Beyond Translation: NLP for Cross-Lingual Communication

Marine Carpuat

Computer Science

University of Maryland

INRIA ALMAnaCH seminar, October 11 2024

Rethinking machine translation in the age of large language models

99.97%

of machine translation users
are not translators.

Nurminen, Mary (2021) ‘Investigating the Influence of Context in the Use and Reception of Raw Machine Translation’
Tampere University Dissertation 515/2021.

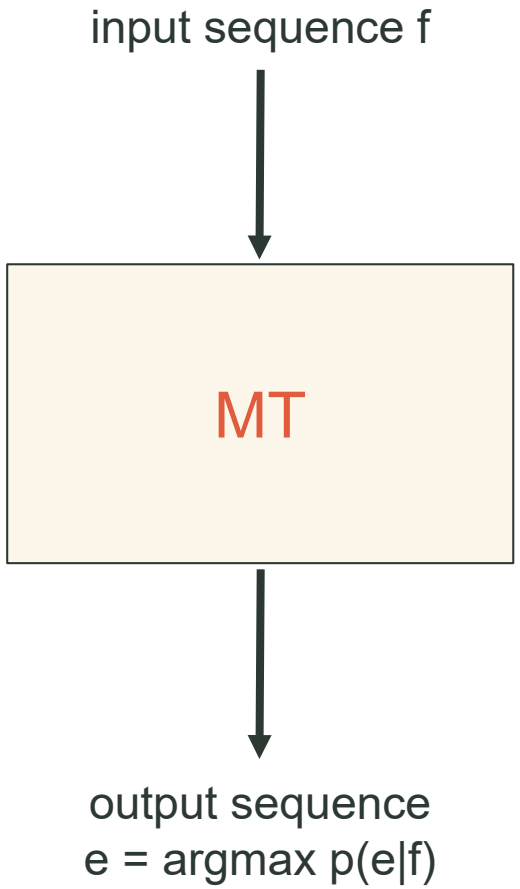
Who uses MT? What for?

Survey of 1200 UK residents found that often used MT for non-essential purposes that do not justify professional translations, but also expressed desire for higher MT quality [Nunes Vieira et al. 2022]

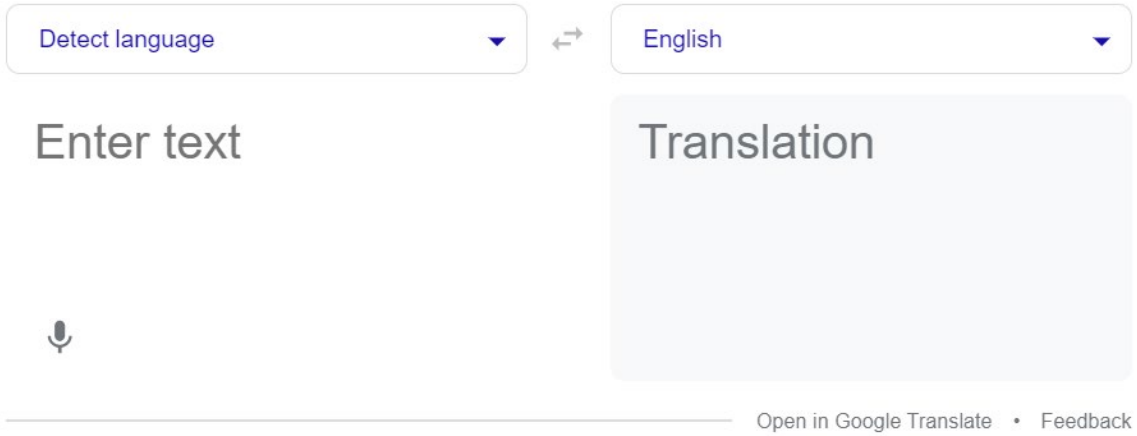
Migrant workers when seeking employment, navigating employer relations, or looking for healthcare information [Liebling et al. 2021]

In France, 85% of 13-21 year old surveyed used genAI in the past 6 months, primarily for summarization, generation, translation [bornAI survey, June 2024]

Rethinking MT: Beyond sequence transduction?



Rethinking MT: Beyond 2 box interfaces?



Beyond Translation: NLP for Cross-Lingual Communication

Part I Reliability in MT-mediated communication

Part II MT as controllable language generation

Users often cannot assess risk in MT outputs

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

Facebook's machine translation mix-up sees man questic
attack threat. Photograph: Thibault Camus/AP

Alex Hern

Tue 24 Oct 2017 07.24 EDT

The Korea Herald

21:29 Spain, Ireland and Norway recognize Palestinian state

National | Social Affairs

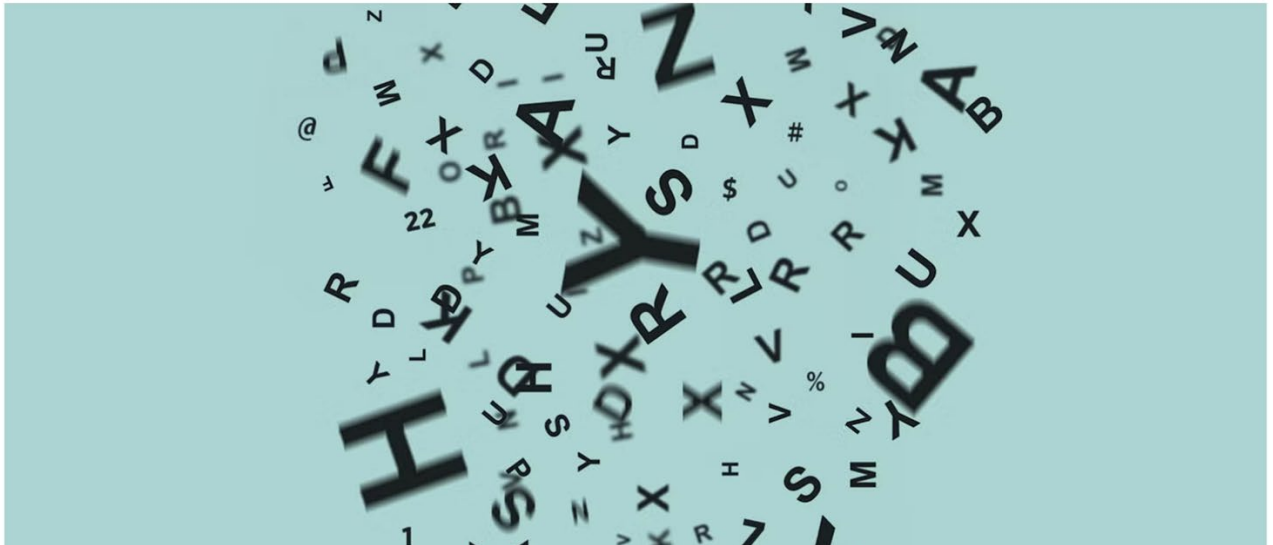
Hidden meaning of Korean term “agassi” leads to murder

