Prediction of *to*-infinitives and gerunds as complements using Machine Learning

Zi Huang
Hamit Kavas
Albert Lleó
Clàudia Martínez

INDEX

Goal of the project

To predict the use of to-infinitives and gerunds as complements of a predefined set of verbs: *start doing* vs. *start to do*

	to-inf	*-ing*
like (in all forms)	11643	741
like (not in base form)[1]	2897	541
begin	18386	2617
continue	9742	911
hate	322	234
love	914	373
prefer	1947	96
propose	1157	191
start	6019	6666
attempt	6326	22
intend	6297	272
not bother	143	11

Table 1: Raw frequency in BNC (SketchEngine)

Methodology

Extracting features from BNC:

- Length of the non-finite verb
  - *Begin to work*
- Adjectives or nouns surrounding the non-finite verb
  - *Wrong doing* vs. *To do wrong*
- Argument structure of the non-finite verb.
  - *Start to eat, start to eat an apple*
- Temporal anchoring
  - *Would like to do* vs. ?*would like working*

Apply Machine Learning

First results: extraction of features

Two verbs **start** and **hate** and four features:

GOOD SLIDE

Feature	Values
Length	Number of characters
Adjs/N surrounding the verb	Before | After
Argument structure	No complement | Noun Phrase |Prep. Phrase
Tense	Present | Past | Future | Conditional

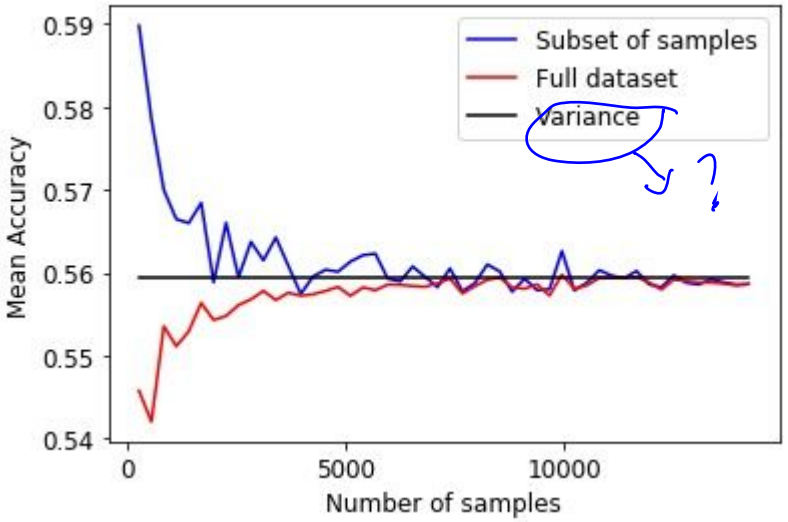
HAVE YOU THOUGHT ABOUT USING THE NON-FINITE VERB AS FEATURE? (SEMANTIC CLASS, ...)

