STAPLE: Simultaneous Translation And Paraphrase for Language Education

2020 Duolingo Shared Task

Xiao Fei;
Lihong Ji

Content

Task Introduction

Duolingo is a online language learning platform.

One learning exercises is to translate sentences from **English to users native language**.

The accepted translation is **not restricted to only one sentence**.

Task Introduction

Organizer: Duolingo

Year: 2020

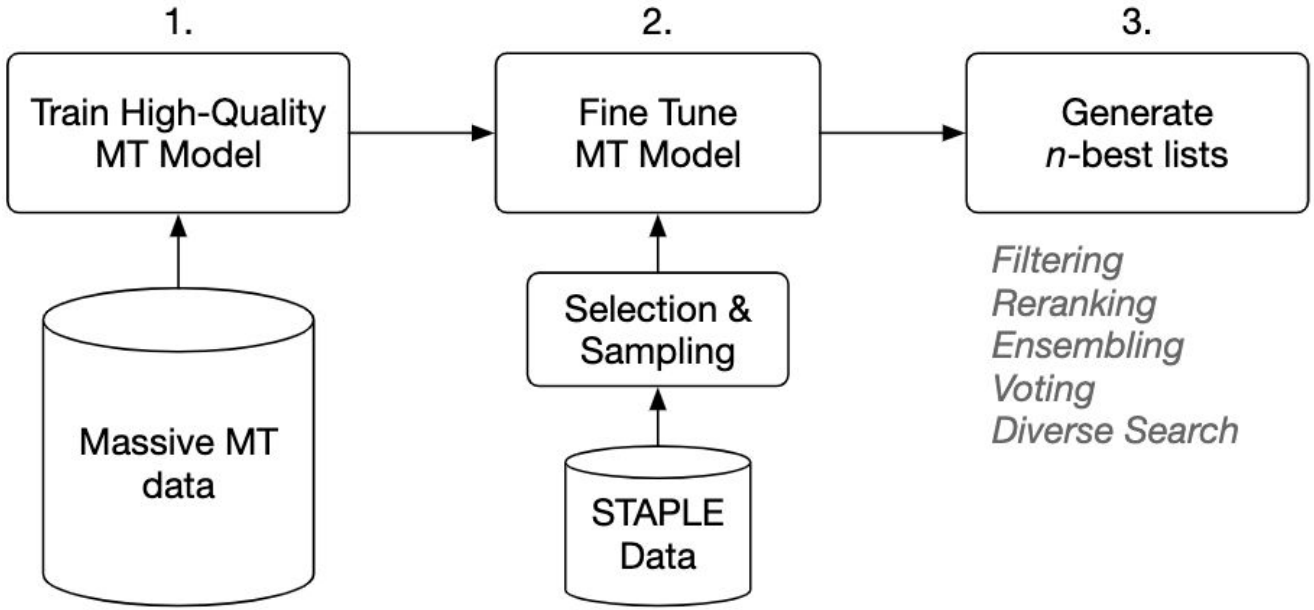
English `->` multilingual target outputs (**high-coverage sets of plausible translations**)

- `en_pt` — **Portuguese**
- `en_hu` — **Hungarian**
- `en_ja` — **Japanese**
- `en_ko` — **Korean**
- `en_vi` — **Vietnamese**

With one English prompt sentence, requires to generate **as many as possible accepted** translations in 5 target languages

Pipeline

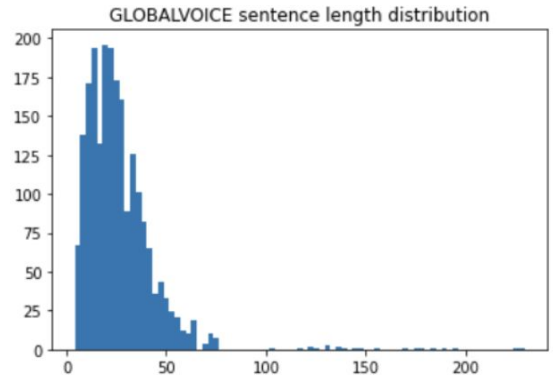
Pretrained transformer frameworks are widely used to deal with translation tasks.