From young woman to bar hostess, a term causes a fatal misunderstanding between a Chinese and Korean man

future tense

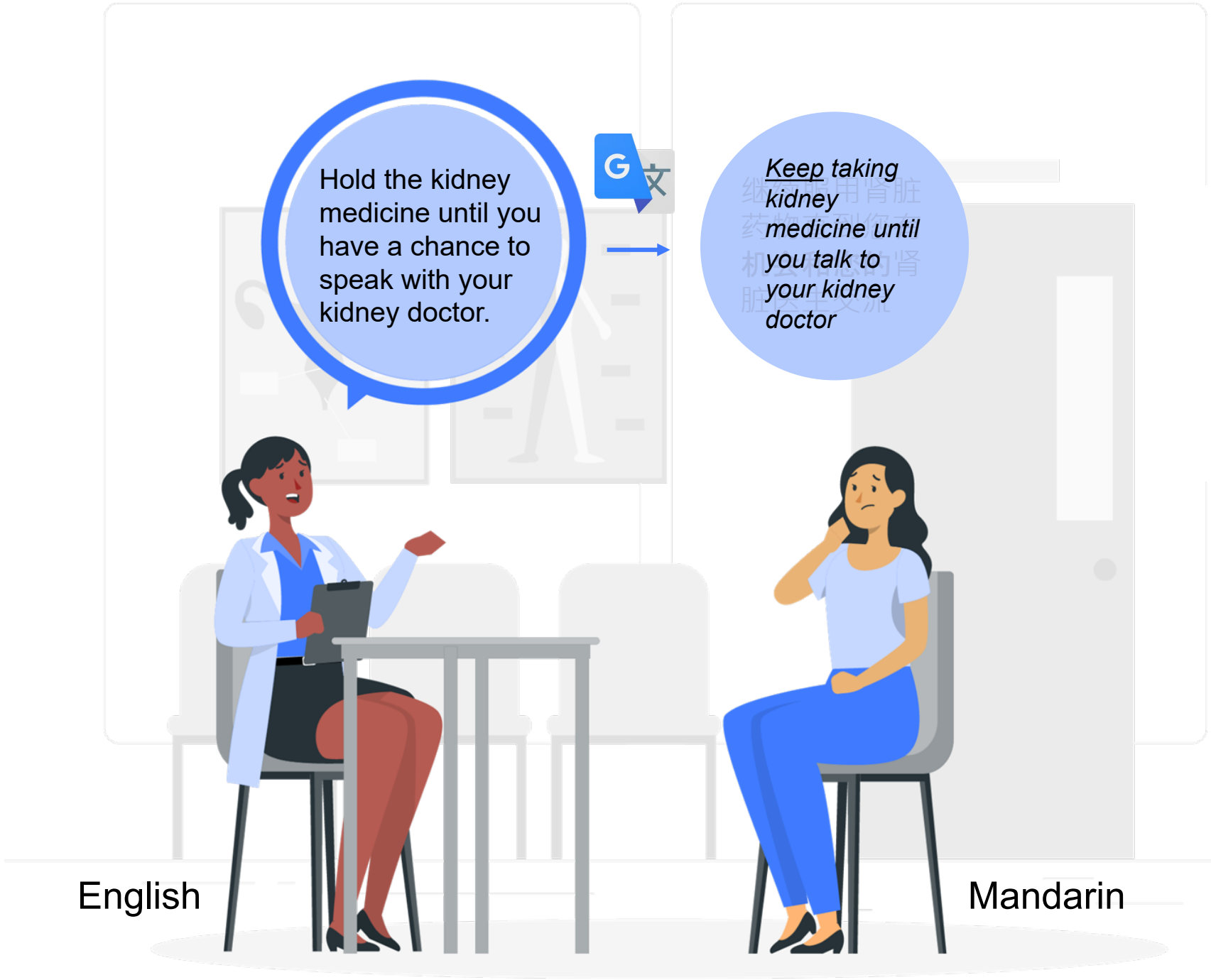
Death by Machine Translation?

BY SOFIA QUAGLIA SEPT 21, 2022 • 5:50 AM

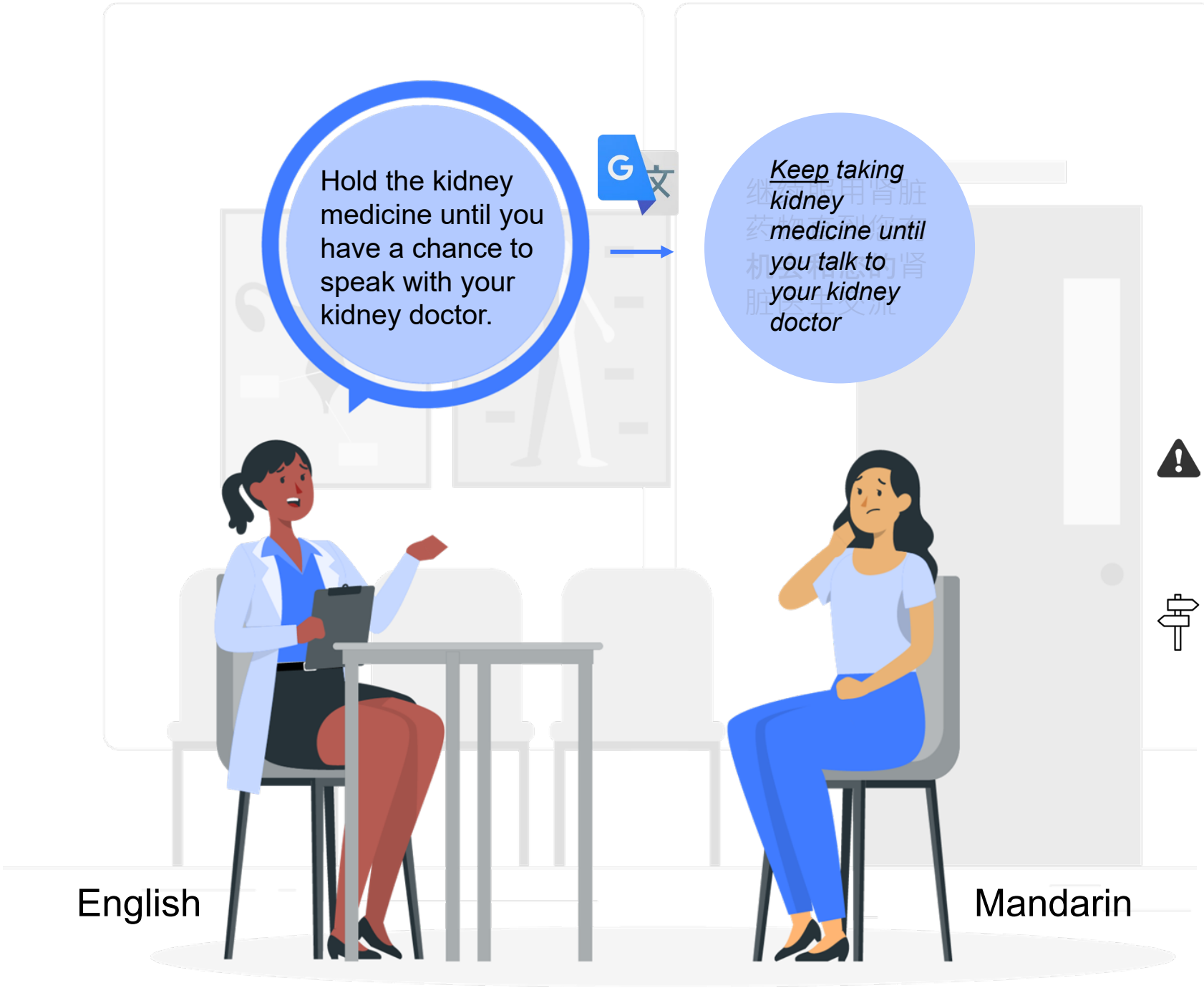


ge translation
s jeopardizing
m-seekers

[Khoong et al., 2019]



[Khoong et al., 2019]



Potentially life threatening error.

Hard to identify errors.

No actionable strategies to recover.

