Sequence Tagging for Verb Conjugation in Romanian

Liviu P. Dinu Vlad Niculae Octavia-Maria Șulea

Center for Computational Linguistics
University of Bucharest
http://nlp.unibuc.ro

September 2013

Verbs in Romanian

Regularity is not black and white

		1st	2nd	3rd
Regular	sg.	merg	mergi	merge
a merge *(to walk)*	pl.	mergem	mergeți	merg
Irregular	sg.	sunt	ești	este
a fi *(to be)*	pl.	suntem	sunteți	sunt

Verbs in Romanian

Regularity is not black and white

		1st	2nd	3rd
Regular	sg.	merg	mergi	merge
a merge *(to walk)*	pl.	mergem	mergeți	merg
Irregular	sg.	sunt	ești	este
a fi *(to be)*	pl.	suntem	sunteți	sunt
Partially irregular	sg.	port	porți	poartă
a purta *(to wear)*	pl.	purtăm	purtați	poartă

Previous work

Dinu et al, RANLP 2011, EACL 2012

- Hand-crafted sets of regular expressions fully describing conjugation of most verbs
- Predictive model h(infinitive) = regular expression set

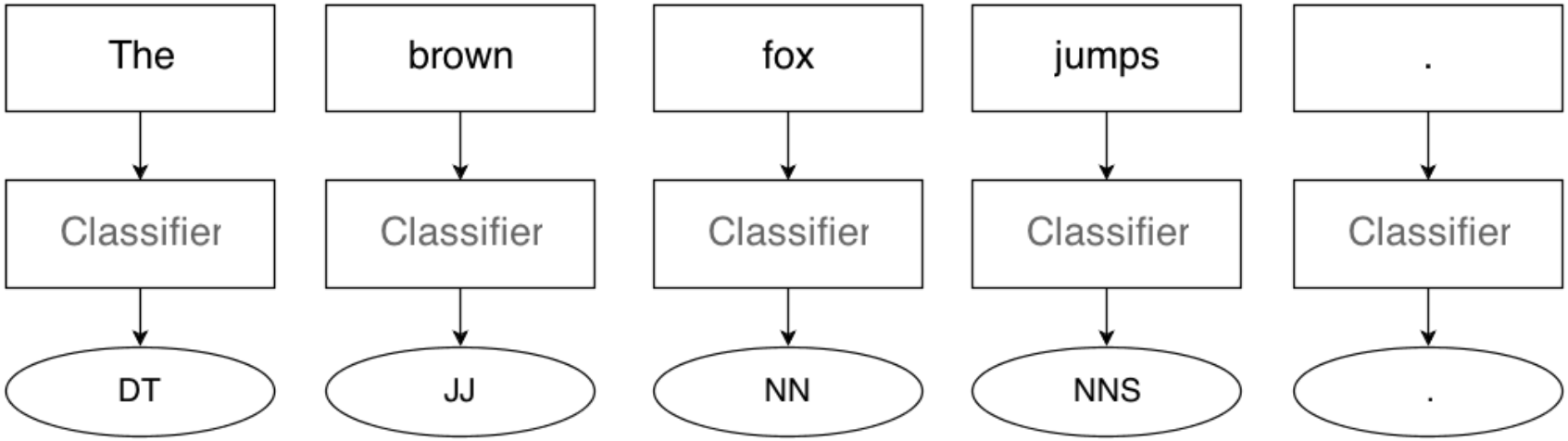
Running example

	sg.	port	porți	poartă
a purta *(to wear)*	pl.	purtăm	purtați	poartă

Regular expression set

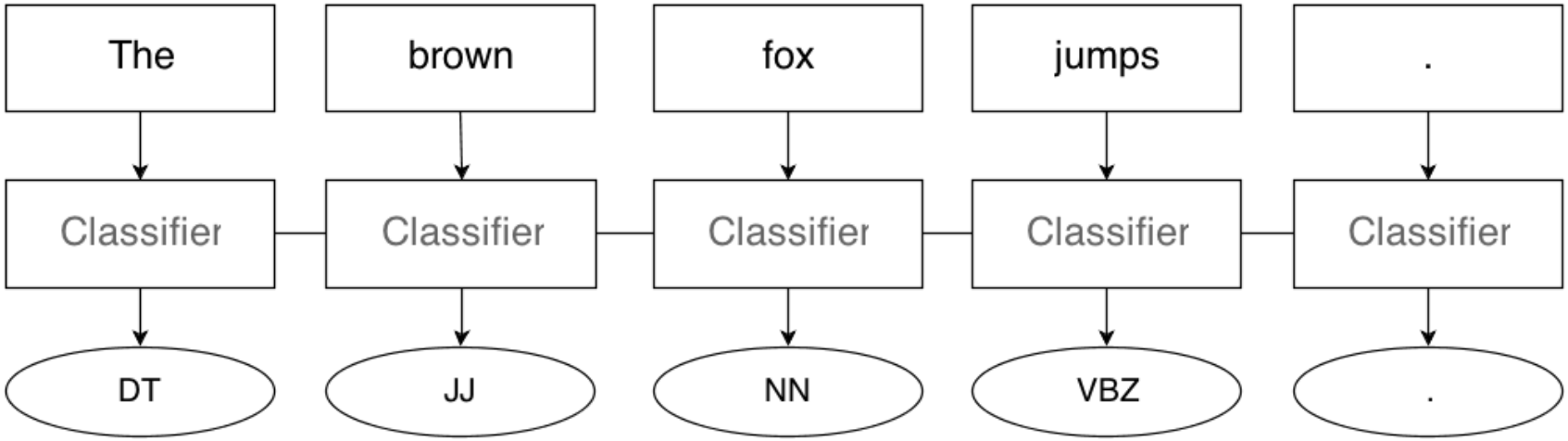
sg.	`^(.*)o(.*)t$`	`^(.*)o(.*)ți$`	`^(.*)oa(.*)tă$`
pl.	`^(.*)u(.*)tăm$`	`^(.*)u(.*)tați$`	`^(.*)oa(.*)tă$`

Sequence tagging: POS tagging example



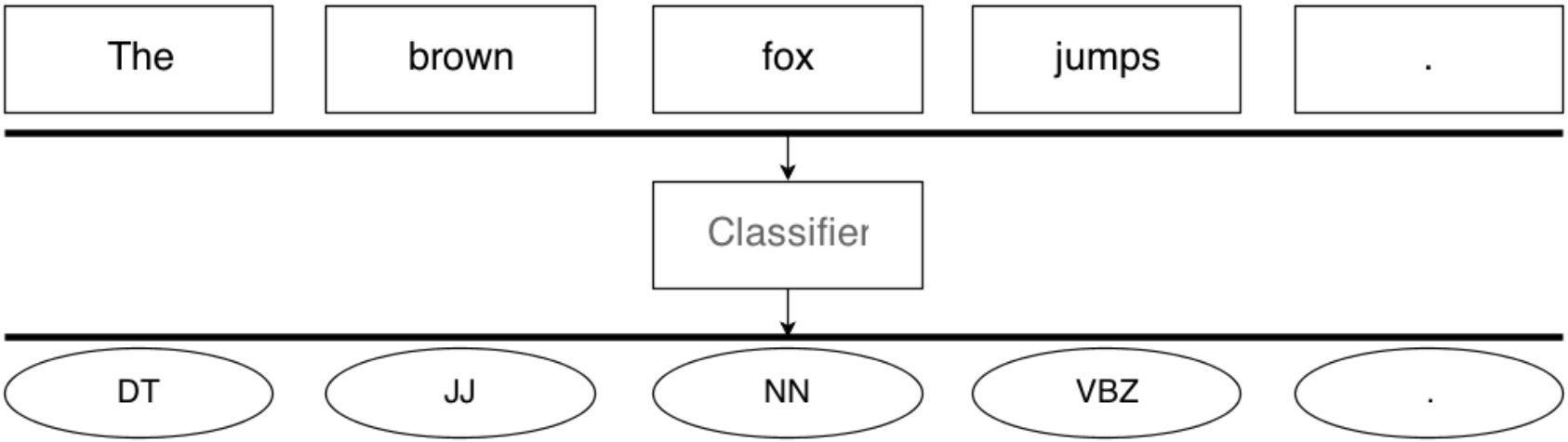
$$\prod \phi(y_i, x_i)$$

Sequence tagging: POS tagging example (better)