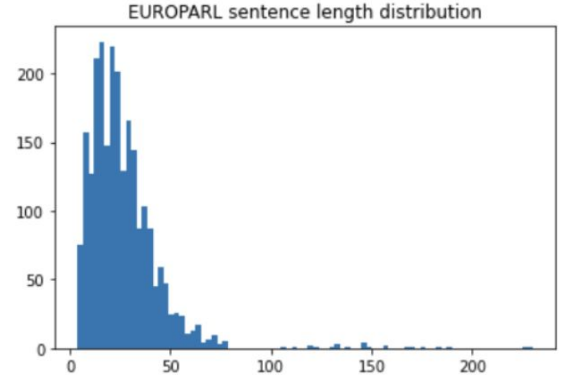


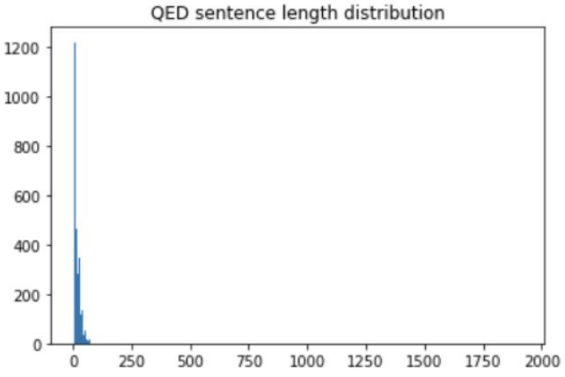
Dataset & Statistical Data

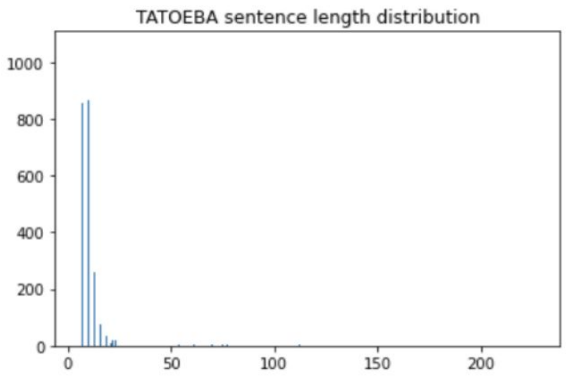
Dataset	doc's	sent's	en tokens	pt tokens
Europarl	10344	2.0M	59.5M	61.1M
QED	4618	0.5M	7.4M	8.7M
Tatoeba	6	0.2M	1.7M	1.7M
GlobalVoices v2018q4	5464	86.6K	2.7M	2.7M

