LJGG @ CLEF JOKER Task 3: An improved solution joining with dataset from task 1

Leopoldo Jesús Gutiérrez Galeano

UCA Universidad de Cádiz

7 September 2022
Bologna - Italy

JOKER

CLEF 2022 BOLOGNA

Contents

1. Introduction
2. Model
3. Dataset
4. Experiments
5. Manual translations with DeepL
6. Conclusions and Future work

Introduction

Proposed solution for

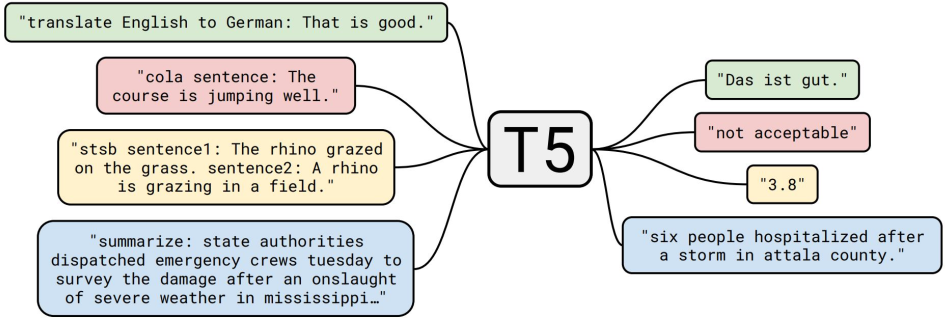
Task 3: Translate entire phrases containing wordplay

Example:

```
{
  "id":"pun_532_2",
  "en":"I phoned the zoo but the lion was busy.",
  "fr":"J'ai appelé le zoo, mais la lionne était occupé."
}
```

Model: T5

- Selected models for experiments: T5 & mT5
- Transformer based architecture
- Sequence to sequence model
- Text-to-Text tasks
- Pre-trained models
- Different sizes
- Prepared for many tasks
  - Translation
  - Classification
  - Summarization
  - Question & Answering
  - And many more…
- Used SimpleT5 for simplicity



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

Dataset: Task 3 data

Used the dataset provided by Project JokeR for Task 3

5114 rows and 3 features:

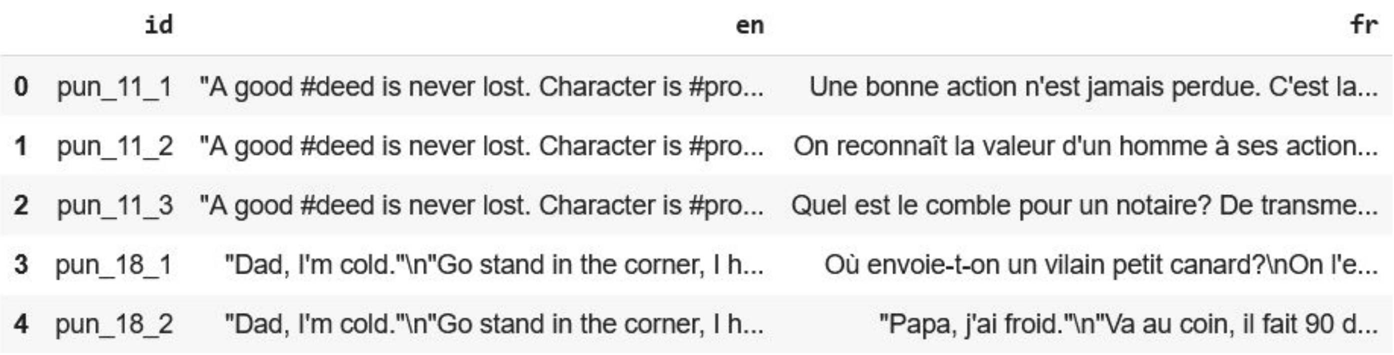
- id
- en
- fr

Example:

```
{
  "id":"pun_532_2",
  "en":"I phoned the zoo but the lion was busy.",
  "fr":"J'ai appelé le zoo, mais la lionne était occupé."
}
```

Dataset: More features?

