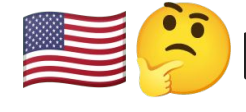


HyoJung Han, Jordan Boyd-Graber, Marine Carpuat  
University of Maryland, College Park in Department of Computer Science



Source

- 1 ...frère de [Dominique de Villepin](#)...
- 2 La veille de Noël [1800](#), sachant qu'...
- 3 ...l'attentat contre [Charlie Hebdo](#)...



Literal Translation

- 1 ...brother of [Dominique de Villepin](#) ...
- 2 The day before Christmas [1800](#), knowing that
- 3 ...the attack against [Charlie Hebdo](#)...



with Explicitation

- 1 ...brother of the former **French Prime Minister** [Dominique de Villepin](#) ...
- 2 On Christmas Eve in [1800](#), amid the **French Revolution**,
- 3 ...the attack against the **French satirical newspaper** [Charlie Hebdo](#)...




## Detecting Explicitation

- We first **extract candidates** from the noisy parallel data [3]
- Then, ask human annotators to label if it is Explicitation or not.

### How to Extract Explicitation Candidates

1. Find **unaligned** segments via word alignment among bitext.
2. Pair segment with the **adjacent and related** named **entity**  $e$ .
3. See if the paired entity is culturally distant entities.  
→ measure property values shift of entity  $e$  conditioned on each language of bitext pairs  $\text{prop}(e|l_{src}) - \text{prop}(e|l_{tgt}) > \tau$   
e.g. the number of hops from an entity to lang speaking country

### Collected "WIKIEXPL" Statistics and Examples

Source Language	French	Polish	Spanish
WikiMatrix	29826	21392	28900
Candidates	791	245	307
Top 1 country			
Annotated	460	244	220
Average $\kappa$	0.66	0.72	0.74
At Least one vote	236	111	73
Explicitation	116	67	44

Source	Target
la Sambre	the Sambre <i>river</i>
Javier Gurruchaga	<i>showman</i> Javier Gurruchaga
PP	<i>People 's Party</i> (PP)
Cervantes	<i>Miguel</i> de Cervantes
Felipe II	Philip II <i>of Spain</i>
Dominique de Villepin	<i>former French Prime Minister</i> Dominique...

## Automating Explicitation

### 1. Deciding if Explicitation is Needed

- Find entities that are tightly bound to the socio-linguistic context.
- Adjust the candidate extraction heuristics with collected data

### 2. Generating the Explanation

- Fetch text from Wikidata and Wikipedia page linked to entity
- Integrate in the form of Appositive, Parenthetical or Footnote
- Three types of generation by the length

Type	Length	Example of "Sambre,"
SHORT	1–2 words	Sambre river,
MID	3 words—a phrase	Sambre, river in France and Belgium,
LONG	1–3 sentences	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.)

[1] Stylistique Comparee Du Francais Et De L'anglais (Vinay and Darbelnet, Bibliotheque de stylistique comparee 1958)

[2] On explicitation Hypothesis (Klaudy, Dániel Berzsenyi College 1993)

[3] WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia (Schwenk et al., EACL 2021)

[4] SimQA: Detecting Simultaneous MT Errors through Word-by-Word Question Answering (Han et al., EMNLP 2022)

## Explicitation of Implicit Background Information

### What is Explicitation?

"The process of *explicitly* introducing details into the target language which remain *implicit* in the source language." [1]

- The term covers broad types of details that is newly introduced in the target language caused by difference of syntactic/semantic/discourse structure or by fluency/naturalness.
- We focus on explicitation that **explicitly explains implicit background knowledge** (Pragmatic Explicitation [2])

## Experimental Settings

- Dataset** XQB, Cross-lingual parallel Quizbowl QA set [4]
- Model** LLAMA for multilingual QA system
- Metric** Expected Wins (EW), which considers both speed and accuracy at the same time [4]