Dataset & Statistical Data



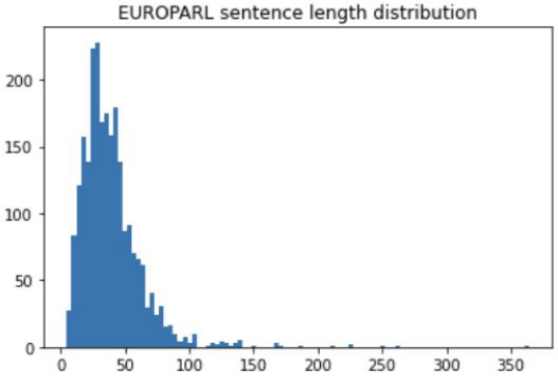


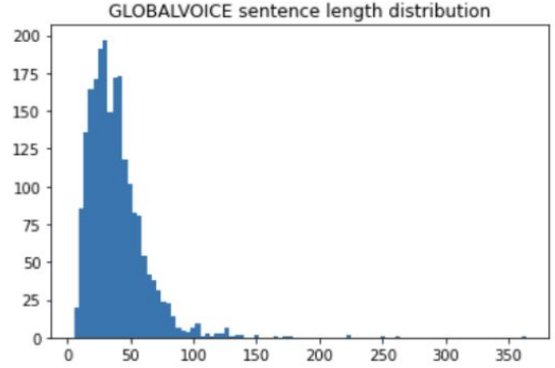


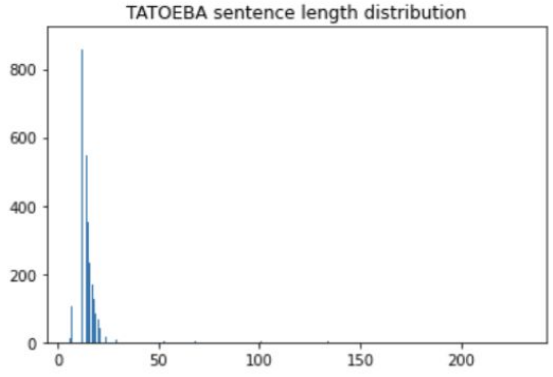


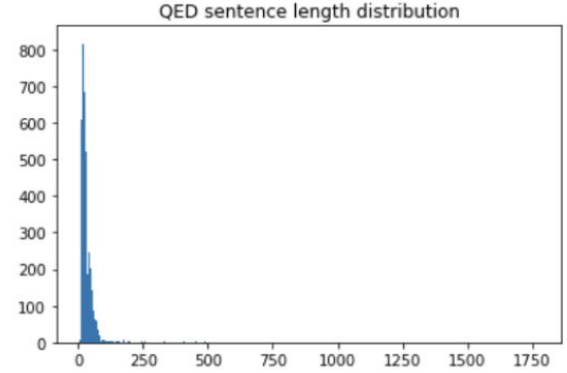
English corpus length distribution

Dataset & Statistical Data









Portuguese corpus length distribution

We set max length to **128**

Dataset & Statistical Data

The data for this task comes from five Duolingo courses.

Each English prompt has many accepted translations and their weights.

It also provide reference translation from Amazon with same prompt.

INPUT: ID	prompt_65c64c31d672de7ed6e084757731dc60	
INPUT: Prompt	**is my explanation clear?**	
INPUT: Reference Translation (from Amazon)	**a minha explicação está clara?**	
OUTPUTS: Accepted Translations (without score) **The weights are the LRF(learner response frequency) which is the frequency of the answer in the translation exercise**	**Translation**: minha explicação está clara? minha explicação é clara? a minha explicação está clara? a minha explicação é clara? minha explanação está clara? está clara minha explicação? minha explanação é clara? a minha explanação está clara? é clara minha explicação? a minha explanação é clara? está clara a minha explicação? …	**Weight**: 0.2673961621319991 0.16168857102956694 0.11109168316077477 0.08778538443694518 0.05717261411615019 0.04428213944745337 0.039226356284927835 0.03634909976199087 0.03634909976199087 0.03134646169709749 0.03134646169709749 …

Dataset & Statistical Data

	Train			Dev			Test		
Language	prompts	trans.	ratio	prompts	trans.	ratio	prompts	trans.	ratio
Hungarian	4,000	251,442	62.9	500	27,647	55.3	500	33,578	67.2
Japanese	2,500	855,941	342.4	500	172,817	345.6	500	165,095	330.2
Korean	2,500	700,410	280.2	500	140,353	280.7	500	150,477	301.0
Portuguese	4,000	526,466	131.6	500	60,294	120.6	500	67,865	135.7
Vietnamese	3,500	194,720	55.6	500	29,637	59.3	500	28,242	56.5