First 5 rows:



	id	en	fr
0	pun_11_1	"A good #deed is never lost. Character is #pro...	Une bonne action n'est jamais perdue. C'est la...
1	pun_11_2	"A good #deed is never lost. Character is #pro...	On reconnaît la valeur d'un homme à ses action...
2	pun_11_3	"A good #deed is never lost. Character is #pro...	Quel est le comble pour un notaire? De transme...
3	pun_18_1	"Dad, I'm cold."\n"Go stand in the corner, I h...	Où envoie-t-on un vilain petit canard?\nOn l'e...
4	pun_18_2	"Dad, I'm cold."\n"Go stand in the corner, I h...	"Papa, j'ai froid."\n"Va au coin, il fait 90 d...

Is the provided dataset enough for this kind of task?

Dataset: Let’s complete it with more information!

The dataset provided by Project JokeR for Task 1 contains useful information.

10 features:

- ID
- WORDPLAY
- TARGET_WORD
- DISAMBIGUATION
- HORIZONTAL/VERTICAL
- MANIPULATION_TYPE
- MANIPULATION_LEVEL
- CULTURAL_REFERENCE
- CONVENTIONAL_FORM
- OFFENSIVE

Dataset: Adding useful features!

Useful features:

- WORDPLAY
- TARGET_WORD
- DISAMBIGUATION

How do we join the data?

- en == WORDPLAY

More information!

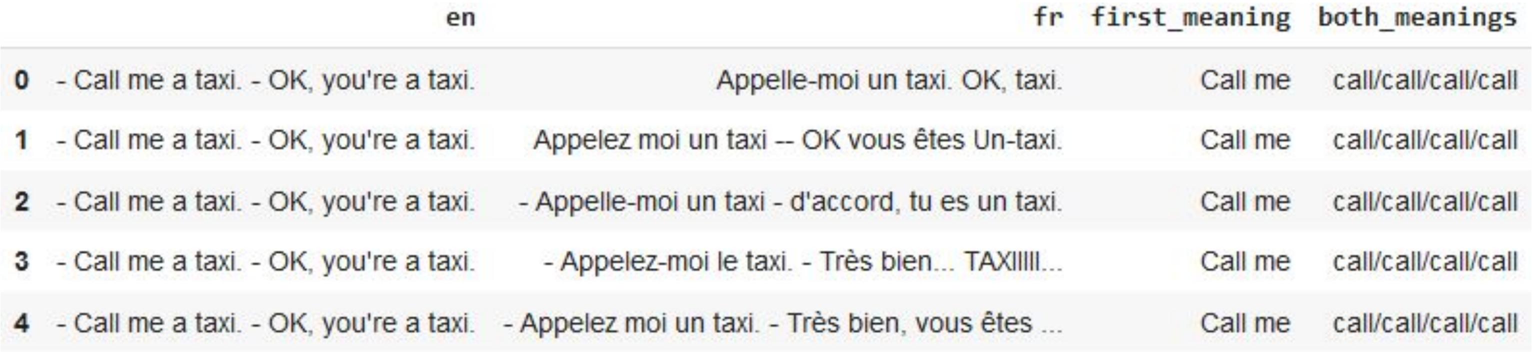
```
{
  "id":"pun_532_2",
  "en":"I phoned the zoo but the lion was busy.",
  "fr":"J'ai appelé le zoo, mais la lionne était occupé."
}

{
  "ID":"pun_532",
  "WORDPLAY":"I phoned the zoo but the lion was busy.",
  "TARGET_WORD":"lion",
  "DISAMBIGUATION":"lion\/line",
  "HORIZONTAL\/VERTICAL":"vertical",
  "MANIPULATION_TYPE":"Similarity",
  "MANIPULATION_LEVEL":"Sound",
  "CULTURAL_REFERENCE":false,
  "CONVENTIONAL_FORM":false,
  "OFFENSIVE":null
}
```

Dataset: Joined datasets

Our new dataset:

- en/WORDPLAY → en
- fr
- TARGET_WORD → first_meaning
- DISAMBIGUATION → both_meanings