[Khoong et al., 2019]

Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors

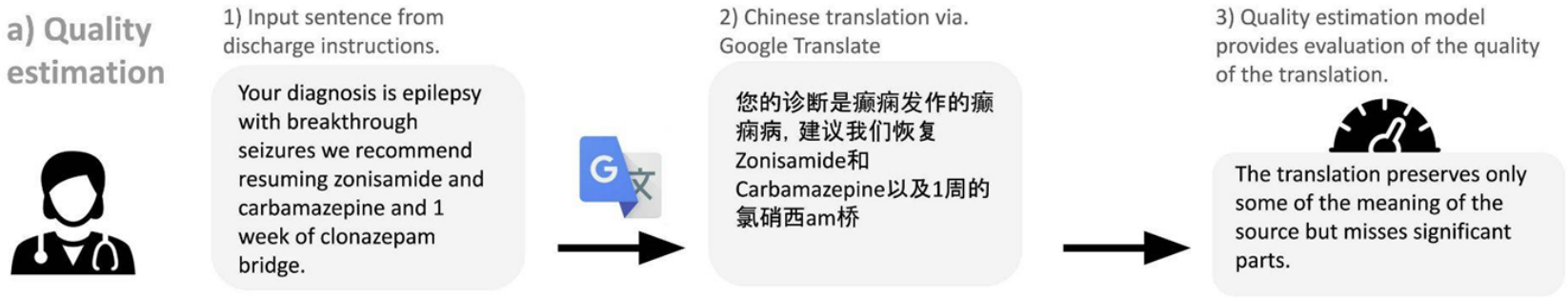
Nikita Mehandru[1], **Sweta Agrawal**[2], **Yimin Xiao**[2],
Elaine C Khoong[3], **Ge Gao**[2], **Marine Carpuat**[2], **Niloufar Salehi**[1]
[1]University of California, Berkeley [2]University of Maryland
[3]University of California, San Francisco

How can we help physicians decide whether MT is good enough to show patients who do not speak their language?

A case study toward **reliable MT,** motivated by a **real world use case**:

- Physicians routinely use MT to communicate with patients [Mehandru et al., 2022]
- even though errors can have severe clinical consequences [Khoong et al., 2019]

User Study of Physician Reliance on MT: 2 Interventions to give feedback on MT quality



Uses **COMET quality estimation** scores

[Rei et al. 2020]

- Correlate well with human ratings at WMT
- Converted to 3 level rating for readability

RAW QE SCORE	OUR QE LABEL
$[0.101; 1]$	The meaning of the translation is consistent with the source.
$[0.072, 0.101]$	The translation retains most of the meaning of the source.
$[-1; 0.072]$	The translation preserves only some of the meaning of the source but misses significant parts.

User Study of Physician Reliance on MT: 2 Interventions to give feedback on MT quality

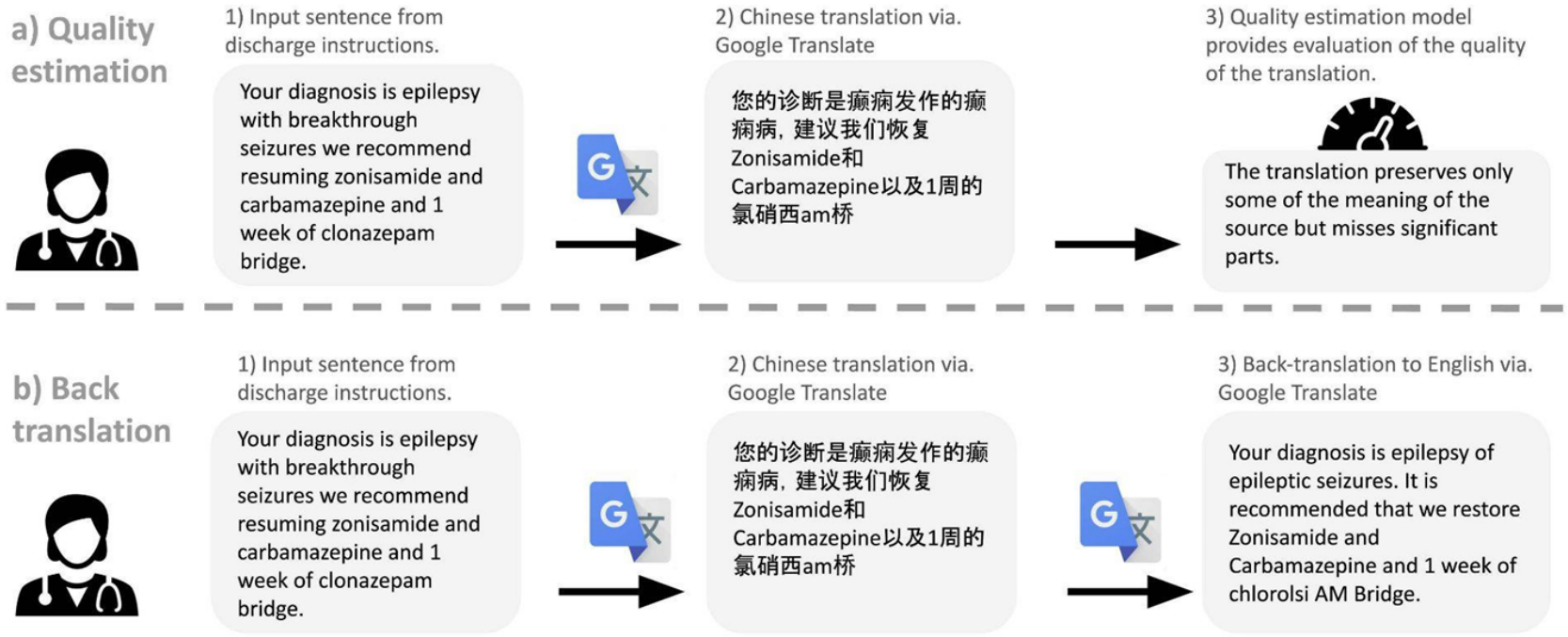
Backtranslation

Using the same translation engine

A noisy signal [Agrawal et al. WMT 2022]

But lay people do this all the time!

User Study of Physician Reliance on MT: Task



Would you give the Chinese translation (in bold) to a patient who only reads Chinese?

Yes

No

D1. How confident are you in your decision to give (or not give) the above translation to a patient?

1 2 3 4 5

Confidence level (1=Not Confident; 5=Very Confident)

User Study of Physician Reliance on MT: Study Setup

Randomized controlled experiment
with n=65 English-speaking physicians

- 45% interact with low English proficiency patients daily
- 17% use Google Translate on a monthly basis when writing discharge instructions

Within subjects: baseline then quality feedback

Between subjects: quality estimation vs. backtranslation

User Study of Physician Reliance on MT: Study Setup

Data from UCSF Emergency Department discharge instructions

Machine translated into Chinese via Google Translate

We collect reference adequacy and clinical risk ratings from 3 bilingual physicians