The number of prompts/translations in three split sets for each language pair

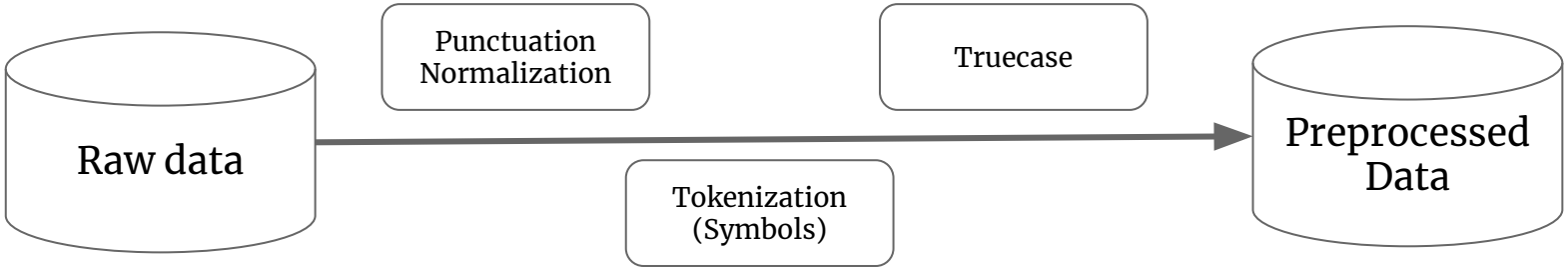
Our approach

Data preprocessing

Mosesdecoder toolkit

。 ---> .