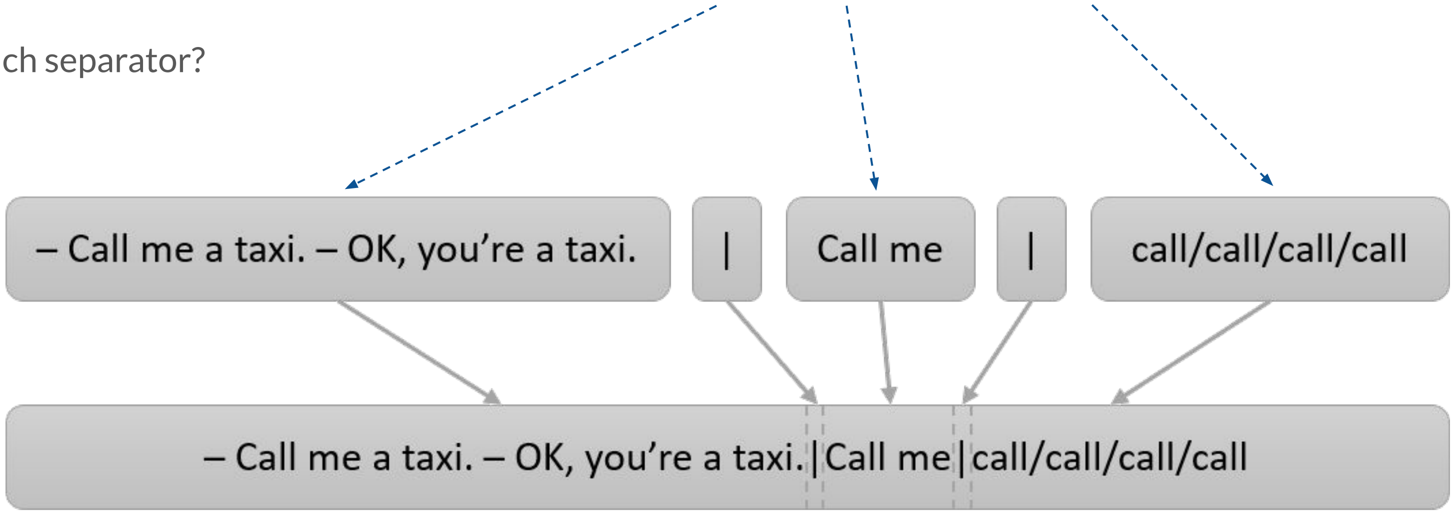


	en	fr	first_meaning	both_meanings
0	- Call me a taxi. - OK, you're a taxi.	Appelle-moi un taxi. OK, taxi.	Call me	call/call/call/call
1	- Call me a taxi. - OK, you're a taxi.	Appelez moi un taxi -- OK vous êtes Un-taxi.	Call me	call/call/call/call
2	- Call me a taxi. - OK, you're a taxi.	- Appelle-moi un taxi - d'accord, tu es un taxi.	Call me	call/call/call/call
3	- Call me a taxi. - OK, you're a taxi.	- Appelez-moi le taxi. - Très bien... TAXIIIII...	Call me	call/call/call/call
4	- Call me a taxi. - OK, you're a taxi.	- Appelez moi un taxi. - Très bien, vous êtes ...	Call me	call/call/call/call

Approach: How do we pass data to T5?

The input requires a string, so we will concatenate **en**, **first_meaning**, **both_meanings** with a separator.

Which separator?


– Call me a taxi. – OK, you’re a taxi.
|
Call me
|
call/call/call/call
– Call me a taxi. – OK, you’re a taxi.|Call me|call/call/call/call

Hardware

Specs:

- GPU Nvidia Quadro P6000
- 24GB of GPU memory
- 30GB of RAM
- 8 vCPUs

Paperspace

Software

- Jupyter Notebook
- SimpleT5
- Pandas
- sacreBLEU

Experiments

Why more data?