Measure of Appropriate Reliance: Confidence Weighted Accuracy

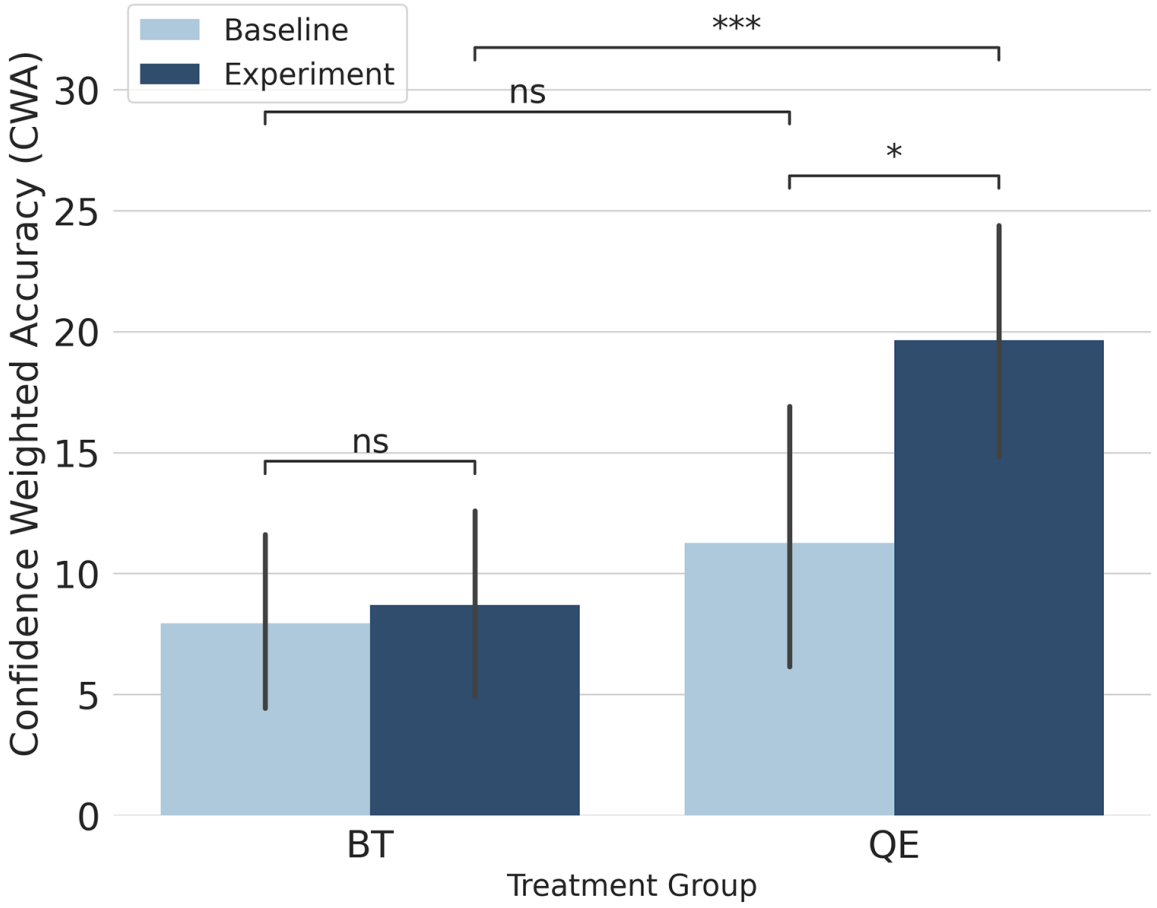
c_s: participant confidence in their decision s on 5-point scale

$$CWA = \frac{1}{S}\sum_{s \in S} sign(s)\frac{c_s}{5}$$

sign(s): +1 if participant decision s is consistent with reference rating; -1 otherwise

[Ebel, 1965]

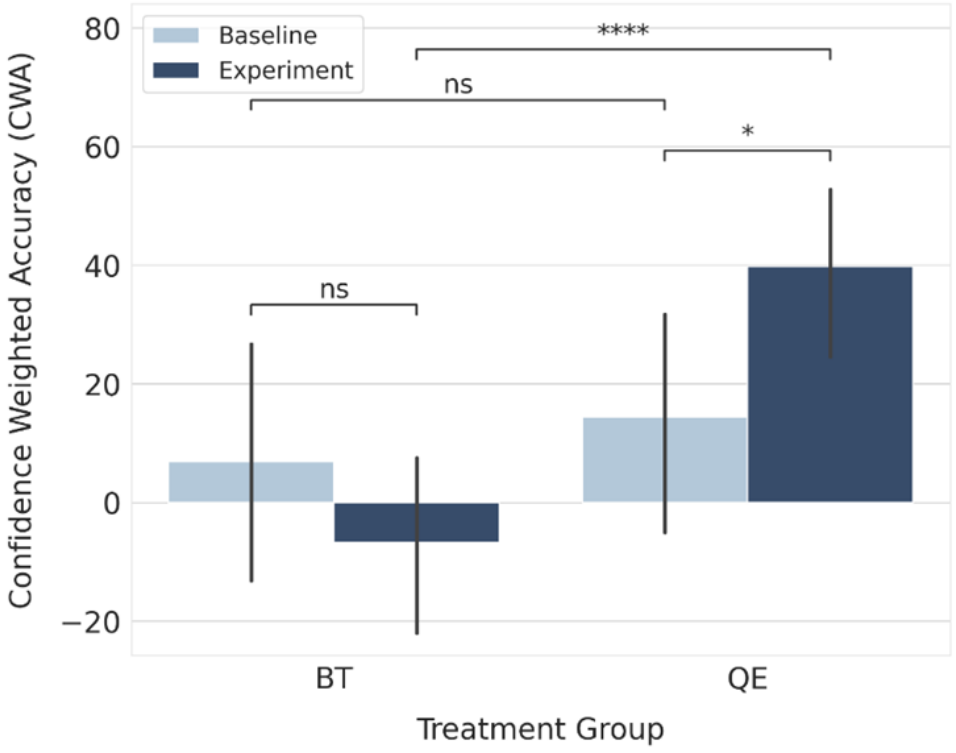
User Study of Physician Reliance on MT: Results



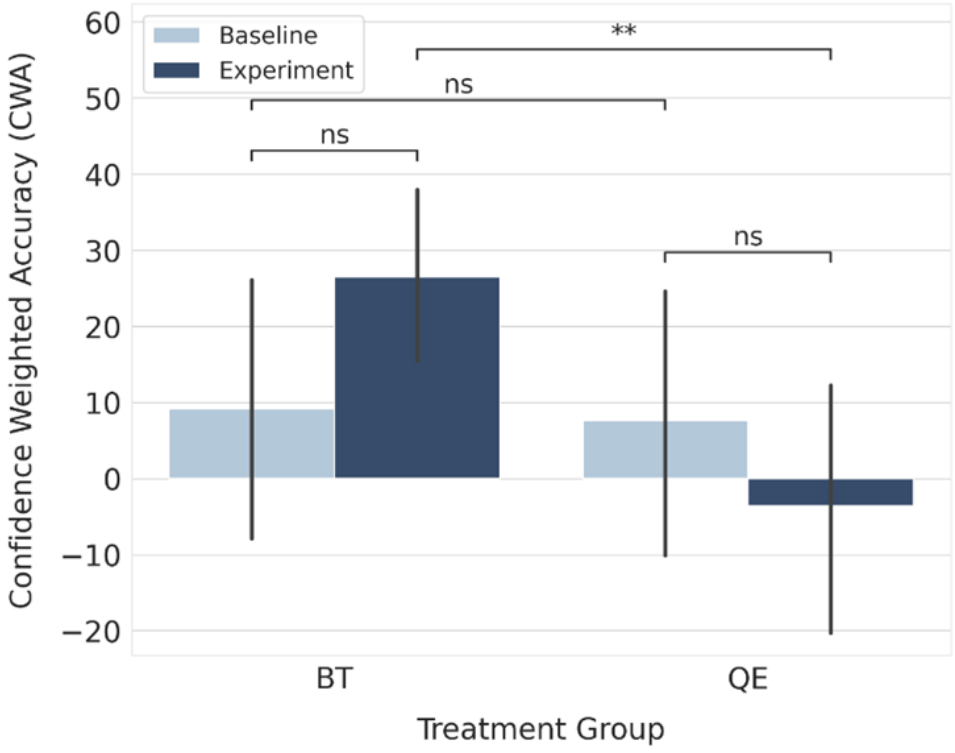
Showing a QE model output made physicians better at deciding *when to rely*.

Backtranslation did not.

User Study of Physician Reliance on MT: Results



(a) Correct Identification of Adequate Translations



(b) Correct Identification of Inadequate Translations

Physicians in the BT condition were better at rejecting errors, and clinically significant errors

User Study of Physician Reliance on MT: Results

Discharge Instruction	Clinical Risk	Accuracy (%) BT	QE
We believe this was due to something called vasovagal response which causes people to feel faint or lose consciousness.	moderately significant	80	23
Please try to drink fluids to make sure you are hydrated today.	moderately significant	89	40
Be sure to avoid taking NSAIDs (ibuprofen, naproxen) and Aspirin for pain.	mildly significant	31	23
The orthopedic surgeons drained your knee.	highly significant	91	37
You may take norco for pain, however, do not drink or drive as it will cause drowsiness.	moderately significant	71	43
Your diagnosis is epilepsy with breakthrough seizures we recommend resuming zonisamide and carbamazepine and 1 week of clonazepam bridge.	life-threatening	97	80
Please START taking and Doxycycline 100mg twice per day for seven days - Keflex 500mg four times per day for seven days	highly significant	94	53

BT group better at rejecting sentences with clinical risk.

Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors

Nikita Mehandru[1], Sweta Agrawal[2], Yimin Xiao[2], Elaine C Khoong[3], Ge Gao[2], Marine Carpuat[2], Niloufar Salehi[1]
[1]University of California, Berkeley [2]University of Maryland
[3]University of California, San Francisco

How can we help physicians decide whether MT is good enough to show patients who do not speak their language?

- Quality estimation models can help users decide *when* to rely on MT
- But backtranslation identifies critical errors better
- Likely need range of interventions for different types of error, and more “in vivo” evaluations

Beyond translation:
NLP for cross-lingual communication

How can we make communication mediated by NLP tools reliable and trustworthy?

Explainable quality estimation

error detection, guided post-editing, speechQE [Briakou et al, 2023, Ki & C. 2024, Han et al. 2024]

MT-mediated conversation dynamics

between native and non-native English speakers [Zhang et al. 2022; Yang et al., under review]

MT and AI literacy in general public

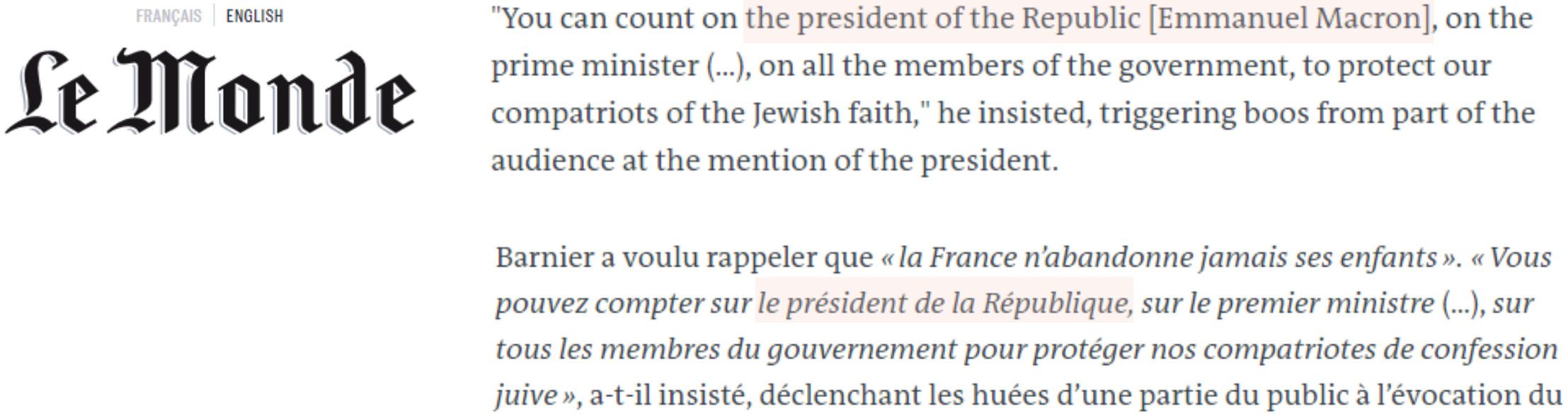
ongoing study at Planet Word museum with Ge Gao, Charlotte Vaughn et al.

Beyond Translation: NLP for Cross-Lingual Communication

Part I Reliability in MT-mediated communication

Part II MT as controllable language generation

When accurate translations are not enough



Explicitation: "*A procedure of* ***explicitly introducing details into the target language*** *which remain* ***implicit in the source language***."

[Vinay & Darbelnet 1958]

https://www.lemonde.fr/en/politics/article/2024/10/08/at-paris-tribute-to-october-7-victims-macron-s-call-to-halt-arms-deliveries-to-israel-earns-him-boos_6728593_5.html

Explicitation in Translation

Explicitations can take several forms [Klaudy, 1983]

- Obligatory: driven by syntactic divergences across languages
- Optional: to improve fluency and naturalness in the target language
- **Pragmatic: explicitly explains implicit background knowledge**

...

We focus on pragmatic explicitations

- enable a communicative act [Snell-Hornby 2006]
- adapt translations to a communication context

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation

HyoJung Han
Computer Science
University of Maryland
hjhan@cs.umd.edu

Jordan Boyd-Graber
CS, UMIACS, iSchool, LCS
University of Maryland
jbg@umiacs.umd.edu

Marine Carpuat
Computer Science, UMIACS
University of Maryland
marine@cs.umd.edu

EMNLP 2023

Can we generate explicitations automatically?

How can we evaluate explicitation generation?

Challenge: explicitations are rare!

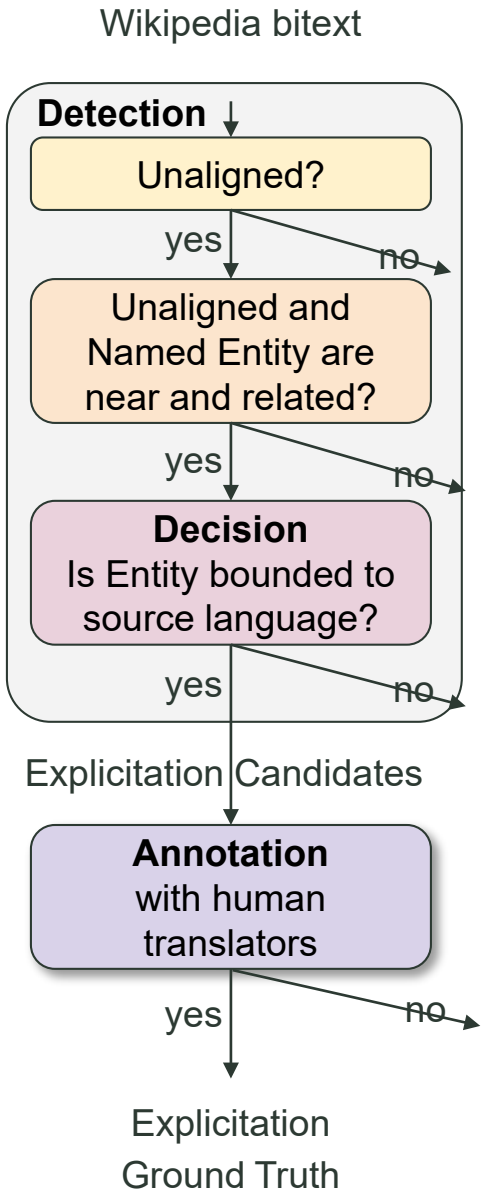
Building WikiExpl

A dataset of explicitations

- drawn from English Wikipedia
- with French, Polish and Spanish source text

Source	**…l'attentat contre Charlie Hebdo…**
Gloss	…the attack against Charlie Hebdo…
Target	**…the attack against the French satirical newspaper Charlie Hebdo…**

Building WikiExpl



Step 1: Collect candidates automatically for recall

Step 2: Human annotation

Building WikiExpl

Step 1: Collect candidates automatically for recall

		Explicitation?
Source	**…avec des vols sans escale vers Antigua,**	
Gloss	…with non-stop flights to Antigua,	X
Target	**…with direct service to Antigua,**	
Source	**…l'attentat contre Charlie Hebdo…**	
Gloss	…the attack against Charlie Hebdo…	
Target (unaligned)	**…the attack against the French satirical newspaper Charlie Hebdo…**	?

