Fill-Mask Task

Use HuggingFace pipelines to replicate tables 3 and 4 from [this paper](). Do this three times with three different models. At least one of the three models should be BERT.

Colab link

https://colab.research.google.com/drive/1MB5AvOyYpDPzjwGLJY2EEyTlozUznsEf?usp=sharing

Table 3. Unmasking: replace each input word with [MASK] and predict fillers. Red is added to highlight differences between the top prediction and the original input

Word	Rank 0	Rank 1	Rank 2
This	this:0.595	it:0.375	there:0.004
is	was:0.628	is:0.334	included:0.007
a	a:0.935	another:0.029	the:0.023
test	part:0.253	subset:0.125	variation:0.082
of	of:0.886	for:0.070	on:0.017
the	the:0.883	an:0.088	our:0.005
emergency	ratio:0.257	television:0.048	satellite:0.031
broadcast	braking:0.620	response:0.070	management:0.024
system	system:0.244	technique:0.077	capability:0.059
.	.:0.969	;:0.029	!:0.001

[Table 3] Model 1: `bert-base-cased`

https://huggingface.co/bert-base-cased
Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. **This model is case-sensitive: it makes a difference between english and English.**

Word	Rank 0	Rank 1	Rank 2
This	It: 0.576	This: 0.368	it: 0.01
is	was: 0.760	is: 0.177	included: 0.005
a	a: 0.878	the: 0.092	another: 0.018
test	part: 0.128	version:0.084	feature: 0.081
of	of: 0.911	for: 0.039	on: 0.016
the	the: 0.544	an: 0.414	any: 0.005
emergency	radio:0.085	FM: 0.073	digital: 0.041
broadcast	response: 0.303	medical: 0.050	management: 0.049
system	system: 0.095	capability: 0.047	power: 0.047
.	.:0.988	;:0.11	!:0.0003

[Table 3] Model 2: `xlm-roberta-base`

https://huggingface.co/xlm-roberta-base
XLM-RoBERTa (base-sized) model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. It was introduced in the paper Unsupervised Cross-lingual Representation Learning at Scale by Conneau et al. and first released in this repository.

Word	Rank 0	Rank 1	Rank 2
This	This: 0.650	Here: 0.238	It: 0.043
is	is: 0.883	was: 0.100	show: 0.005
a	a: 0.792	the: 0.182	another: 0.008
test	screenshot: 0.191	video:0.131	version: 0.066
of	of: 0.751	for: 0.098	with: 0.044
the	an: 0.603	the: 0.231	our: 0.062
emergency	current:0.046	new: 0.042	radio: 0.034
broadcast	response: 0.356	management: 0.136	room: 0.034
system	system: 0.207	ing: 0.082	er: 0.069
.	.:0.804	::0.100	</s>:0.041

[Table 3] Model 3: `bert-base-uncased`

https://huggingface.co/bert-base-uncased

Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. **This model is uncased: it does not make a difference between english and English**.

Word	Rank 0	Rank 1	Rank 2
This	this: 0.595	it: 0.375	there: 0.004
is	was: 0.628	is: 0.334	included: 0.007
a	a: 0.935	another: 0.029	the: 0.023
test	part: 0.253	subset:0.125	variation: 0.082
of	of: 0.886	for: 0.070	on: 0.017
the	the: 0.883	an: 0.088	out: 0.005
emergency	ratio: 0.257	television: 0.048	satellite: 0.031
broadcast	braking: 0.620	response: 0.070	management: 0.024
system	system: 0.244	technique: 0.077	capability: 0.059
.	.:0.969	;:0.029	!:0.001

[Table 4] Model 1: `bert-base-cased`

https://huggingface.co/bert-base-cased
Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. **This model is case-sensitive: it makes a difference between english and English.**