2 experiments:

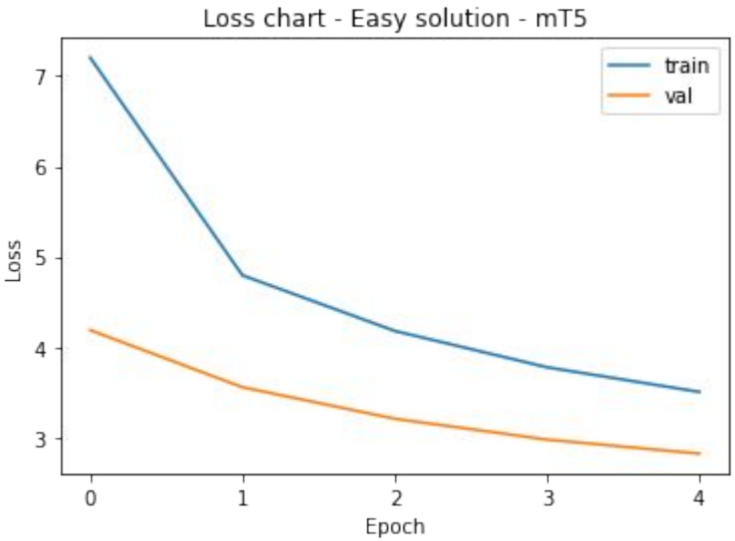
- Single fine-tuning, with the Task 3 dataset.
- Proposed architecture, adding more information from the Task 1 dataset.

Previous experiment

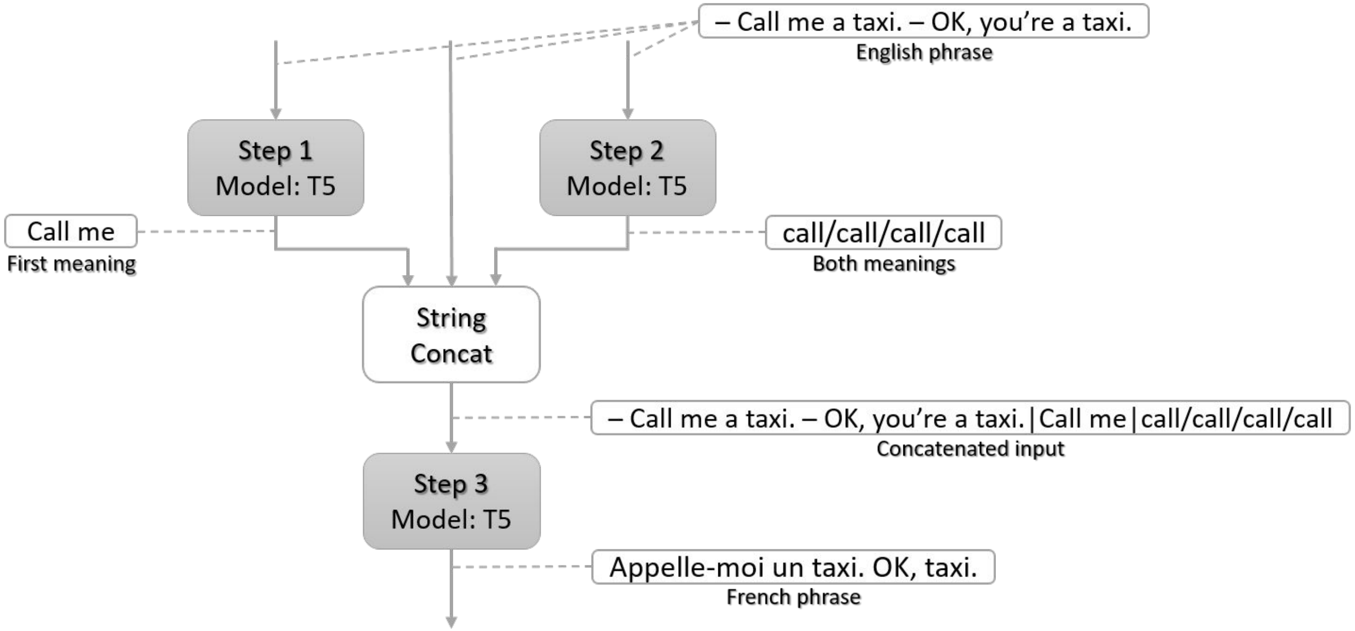
- Basic solution
- Single model
- Fine-tuning

Step 3 (mT5-base):