Assumption 1:
explicitations are unaligned token sequences

Building WikiExpl

Step 1: Collect candidates automatically for recall

		Explicitation?
Source	…comme les persécutions à Metz en 888.	
Gloss	…like the persecution in Metz in 888.	
Target (unaligned)	…such as the one in Metz in 888.	X
Source	…l'attentat contre Charlie Hebdo…	
Gloss	…the attack against Charlie Hebdo…	
Target (unaligned)	…the attack against the French satirical newspaper Charlie Hebdo…	?
Source	À l' adolescence , il écoute de la musique latino-américaine,…	
Gloss	In adolescence, he listened to Latin American music,…	
Target (unaligned)	In his teenage years in Bogotá , he acquired a taste for Latin American music,…	X

Assumption 2: explicitations are adjacent and relevant to named entities

Building WikiExpl

Step 1: Collect candidates automatically for recall

Explicitation?

Source	**…l'attentat contre Charlie Hebdo…**	
Gloss	…the attack against Charlie Hebdo…	?
Target (unaligned)	**…the attack against the French satirical newspaper Charlie Hebdo…**	
Source	**…et au sud du Texas.**	
Gloss	…and in southern Texas,	
Target (unaligned)	**…and into southern Texas in the United States.**	X

Assumption 3: explicitations focus on entities well known in source language/culture but not in target language/culture

measured by

- number of hops in Wikipedia from entity to source vs. target language country
- number of incoming links + length of entity page in source vs. target language

Building WikiExpl

Step 1: Collect candidates automatically for recall

Step 2: Human annotation

Explicitation?

Source	**…l'attentat contre Charlie Hebdo…**
Gloss	…the attack against Charlie Hebdo…
Target (unaligned)	**…the attack against the French satirical newspaper Charlie Hebdo…**

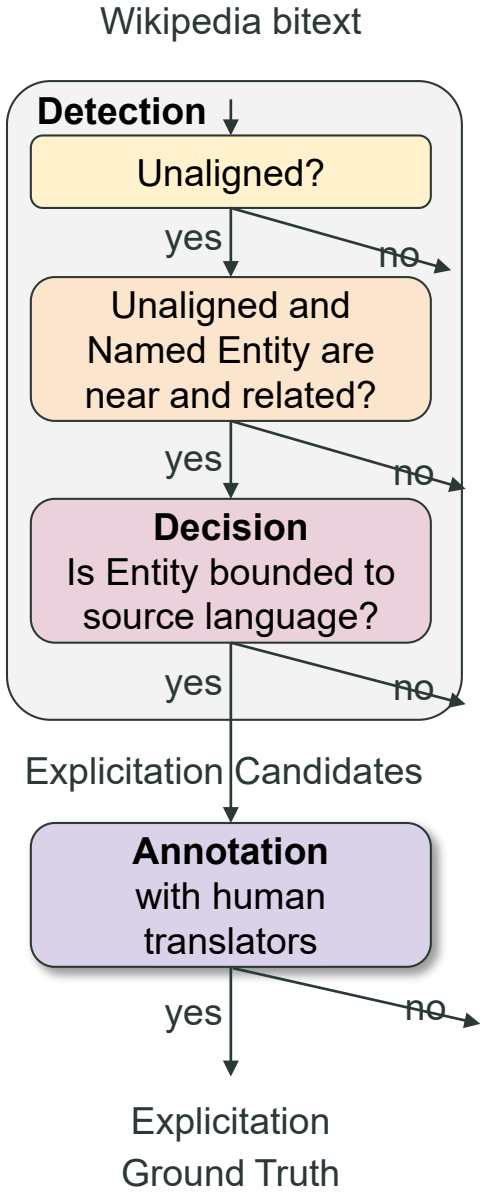
Source	**…nommé d' après Pierre André de Suffren,**
Gloss	…appointed after Pierre André de Suffren,
Target (unaligned)	**…named after the 18th century admiral Pierre André de Suffren,**

Annotation task

- What is the role of the highlighted span? Does it add information to the source?
- If yes, does it explain general knowledge in the source language to target language readers
- Optionally, correct span and justify decision.

3 professional translators per candidate ($\kappa \approx 0.7$)

Building WikiExpl



Step 1: Collect candidates automatically for recall

Step 2: Human annotation

Building WikiExpl

Source Language	French	Polish	Spanish
WikiMatrix	29826	21392	28900
Candidates	791	245	307
Top 1 country			
Annotated	460	244	220
Explicitation	116	67	44

Type	Source	Target
Hypernym (h)	la Sambre	the Sambre ***river***
	Belwederze	Belweder ***Palace***
	Pleszowa	***village of*** Pleszów
Occupation/ Title (o)	Javier Gurruchaga	***showman*** Javier Gurruchaga
	Jana III	***king*** John III ***Sobieski***
	Juan Carlos de España	***Prince*** Juan Carlos of Spain
Acronym Expansion (a)	PP	***People 's Party*** (PP)
	TVE	***Televisión Española of Spain*** (TVE)
	PCE	***Communist Party of Spain***
	PSL	***Polish People 's Party***
Full names (f)	Cervantes	***Miguel*** de Cervantes
	Piłsudskiego	***Józef*** Piłsudski
	Sierakowa	Sieraków ***Wielkopolski***
Nationality (n)	Felipe II	Philip II ***of SPAIN***
	Toruniu	Toruń (***Poland***)
	Cuenca	Cuenca ***in Spain***
	Troyes	Troyes ***, France***
Integrated (i)	Dominique de Villepin	***former French Prime Minister*** Dominique de Villepin
	El País	***Spanish newspaper*** El País
	Zygmunt III Waza	***Polish King*** Sigismund III Vasa
	Przekrój	***Polish weekly magazine*** Przekrój

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation

HyoJung Han
Computer Science
University of Maryland
hjhan@cs.umd.edu

Jordan Boyd-Graber
CS, UMIACS, iSchool, LCS
University of Maryland
jbg@umiacs.umd.edu

Marine Carpuat
Computer Science, UMIACS
University of Maryland
marine@cs.umd.edu

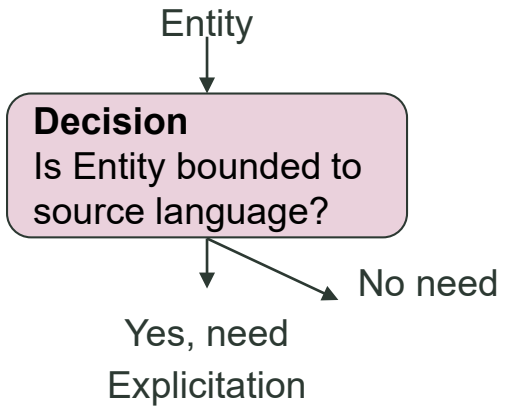
EMNLP 2023

Can we generate explicitations automatically?

How can we evaluate explicitation generation?

Challenge: explicitations are rare!

A method for automatic explicitation generation using Wikipedia

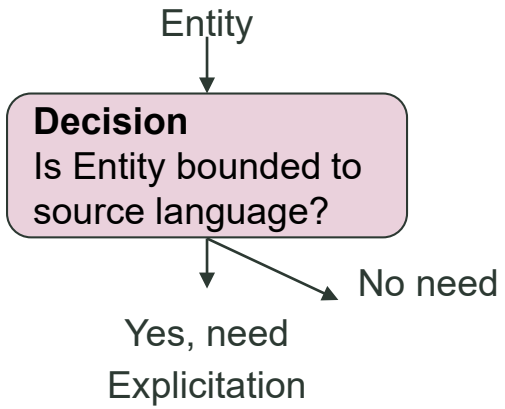


Is explicitation needed?

We predict that an entity needs explicitation if two conditions hold:

1. Closeness to source country: is there a direct relation link to the language speaking country in Wikipedia?
2. Popularity of topic: difference in normalized length of Wikipedia page in source vs. target language

A method for automatic explicitation generation using Wikipedia



Type	Length	Form	Example of "Sambre,"
SHORT	1–2 words	Appositive	Sambre river,
MID	3 words–a phrase	+ Parenthetical	Sambre, river in France and Belgium,
LONG	1–3 sentences	Footnote	Sambre (a river in northern France and in Wallonia, Belgium. It is a left-bank tributary of the Meuse, which it joins in the Wallonian capital Namur.)

