$
3sg	^(.*)tă$	^(.*)ea(.*)tă$	^(.*)a(.*)tă$	^(.*)ată$
1pl	^(.*)tăm$	^(.*)e(.*)tăm$	^(.*)e(.*)tăm$	^(.*)ătăm$
2pl	^(.*)tați$	^(.*)e(.*)tați$	^(.*)e(.*)tați$	^(.*)ătați$
3pl	^(.*)tă$	^(.*)ea(.*)tă$	^(.*)a(.*)tă$	^(.*)ată$

Rule 13 is much more productive than 10, 12 and 15, but we miss the importance of the t-ț alternation itself. It also occurs in rule 14, 21 and also verbs with too rare conjugation patterns to generalize ('a purta' is actually a singleton!)

IV. Classification methodology



V. Conclusion and perspectives

Estimated scores:

Parameters chosen by grid search: n=5, append '$', do not binarize, C=0.1
Correct classification rate: 90.64% (baseline choosing most probable class: 48%)
Weighted averaged precision: 80.90%, recall: 90.64%, F_1 score: 89.89%.
Appending the artificial terminator marker '$' consistently improves accuracy by around 0.7% irrelevant of the other parameters.
Frequency features perform slightly better than binarized ones for this task

What does this mean?

Verb conjugation can be learned with good scores, even with the assumption that classes don't interact. Our classes are **coarse-grained**. An **exhaustive model**, at least for the training data, will need to have many classes for unique and near-unique conjugation patterns. For better generalization: we need a finer-grained system.

Future work and collaboration ideas:

Build a more compact model by eliminating rule interaction: (see discussion above)
Compare with hand-crafted rule based conjugation
Try human evaluation on unseen, unlabeled verbs
Actually build a verb conjugation using classification output (trivial)
Extend to other languages with similar behaviour (Hebrew)