- Best epoch: 4
- Accuracy: 0%
- BLEU Score: 4,88 → almost useless



Main experiment: Architecture



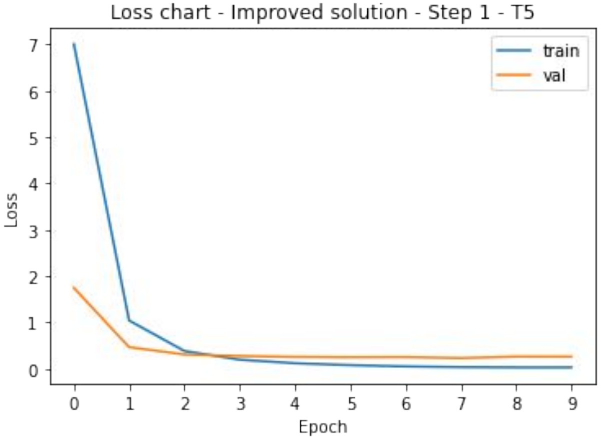
Main experiment: Step 1 & Step 2 Models Loss Charts

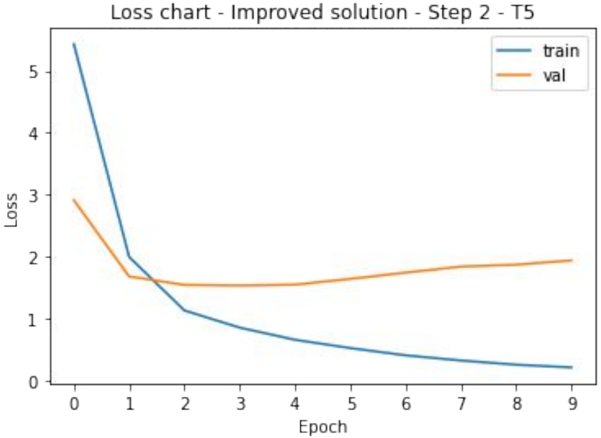
Step 1 (t5-large):

- Best epoch: 7
- Accuracy: 79,36%

Step 2 (t5-large):

- Best epoch: 3
- Accuracy: 19,04%





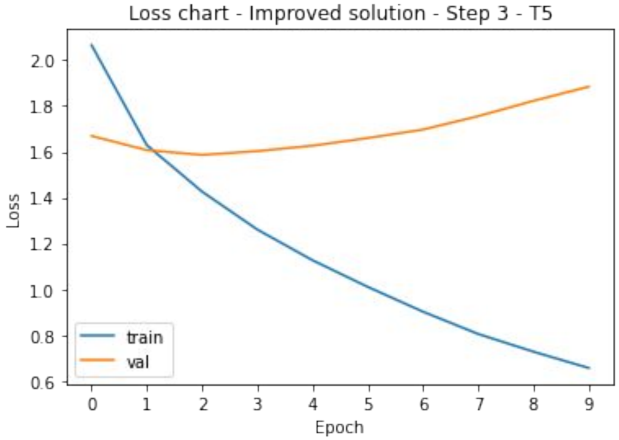
Main experiment: Step 3 Model, T5 or mT5?

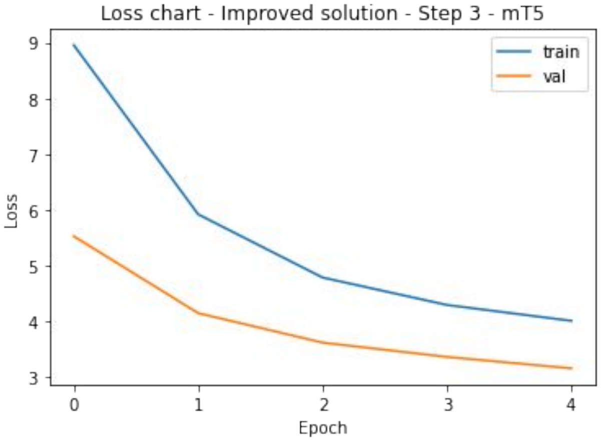
Step 3 (t5-large):

- Best epoch: 2
- Accuracy: 0,17%
- BLEU Score: 19,99 → hard to get the gist / the gist is clear but has significant errors