Is explicitation needed?

If yes, extract explicitation from Wikipedia and integrate in translation

How can we evaluate explicitation generation?

Intrinsic human evaluation

By professional translators

Extrinsic automatic evaluation

How useful are explicitations in multilingual Question Answering tasks?

Intrinsic Human Evaluation

- Done by same professional translators as WikiExpl validation
- 3 point Likert scale mapped to [0,1], higher is better
- On ~2000 Polish segments

Is the explicitation needed?

Is the explanation appropriate?

Is the explanation well integrated?

Decision	Type	Generation	Integration
0.71	SHORT	0.63	0.79
	MID	0.82	0.92
	LONG	0.95	_

Extrinsic Evaluation of Explicitations via Multilingual Question Answering

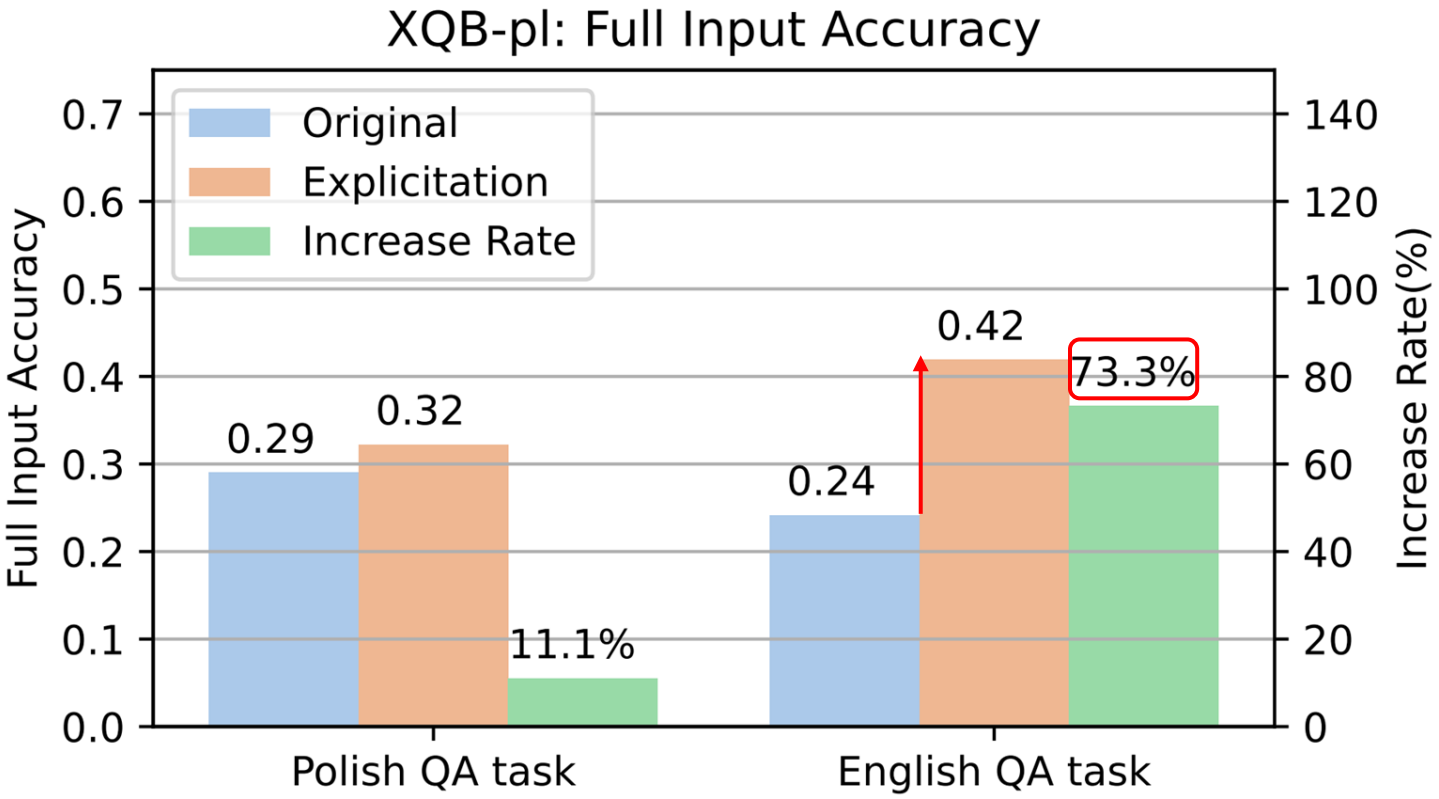
Question Type	Input text	System Answer
Polish Question	...Aby otrzymać punkt, nazwij łańcuch górski, w którego skład wchodzą między innymi Tatry,...	Karpaty (O)
with Explanation (**Explicitation**)	...To get a point, name the mountain range of which the Tatras (mountain range on the Polish–Slovak border) are a part,...	Carpathian Mountain (O)

Extrinsic Evaluation of Explicitations via Multilingual Question Answering

Question Type	Input text	System Answer
Polish Question	...Aby otrzymać punkt, nazwij łańcuch górski, w którego skład wchodzą między innymi Tatry,...	Karpaty (O)
with Explanation	...Aby otrzymać punkt, nazwij łańcuch górski, w którego skład wchodzą między innymi Tatry (pasmo górskie na granicy polsko-słowackiej),...	Karpaty (O)
English Question	...To get a point, name the mountain range of which the Tatras are a part,...	(wrong answer) (X)
with Explanation (**Explicitation**)	...To get a point, name the mountain range of which the Tatras (mountain range on the Polish–Slovak border) are a part,...	Carpathian Mountain (O)

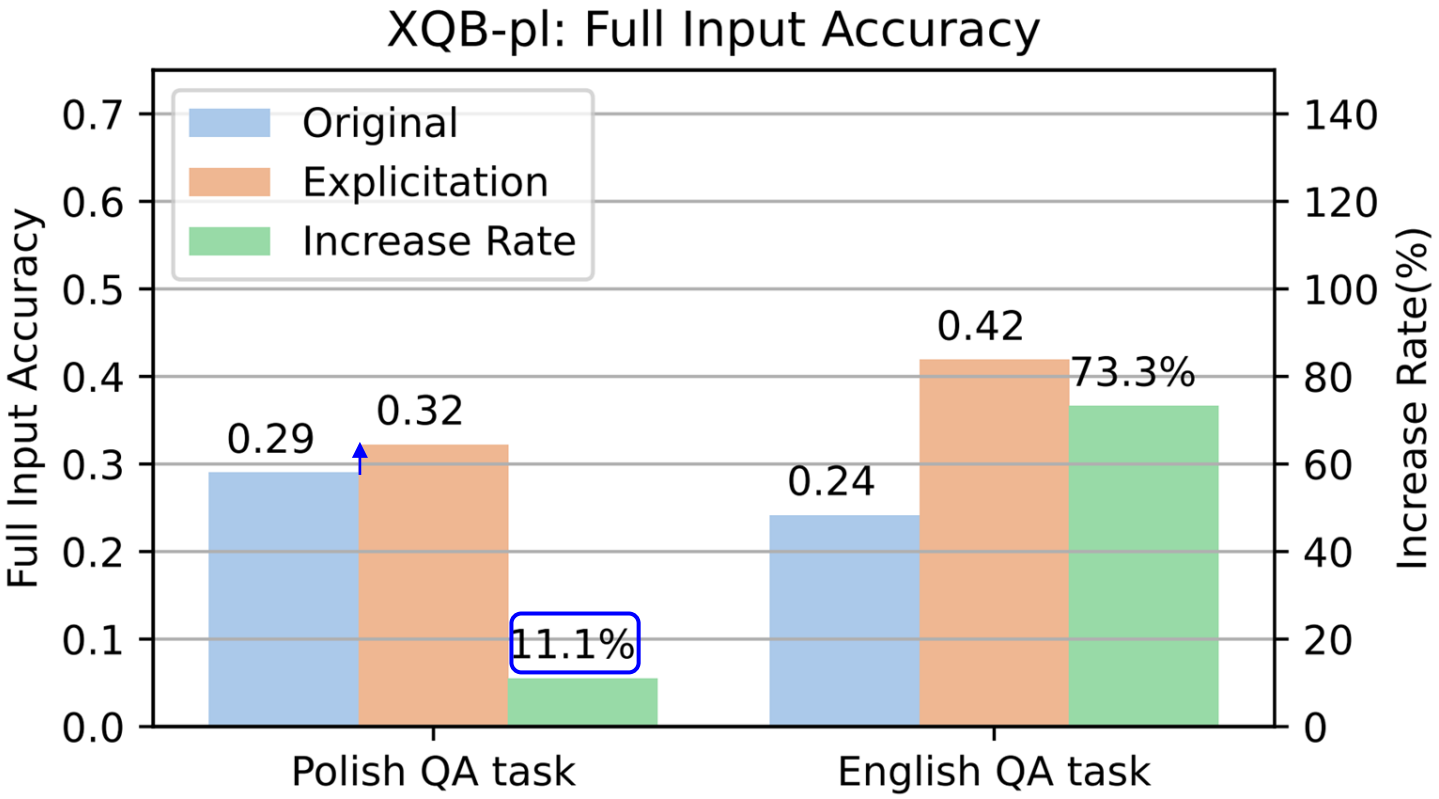
If the automatic explanation ...	Helps target QA (English)	Doesn't help target QA (English)
Helps source QA (Polish)	useful but not explicitation ✗	useful but not explicitation ✗
Doesn't help source QA (Polish)	useful explicitation ✓	not useful ✗