## Evaluating Explicitation

### Intrinsic/Direct Evaluation

- Ask human annotators quality of our automatic explicitation

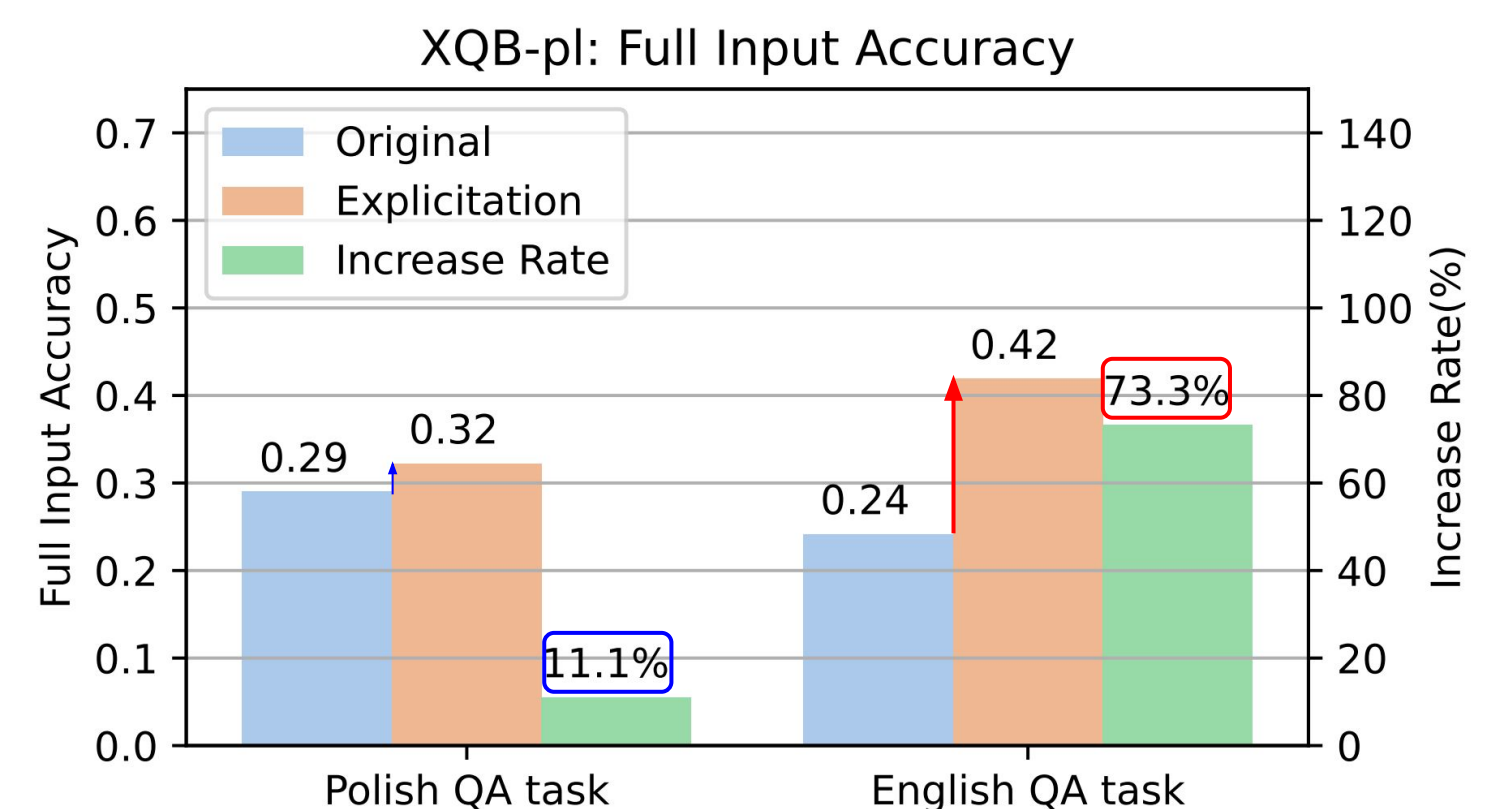
Decision	Type	Generation	Integration
0.71	SHORT	0.63	0.79
	MID	0.82	0.92
	LONG	0.95	–

### Extrinsic Evaluation using Multilingual QA system

- Assumption: given source Polish questions, well-generated explicitations in English will improve English QA

	Useful in target QA	Not Useful in target QA
Useful in source QA	Subject we don't know commonly	Subject is not popular in source community
Not Useful in source QA	Subject is well-known only in source community → <b>Need Explicitation!</b>	Subject is well-known globally

- Higher increase rate in English QA task indicates the effectiveness of our automated explicitation methods



## Related Works

- Explicitation in Contemporary Works [5][6] (discourse)
- Culturally aware MT [7][8] (LLM prompting)
- Elaborative Text Simplification [9][10]

[5] Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English-German Translation (Lapshinova-Koltunski & Hardmeier, DiscoMT 2017)

[6] The Role of Expectedness in the Implication and Explicitation of Discourse Relations (Hoek et al., DiscoMT 2015)

[7] Empowering LLM-based Machine Translation with Cultural Awareness (Yao et al., 2023)

[8] Audience-specific Explanations for Machine Translation (Lou and Niehues, 2023)

[9] Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification (Srikanth & Li, Findings 2021)

[10] Elaborative Simplification as Implicit Questions Under Discussion (Wu et al., EMNLP 2023)