PARIS/paris ---> Paris



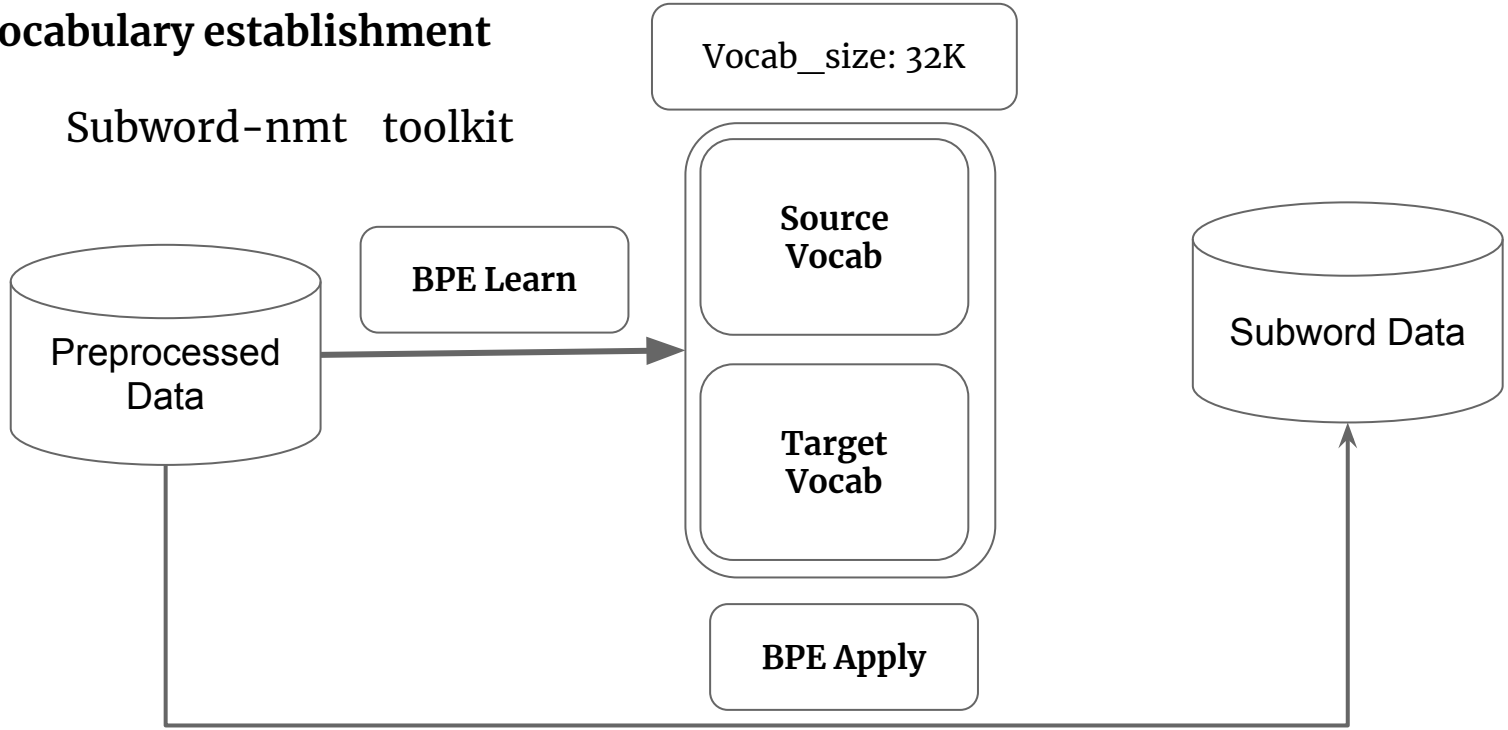
__ ---> _

: ---> 't

Our approach

Vocabulary establishment

Subword-nmt toolkit



Our approach

Duolingo Data

We propose 4 sorts of copying schedules.

Prompt	is my explanation clear?	
Output	minha explicação está clara?	0.267
	minha explicação é clara?	0.161
	a minha explicação está clara?	0.111
	a minha explicação é clara?	0.088
	minha explanação está clara?	0.057
	está clara minha explicação?	0.044
	minha explanação é clara?	0.039

27
16
11
9
6
5
4

1. 5-level division (1/5/10/25/50)
2. Copying based on frequency
3. Copying best 1 based on frequency
4. Copying best 3 based on frequency

Our approach

Output Generation

Beam search & best 10

Our results

Pretrain model

En -> Pt Model Multi-BLEU : about 43.5

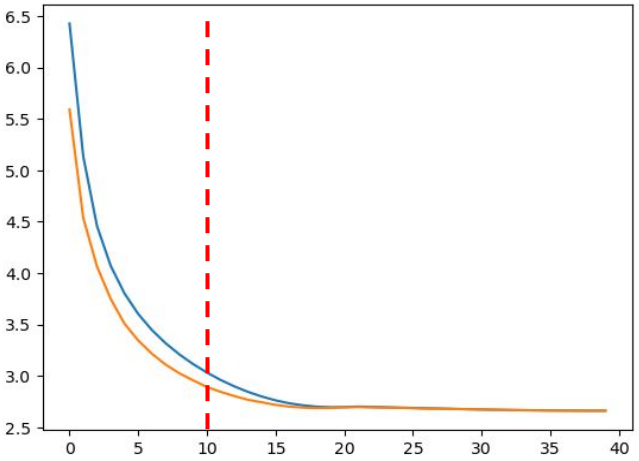
** En -> Ja Model Multi-BLEU : about 21.6 (Lost)

Our results

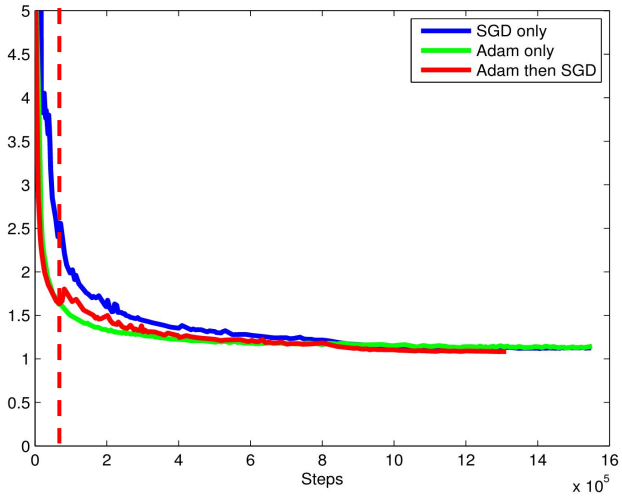
Pretrain model

We conduct google's training schedule AdamW to SGD to achieve better convergence.