$$\prod \phi_1(y_i, x_i)\phi_2(y_i, y_{i+1})$$

Sequence tagging: POS tagging example (worse?)



$$\phi(y_1, y_2, \ldots y_n, x_1, x_2, \ldots x_n)$$

Ignored structure: interaction between classes

a cânta	a deștepta	a deșerta
to sing	*to rise*	*to empty*
`^(.*)t$`	`^(.*)e(.*)t$`	`^(.*)e(.*)t$`
`^(.*)ți$`	`^(.*)e(.*)ți$`	`^(.*)e(.*)ți$`
`^(.*)tă$`	`^(.*)ea(.*)tă$`	`^(.*)a(.*)tă$`
`^(.*)tăm$`	`^(.*)e(.*)tăm$`	`^(.*)e(.*)tăm$`
`^(.*)tați$`	`^(.*)e(.*)tați$`	`^(.*)e(.*)tați$`
`^(.*)tă$`	`^(.*)ea(.*)tă$`	`^(.*)a(.*)tă$`

Conjugation as sequence tagging

Running example

	sg.	port	porți	poartă
a purta *(to wear)*	pl.	purtăm	purtați	poartă

Variable letters (Moisil)

$$\text{form}(u_0|1sg) = \text{o} \qquad \text{form}(t_0|1sg) = \text{t}$$
$$\text{form}(u_0|3sg) = \text{o}a \qquad \text{form}(t_0|2sg) = \text{ț}$$
$$\text{form}(u_0|1pl) = \text{u}$$

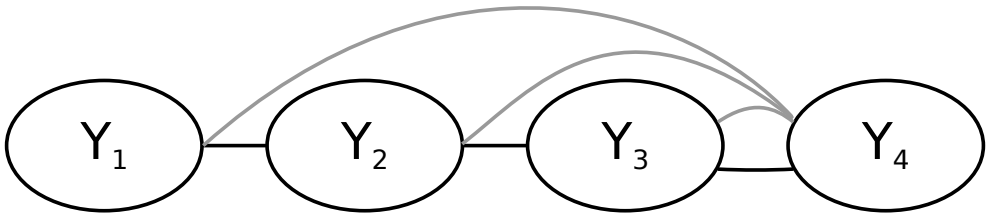
Tagging example

p	u	r	t	a
0	u_0	0	t_0	T_4

Models, features, training

- Features: character n-grams to the left and right size up to n
- Dataset: RoMorphoDict (lemmas and forms) labeled using the RegEx sets
  16 ending patterns, 17 variable letters
  $4,699$ train / $2,257$ test / 339 unlabeled
- Grid search, 10-fold cross validation

Skip-edge CRF



- An extra factor template allowing the ending to influence all positions
- Inference becomes more complex
- Out-of-the-box sequence tagging no longer appropriate

Results

	Cross-val. accuracy			Test accuracy		
method	word	char	char′	word	char	char′
SVM	0.886	-	-	0.896	-	-
ML	0.924	0.987	0.913	0.914	0.985	0.900
AP	0.923	0.987	0.917	0.912	0.985	0.900
PA	0.925	0.987	0.917	0.912	0.984	0.900
AROW	0.916	0.986	0.912	0.908	0.984	0.895
SKIP	-	0.984	-	0.906	0.983	0.896

Generalization on 105 of the unlabeled verbs:

- many termination patterns are correctly found (30)
- some alternations are found (3)