MWEs	**are**	**a**	**pain**	**in**	**the**	**neck**	**for**	**NLP**	**.**
There:0.5	cause:0.3	a:0.4	hole:0.2	in:1.0	the:1.0	ass:1.0	of:0.3	children:0.1	.:1.0
		better	**safe**	**than**	**sorry**	**.**			
		better:1.0	safe:1.0	than:0.9	sorry:0.9	.:0.9			
		read	**between**	**the**	**lines**	**.**			
		in: 0.1	between:0.8	the:0.9	lines:0.6	.:0.8			
		play	**your**	**cards**	**right**	**.**			
		Play:0.8	your:0.7	cards:0.7	right:0.1	.:0.6			
		it's	**an**	**uphill**	**battle**	**.**			
		What:0.3	an:1.0	endless:0.2	hike:0.2	.:0.9			
you	**can't**	**judge**	**a**	**book**	**by**	**its**	**cover**	**.**	
You:0.7	can:0.5	judge:0.5	a:0.6	book:0.8	by:1.0	its:0.8	title:0.2	.:0.9	
		ignorance	**is**	**bliss**	**.**				
		This:0.4	is:0.4	essential:0.04	.:0.9				
the	**grass**	**is**	**always**	**greener**	**on**	**the**	**other**	**side**	**.**
The:0.9	grass:0.1	was:0.7	much:0.6	green:0.04	on:0.9	the:1.0	other:0.2	side:0.9	.:0.8

[Table 4] Model 2: `albert-base-v2`

https://huggingface.co/albert-base-v2

Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. **This model, as all ALBERT models, is uncased: it does not make a difference between english and English**.

MWEs	**are**	**a**	**pain**	**in**	**the**	**neck**	**for**	**NLP**	**.**
They:0.1	is:0.1	a:0.2	hole:0.07	in:0.9	the:1.0	neck:0.4	joyah:0.2	!!:0.3	.:0.5
		better	**safe**	**than**	**sorry**	**.**			
		better:0.6	done:0.3	than:1.0	sorry:0.1	!:0.5			
		read	**between**	**the**	**lines**	**.**			
		evalle: 0.1	between:0.2	the:0.1	stars:0.06	!:0.4			
		play	**your**	**cards**	**right**	**.**			
		write:0.2	your:0.8	game:0.05	now:0.1	?:0.7			
		it's	**an**	**uphill**	**battle**	**.**			
		what:0.2	an:1.0	ongoing:0.1	ride:0.1	.:0.4			
you	**can't**	**judge**	**a**	**book**	**by**	**its**	**cover**	**.**	
You:0.8	cannot:0.6	judge:0.6	a:0.6	book:0.2	by:0.8	its:1.0	characters:0.05	.:0.5	
		ignorance	**is**	**bliss**	**.**				
		evalle:0.07	constitutes:0.1	ignorance:0.05	!:0.5				
the	**grass**	**is**	**always**	**greener**	**on**	**the**	**other**	**side**	**.**
the:0.3	grass:0.3	was:0.5	much:0.2	growing:0.08	on:0.6	the:0.9	grassy:0.1	side:0.06	your: 0.01

[Table 4] Model 3: `bert-base-uncased`

https://huggingface.co/bert-base-uncased

Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. **This model is uncased: it does not make a difference between english and English**.

MWEs	**are**	**a**	**pain**	**in**	**the**	**neck**	**for**	**NLP**	**.**
there:0.5	is:0.4	a:0.4	pain:0.2	in:1.0	the:1.0	ass:0.9	in:0.6	me:0.1	.:1.0
		better	**safe**	**than**	**sorry**	**.**			
		better:1.0	safe:1.0	than:1.0	sorry:1.0	.:1.0			
		read	**between**	**the**	**lines**	**.**			
		in: 0.1	between:0.8	the:0.9	lines:0.9	.:1.0			
		play	**your**	**cards**	**right**	**.**			
		play:0.8	the:0.5	cards:0.8	right:0.4	.:1.0			
		it's	**an**	**uphill**	**battle**	**.**			
		in:0.3	an:1.0	epic:0.5	battle:0.4	.:1.0			
you	**can't**	**judge**	**a**	**book**	**by**	**its**	**cover**	**.**	
you:0.8	can:0.6	judge:0.8	a:0.8	book:0.8	by:1.0	its:0.9	quality:0.1	.:0.9	
		ignorance	**is**	**bliss**	**.**				
		this:04	is:0.4	bliss:0.4.	.:1.0				
the	**grass**	**is**	**always**	**greener**	**on**	**the**	**other**	**side**	**.**
the:0.3	land:0.2	was:0.8	much:0.6	planted:0.1	on:0.9	the:1.0	other:0.2	side:8	.:0.9