Our model



Google NMT



Our results

	weighted macro F1	multi-BLEU	BLEU(n=1/2/3/4)
5-level division(1/5/10/25/50)	**0.4829**	**63.16**	83.6/68.3/57.5/48.4
copying with frequency	0.4710	59.27	79.7/64.8/53.6/44.6
only copying best 1	0.4615	54.81	76.8/61.2/48.9/39.2
copying best 3	0.4605	58.18	79.1/63.7/52.5/43.3
original training set	0.426	60.15	81.2/66.8/55.4/46.7
baseline(no fine-tune)	0.213	NULL	NULL

Our results

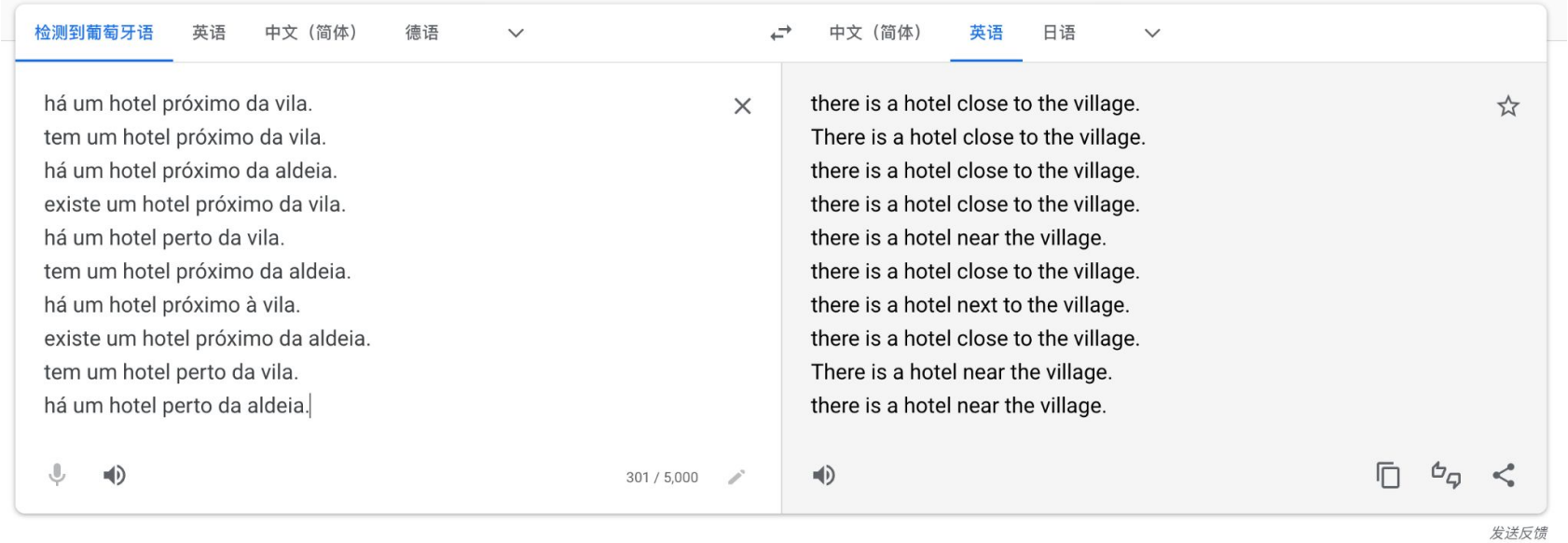
user	hu	ja	ko	pt	vi
jbrem	0.555	0.318	0.404	0.552	0.558
nickeilf	--	--	--	0.551	--
rakchada	0.552	--	--	0.544	--
jspak3	--	--	0.312	--	--
sweagraw	0.469	0.294	0.255	0.525	0.539
masahiro	--	0.283	--	**0.4829**	--
mzy	--	0.260	--		--
dcu	--	--	--	0.460	--
jindra.helcl	0.435	0.213	0.206	0.412	0.377
darkside	--	0.194	--	--	--
nagoudi	--	--	--	0.376	--
baseline_aws	0.281	0.043	0.041	0.213	0.198
baseline_fairseq	0.124	0.033	0.049*	0.136	0.254*

We are here.