Step 3 (mT5-base):

- Best epoch: 4
- Accuracy: 0%
- BLEU Score: 3,14 → almost useless

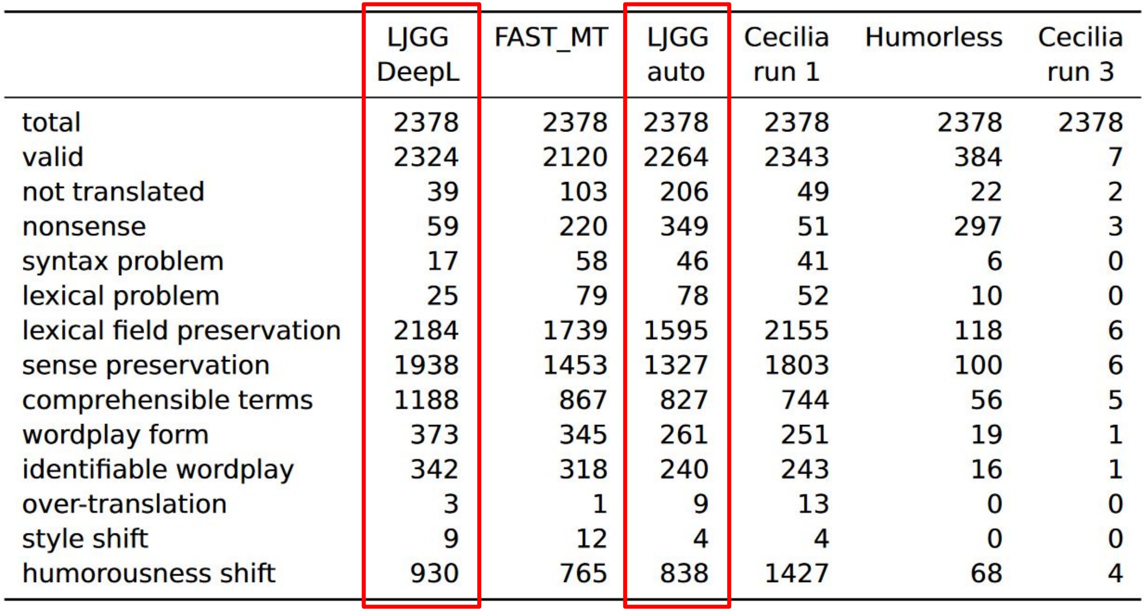




Manual translations with DeepL

- Free translator, with limitations without susbcription
- Custom architecture based on transformers
- Some parts contain attention mechanisms
- Free version: up to 5000 characters (~70/80 phrases)
- Better results than proposed architecture

Analysis of results



	#1		#3			
	LJGG DeepL	FAST_MT	LJGG auto	Cecilia run 1	Humorless	Cecilia run 3
total	2378	2378	2378	2378	2378	2378
valid	2324	2120	2264	2343	384	7
not translated	39	103	206	49	22	2
nonsense	59	220	349	51	297	3
syntax problem	17	58	46	41	6	0
lexical problem	25	79	78	52	10	0
lexical field preservation	2184	1739	1595	2155	118	6
sense preservation	1938	1453	1327	1803	100	6
comprehensible terms	1188	867	827	744	56	5
wordplay form	373	345	261	251	19	1
identifiable wordplay	342	318	240	243	16	1
over-translation	3	1	9	13	0	0
style shift	9	12	4	4	0	0
humorousness shift	930	765	838	1427	68	4

Conclusions & Future work

Conclusions

- Better results with DeepL instead of custom architecture
- Custom architecture got was ranked in 3rd position

Future work

- Swapping the concatenated information
- Adding more data
- Everything in lowercase
- Maybe separating step 3 in 2 models: one for puns and the other for wordplays
- Different model
- Playing with hyperparameters
- Direct implementation with PyTorch

Thank you for your attention

LJGG @ CLEF JOKER Task 3: An improved solution joining with dataset from task 1

Leopoldo Jesús Gutiérrez Galeano

UCA Universidad de Cádiz

7 September 2022
Bologna - Italy

JOKER

CLEF 2022 BOLOGNA