Extrinsic Evaluation of Explicitations via Multilingual Question Answering



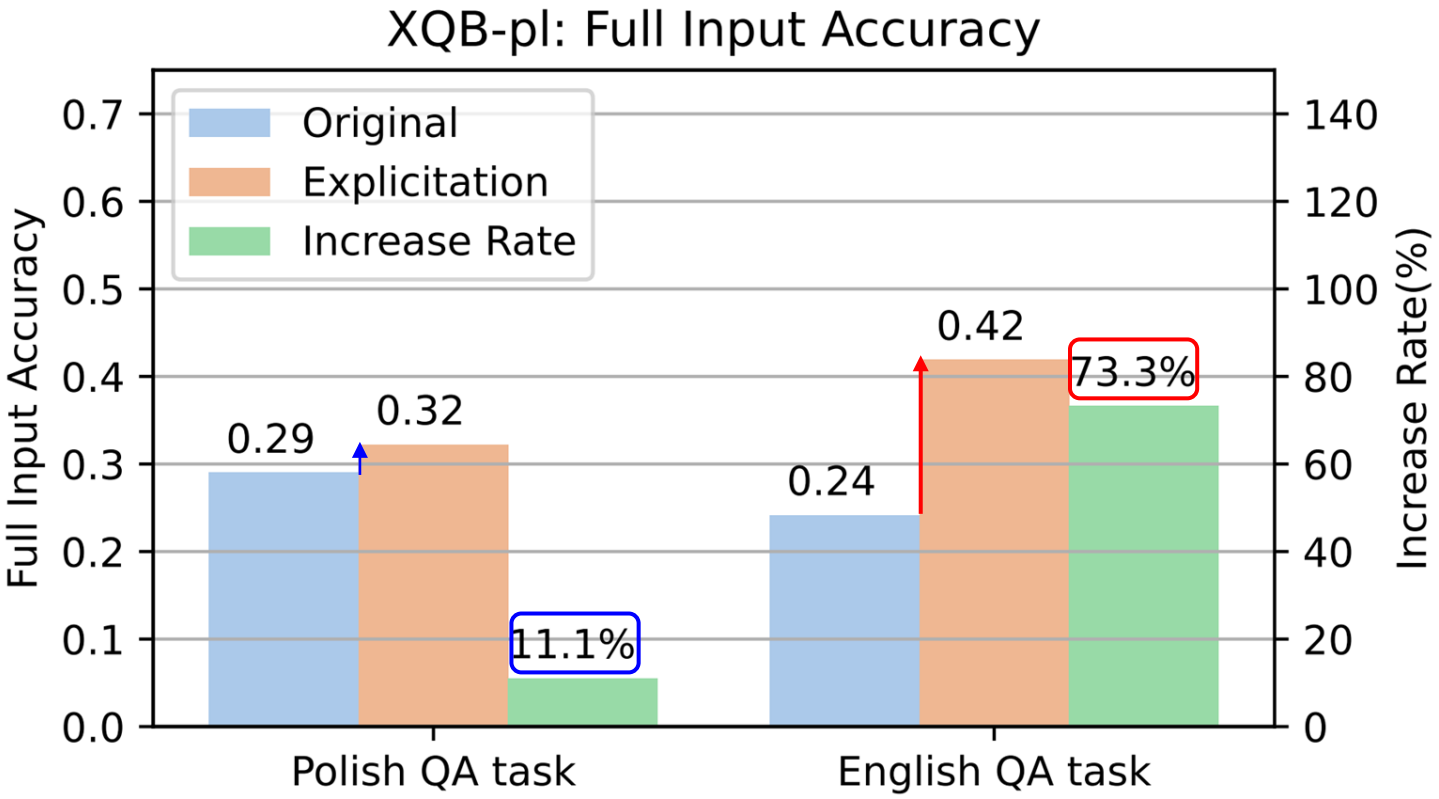
QA accuracy improves by 73% with explicitations in English

Extrinsic Evaluation of Explicitations via Multilingual Question Answering



QA accuracy with symmetric explanations in Polish improves by 11.1%

Extrinsic Evaluation of Explicitations via Multilingual Question Answering



QA accuracy improves with explicitations in English more than with explanations in Polish

So explicitations are useful!

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation

EMNLP 2023

HyoJung Han
Computer Science
University of Maryland
hjhan@cs.umd.edu

Jordan Boyd-Graber
CS, UMIACS, iSchool, LCS
University of Maryland
jbg@umiacs.umd.edu

Marine Carpuat
Computer Science, UMIACS
University of Maryland
marine@cs.umd.edu

To bridge the gap of background knowledge between source and target audiences in translation, we contribute:

- **Automatic Explicitation**: techniques to automatically decide what to explain and how
- **WikiExpl**: dataset semi-automatically collected from Wikipedia
- **Evaluation**: explicitations are valid based on human evaluation, and useful based on extrinsic multilingual QA.

Beyond translation:
NLP for cross-lingual communication

How can we generate controlled outputs tailored to different audiences and use cases?

Background

[Han et al. 2023]

Register

formality controlled MT [Rippeth et al. 2022]

Complexity

controllable text simplification & complexity-controlled MT [Agrawal & C. 2020]

Author Style

authorship analysis & obfuscation [Richburg & C. 2024, Bao & C. 2024]

Beyond Translation: NLP for Cross-Lingual Communication

Part I Reliability in MT-mediated communication

Part II MT as controllable language generation

Beyond Translation:
NLP for Cross-Lingual Communication

Helping people communicate across languages with NLP requires much more than MT.

We need better tools to help them rely on outputs appropriately, adapt outputs to context of use, craft appropriate inputs, assess impact of tools on communication dynamics, use text and speech in many languages …

We need a broader set of evaluations to assess usefulness extrinsically and in human studies.

Beyond Translation: NLP for Cross-Lingual Communication

Part I Reliability in MT-mediated communication

Part II MT as controllable language generation

marine@umd.edu / marine.carpuat@inria.fr