Our results

Prompt: there 's a hotel close to the village.

Tested by Google Translation.



Our results

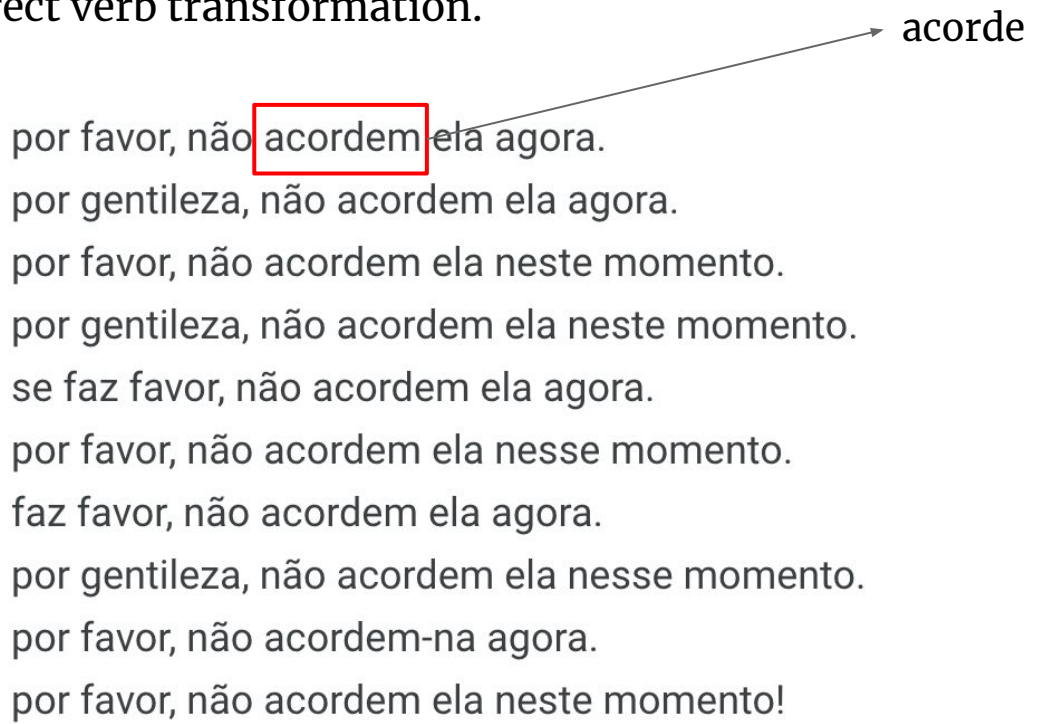
However, there are some problems.

For instance, improper words in this context.

eles gostam um do outro?	do they like each other?
elas gostam um do outro?	do they like each other?
elas gostam umas das outras?	do they like each other?
eles gostam uns dos outros?	do they like each other?
elas gostam uns dos outros?	do they like each other?
elas gostam uma da outra?	do they like each other?
eles estão curtindo um ao outro?	are they enjoying each other?
elas estão curtindo umas às outras?	are they enjoying each other?
elas estão curtindo um ao outro?	are they enjoying each other?
eles curtem um ao outro?	do they like each other?

Our results

Another one, incorrect verb transformation.



Thank you
Q&A

STAPLE: Simultaneous Translation And Paraphrase for Language Education

Xiao Fei; Lihong Ji