A new automatic spelling correction model aimed at improving parsability

Rob van der Goot and Gertjan van Noord

Old approach

- IV/OOV
- Generate candidates
- Rank candidates

New approach

- ~~IV/OOV~~
- Generate candidates
- Rank candidates

Data

- LexNorm v1.2
- 549 tweets / 10,576 tokens
- 2,140 OOV tokens
- 1,184 tokens corrected

17

only	IV	only
3mths	OOV	3mths
left	IV	left
in	IV	in
school	IV	school
.	NO	.
i	IV	i
wil	OOV	will
always	IV	always
mis	OOV	miss
my	IV	my
skull	IV	skull
,	NO	,
frnds	OOV	friends
and	IV	and
my	IV	my
teachrs	OOV	teachers

4

new	IV	new
pix	OOV	pictures
comming	OOV	coming
tomoroe	OOV	tomorrow

IV/OOV

- Aspell dictionary
- IV tokens skipped
- 90% of the errors (Bo Han, 2013)

- Example:
  - I am tiret
  - I am tire

IV/OOV

I am tiret

tired
tire

I am tiret

is amy tired
im aim tire

IV/OOV

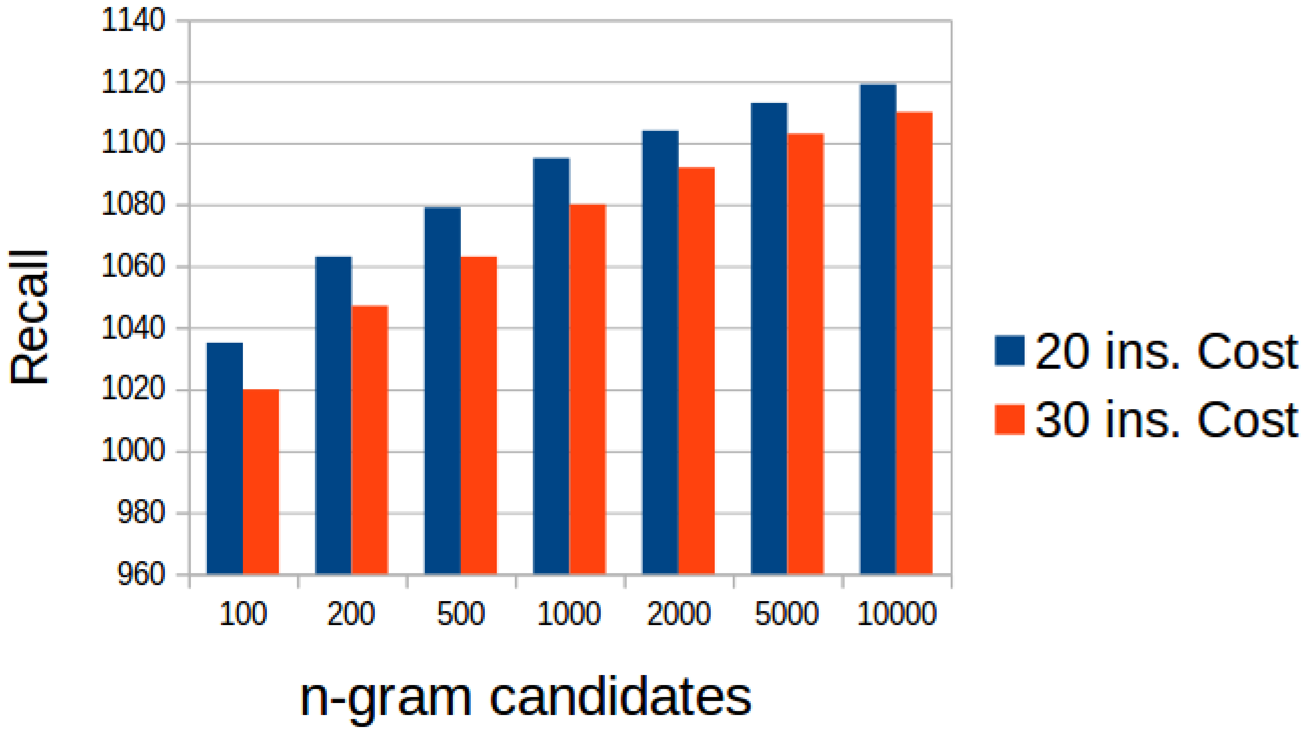
I am tire

I	am	tire
is	amy	tired
im	aim	tire

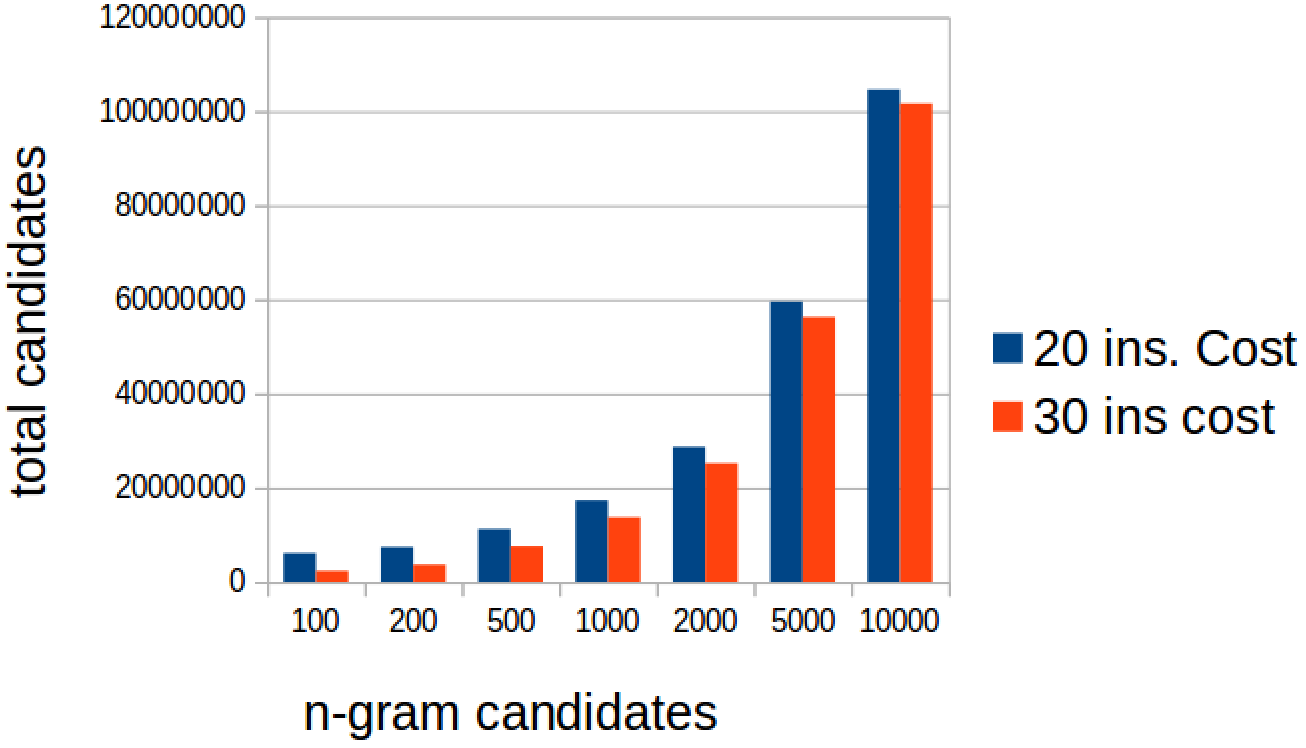
Generate candidates

- Edit distances (Modified Aspell)
- N-grams
- Original token

Generate candidates



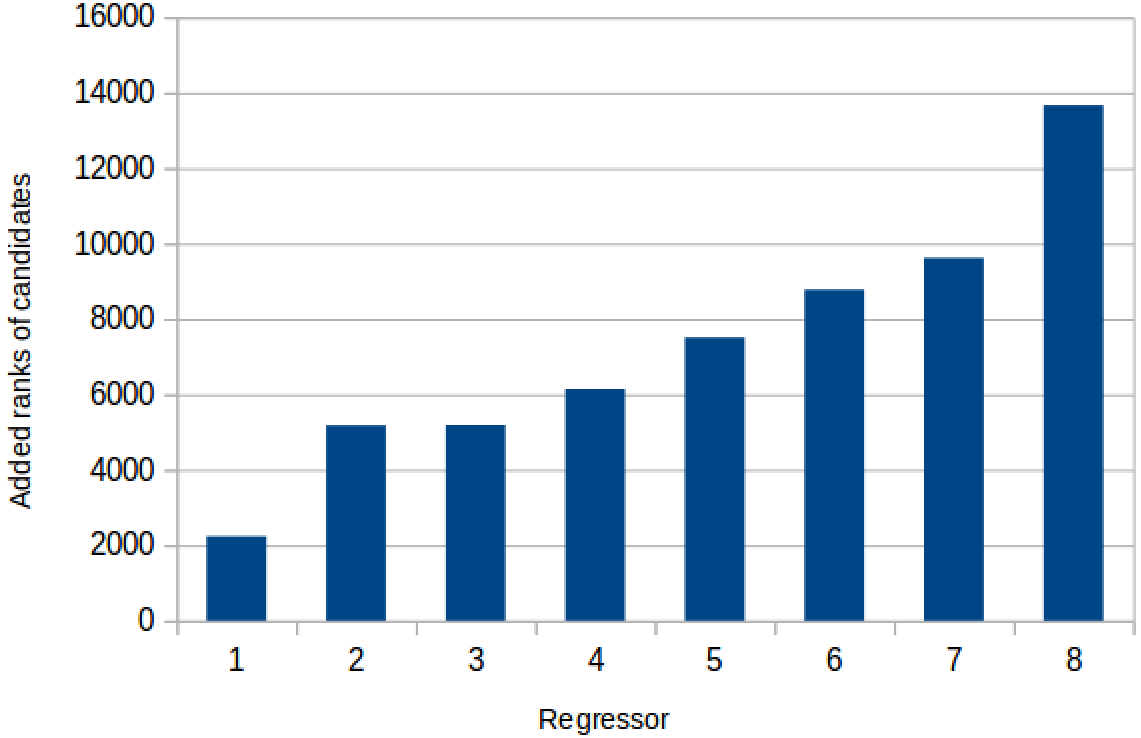
Generate candidates



Rank candidates

- N-grams
- Edit distance
- Occurrence in dictionaries
- Parse probability

Rank candidates



1. Random Forest
2. Coordinate Ascent
3. MART
4. RankBoost
5. RankNet
6. AdaRank
7. LambdaMART
8. ListNet

Rank candidates

- Average 222 candidates

top	Accuracy
1	0.32
5	0.62
10	0.72

(Dis-) Advantages

- Includes IV errors
- More general
- Adaption

- Less efficient
- Training data

Future work

- Rank on sentence level
- Generate different token orders
- Generate multi-word solutions
- New corpus (parses)