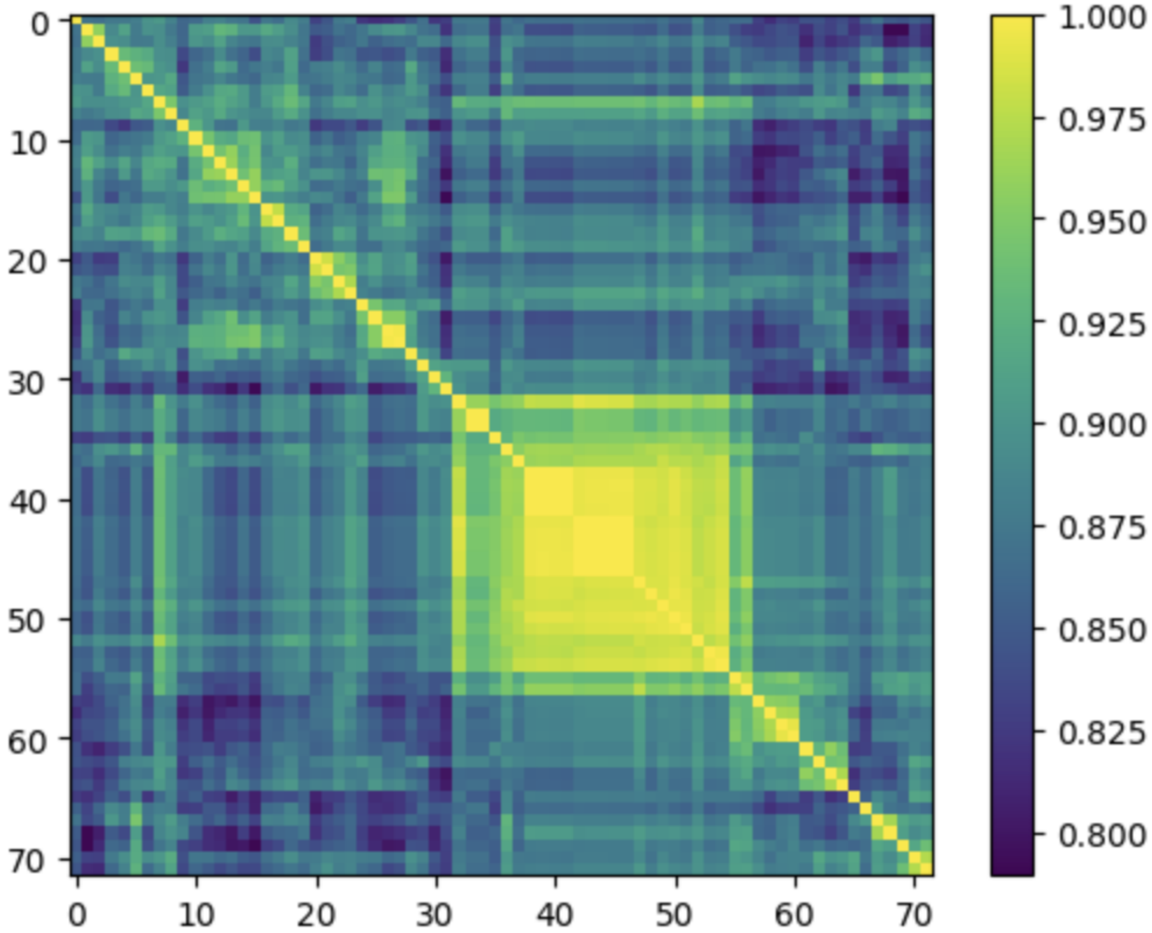
Translation and Feature Extraction

Colab link

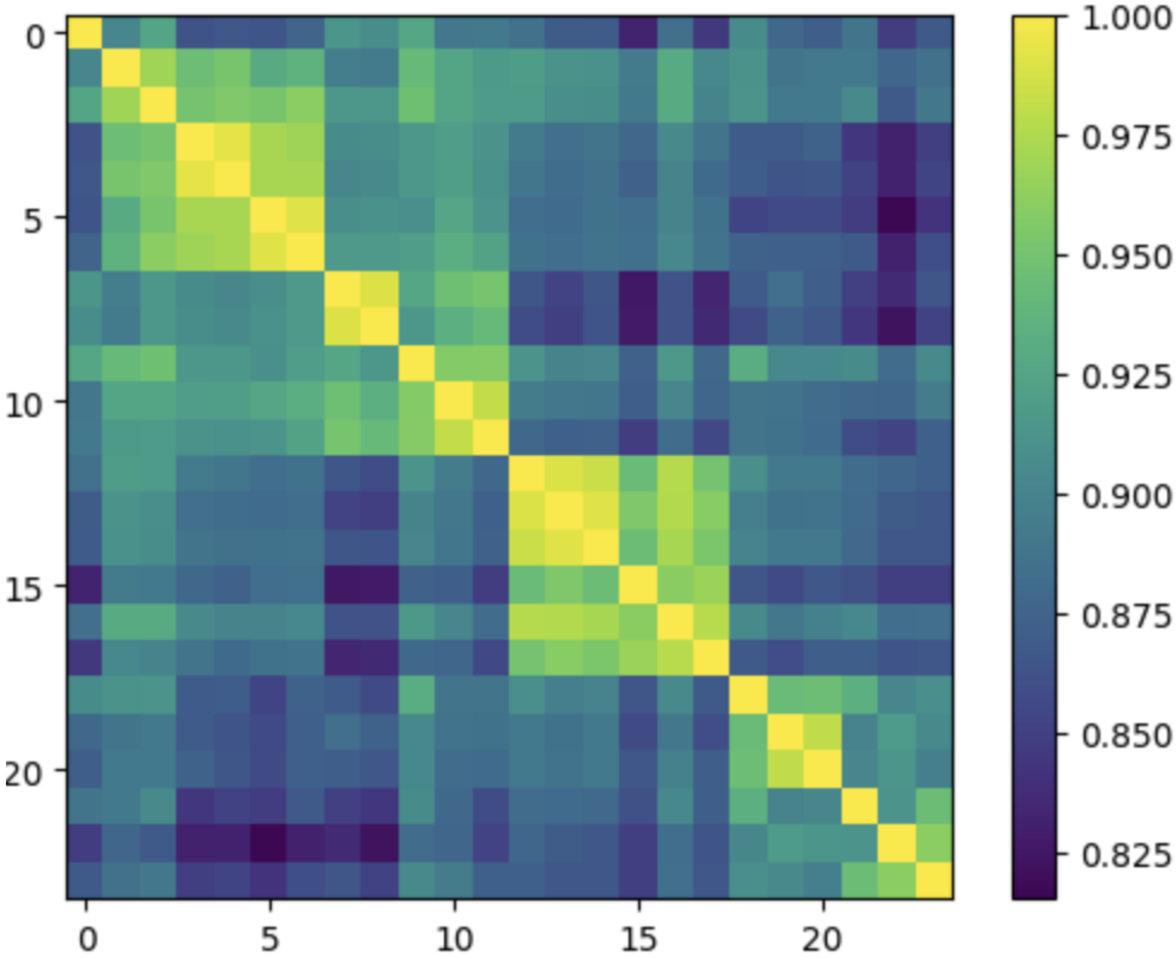
https://colab.research.google.com/drive/1tx-82Pxplh-cYFhxEOLpD__sLtcZPyzV?usp=sharing

Here’s the cosine similarities of the vectors with imshow and boxplot.

Translations between English and Chinese.



Translations between English and French.



My thinking about the open question

Both graphs show that **the median cosine similarity for the translation vectors is both around 0.85**. This indicates that the translation vectors are generally similar to each other.