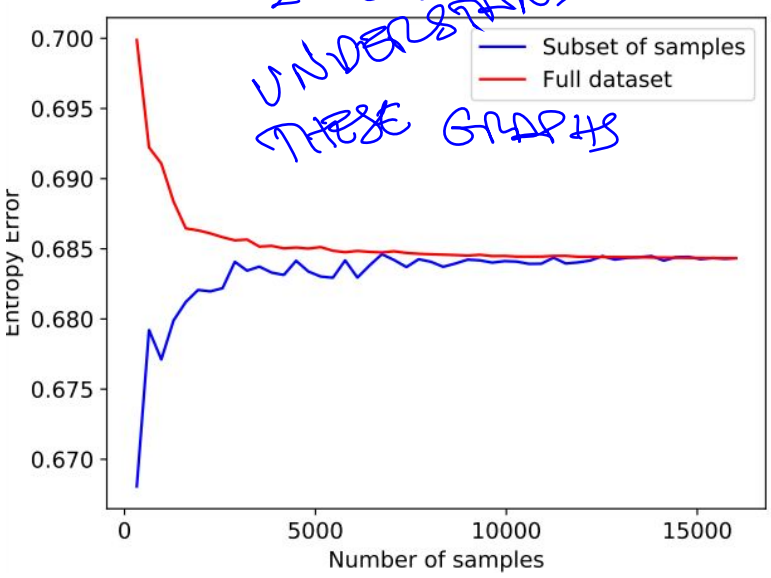
First results: extraction of features

GOOD SLIDE

id	sent_id	finite_verb	non_finite_	length	befaft_bef	befaft_aft	argstr_NC	argstr_NP	argstr_PP	tense_pre	tense_pst	tense_fut	tense_conc	verb_POS	target_form
3	2147	start	flow	4	0	0	0	0	1	0	0	1	0	VVG	1
4	3128	hate	kick	4	0	1	0	1	0	1	0	0	0	VVG	1
4	3262	start	write	5	0	0	1	0	0	1	0	0	0	VVG	1
5	4662	start	come	4	0	0	1	0	0	0	1	0	0	VV	0
5	4941	hate	hear	4	0	0	1	0	0	1	0	0	0	VV	0
6	5641	start	think	5	0	0	1	0	0	0	0	1	0	VV	0
6	6940	start	take	4	0	1	0	1	0	1	0	0	0	VV	0
6	7277	hate	see	3	0	0	0	0	1	0	1	0	0	VVG	1
6	7934	start	act	3	0	0	1	0	0	0	1	0	0	VVG	1
8	9543	start	transcribe	10	0	0	1	0	0	0	1	0	0	VV	0
8	10005	start	go	2	0	1	1	0	0	0	1	0	0	VV	0
8	10726	start	play	4	1	0	0	1	0	1	0	0	0	VV	0
8	10727	start	echo	4	0	0	0	0	1	1	0	0	0	VV	0
8	10730	start	distort	7	1	0	0	0	1	1	0	0	0	VV	0
8	10765	start	hear	4	0	1	0	1	0	1	0	0	0	VV	0
8	11398	start	do	2	0	1	0	1	0	0	1	0	0	VV	0
8	11479	start	make	4	0	1	0	1	0	0	1	0	0	VV	0