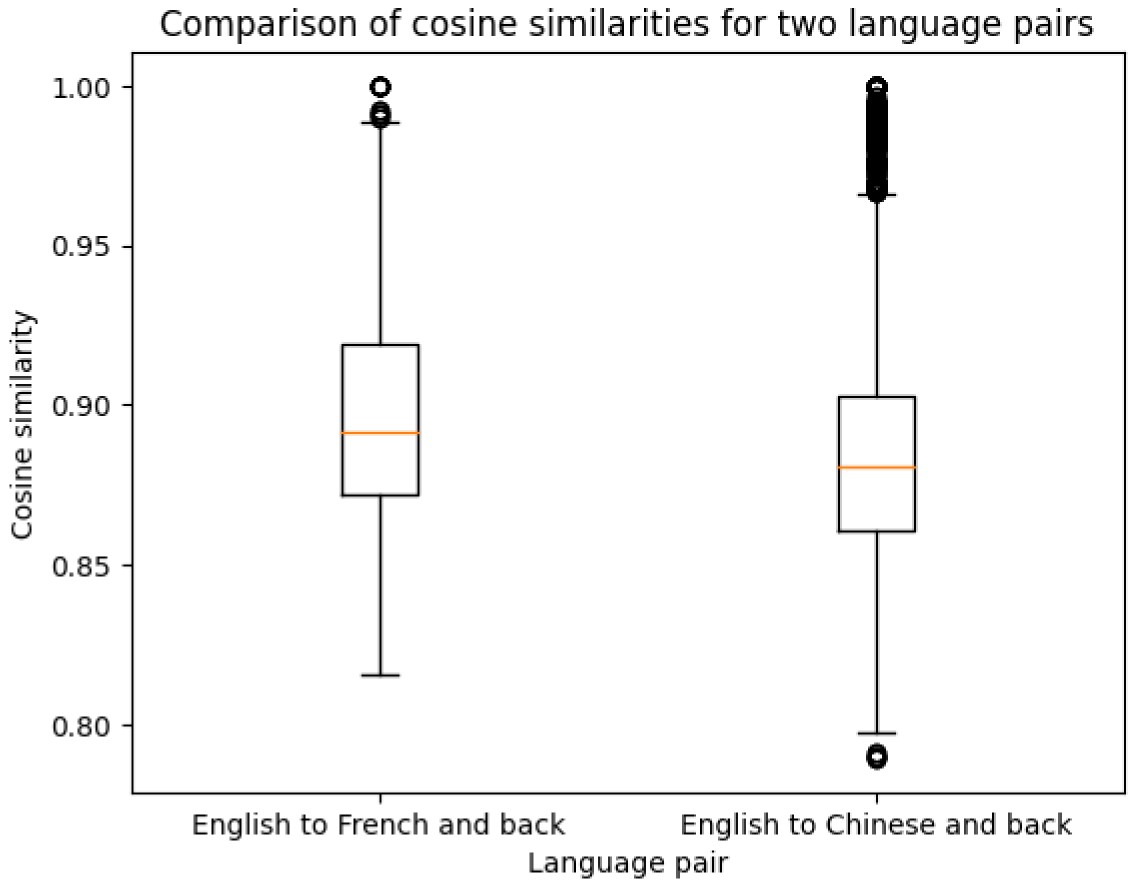
However, **there is a wider range of cosine similarity values in the graph for [translations between English and Chinese]**. This suggests that there is more variation in the accuracy of the translations for this language pair.

Another difference between the two graphs is that **the graph for [translations between English and Chinese] has more outliers**. This indicates that there are a few translation vectors for this language pair that are very different from the other vectors. This could be due to a variety of factors, such as

- The translation model is not very accurate at translating some sentences.

- The translation model is not very consistent, meaning that the same sentence may be translated differently depending on the context.
- The translation model has been trained on a dataset that is not representative of the types of sentences that it will be used to translate.
- There are errors in the translation model.

To solve this problem and make it more precise, we can try to use a different translation model or use the translation model with caution.



From this comparison, it appears that translation does have an impact on the vectors. The cosine similarities between the English vectors and the French vectors are generally higher than the cosine similarities between the English vectors and the Chinese vectors. This suggests that the French translations are more similar to the English vectors than the Chinese translations.

However, I think the quality of a translation also depends on the translation model used and the difficulty of the sentence being translated. Therefore, it might be impossible to say definitively whether the French translations are better or worse than the Chinese translations for comparing papers without more information.

Apart from that, since the task aims to translate titles and abstracts of papers, I think we should also consider the domain knowledge and cultural context when comparing translations of scientific papers.

To make the question more precise, we could probably add more information like "What translation model was used?", "What is the difficulty of the sentences being translated?", "Does the French translation model produce more accurate translations of scientific papers than the Chinese translation model?", "Does the French translation model produce more fluent and natural-sounding translations of scientific papers than the Chinese translation model?" and "Does the French translation model preserve the meaning of scientific papers better than the Chinese translation model?"