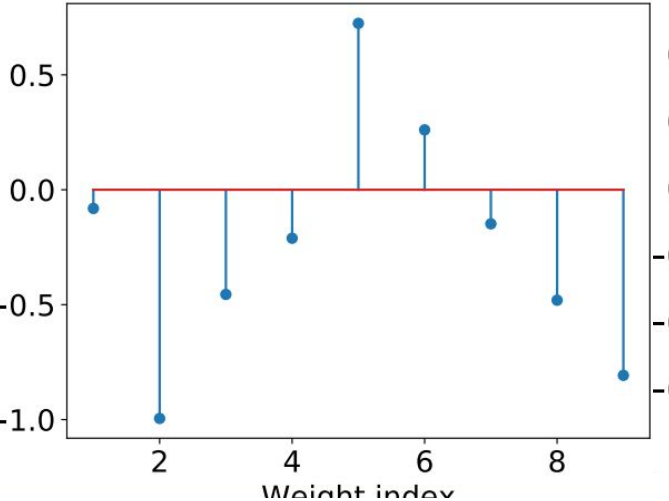
First results: ML Logistic Regression

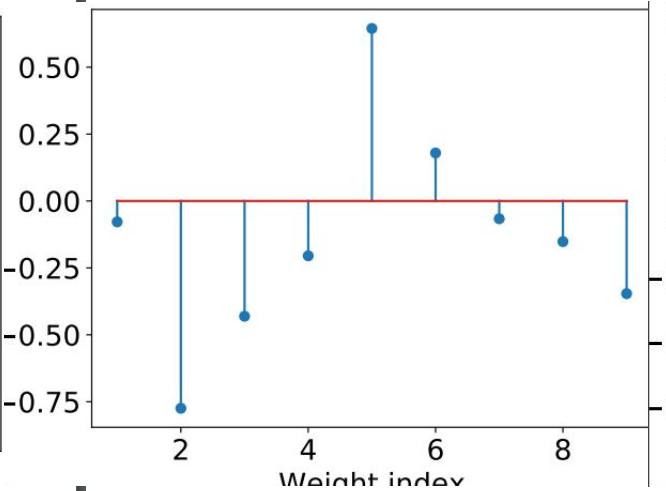


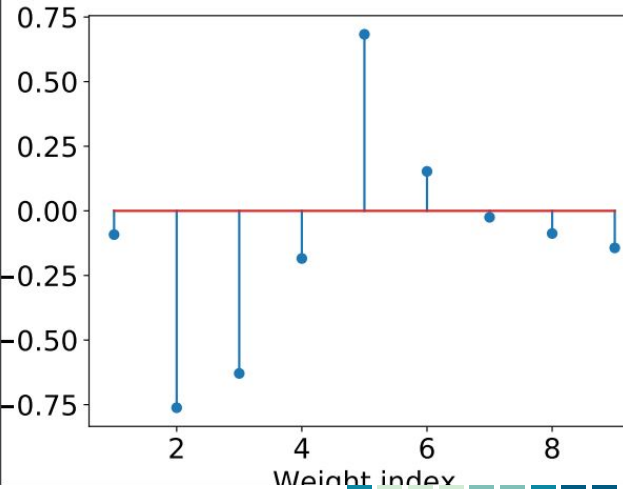


First results: ML Logistic Regression

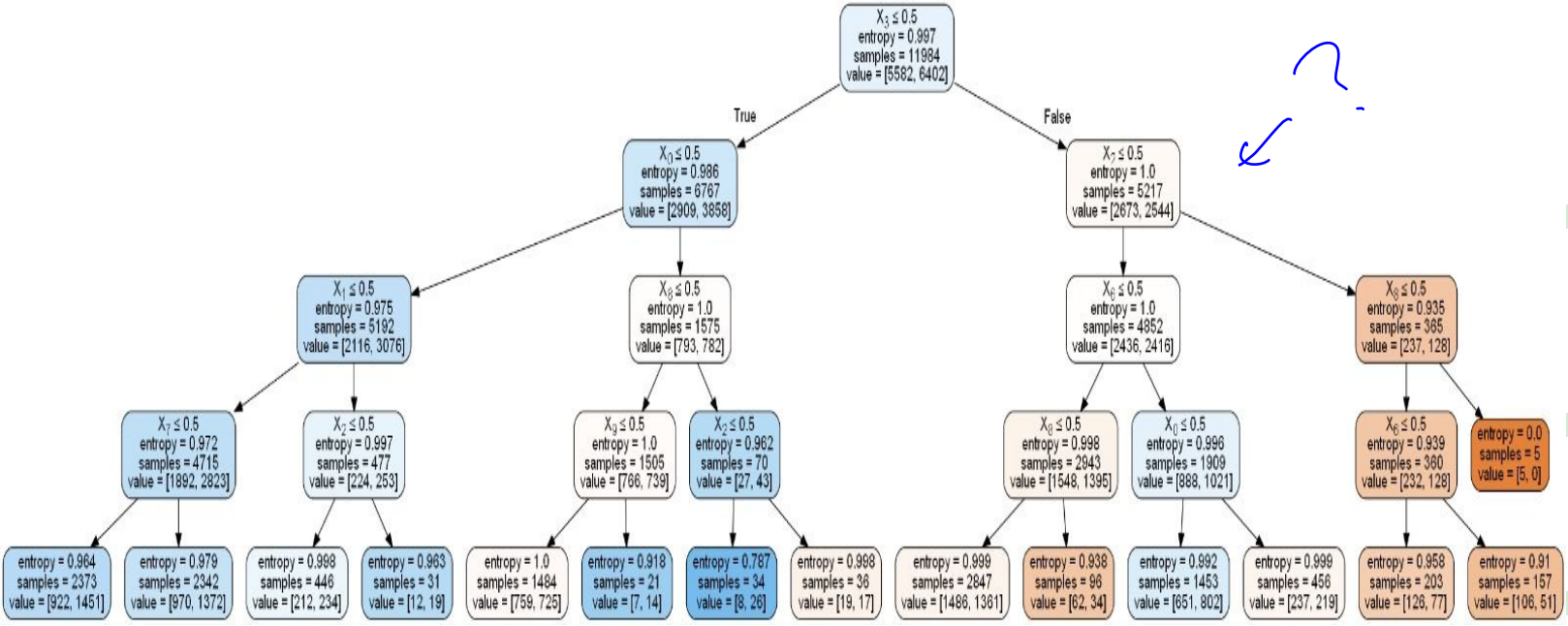
NEITHER THESE



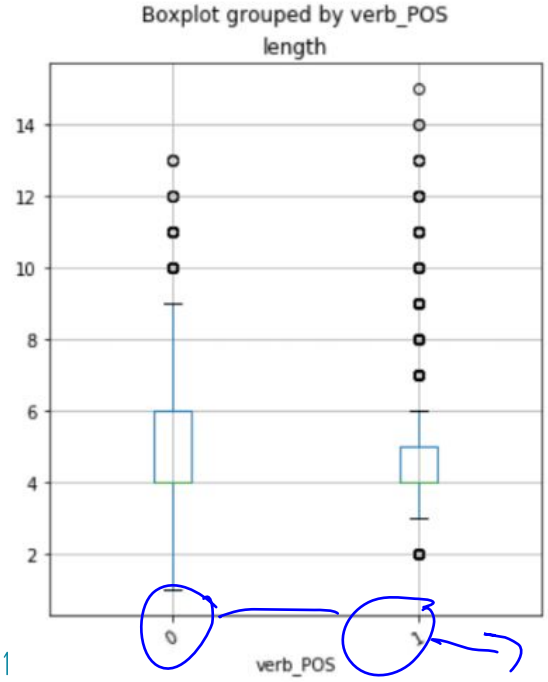


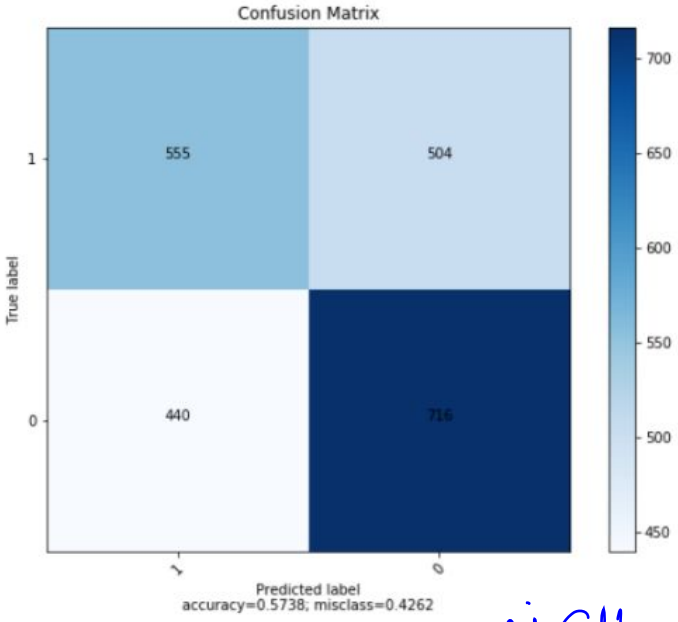


First results: ML decision trees



First results: ML decision trees



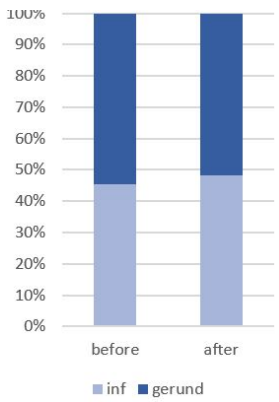


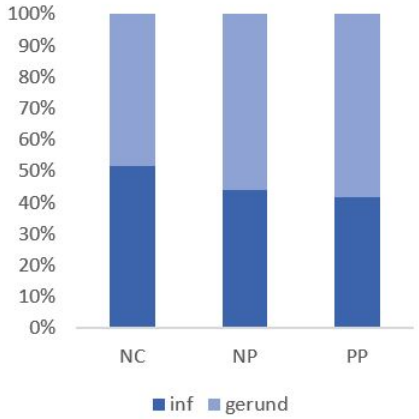
	Feature	Importance
3	argstr_NC	0.371
0	length	0.241
2	befaft_aft	0.172
6	tense_pre	0.084
8	tense_fut	0.066
1	befaft_bef	0.038
7	tense_pst	0.015
9	tense_cond	0.012
4	argstr_NP	0.000
5	argstr_PP	0.000

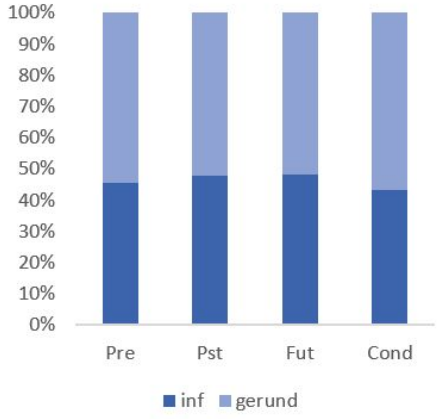
MAKE CLEAR WHICH IS WHICH

Discussions

- Baseline: always predicting gerund, 0.53
- Difference between main verbs
  - *Start* and *hate* provide the most balanced data, other verbs tend to occur with infinitives
- Features may be improved: ✓







Further work

- Apply to more main verbs (*begin*, *like*, etc.)
- Extract more features:
  - Last mention of the non-finite verb
  - Animacy of the subject
  - Semantic class of main verb and non-finite verb
  - Phonological features
- Try more ML methods
  - Neural networks
- **Maybe, the two constructions are just interchangeable all the time.**

INSTEAD, FOR OTHER VERBS IT COULD BE DIFFERENT. FOR INSTANCE, INTUITIVELY I'D SAY THAT "LOVE TO X / LOVE X-ING" ARE NOT USED IN THE SAME CONTEXTS.

ANOTHER QUESTION: WHY IS THE TO-CONSTRUCTION SO MUCH MORE FREQUENT OVERALL?

“ Thank you!

Gràcies!

Teşekkürler!

谢谢！