Translation-based Lexicalization Generation and Lexical Gap Detection: Application to Kinship Terms

Senyu Li, Bradley Hauer, Ning Shi, Grzegorz Kondrak

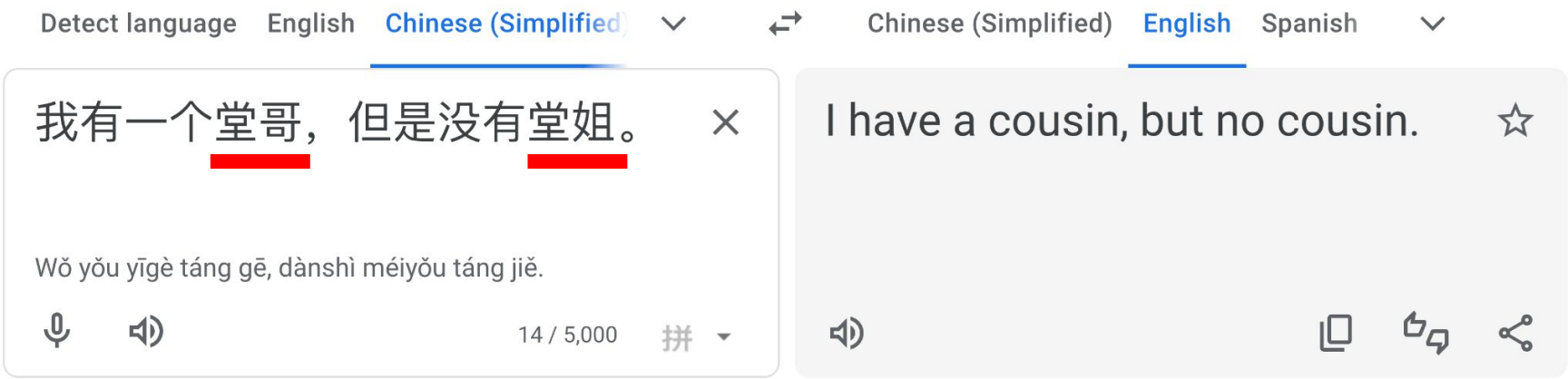
Alberta Machine Intelligence Institute, Dept of Computing Science
University of Alberta, Edmonton, Canada

UNIVERSITY OF ALBERTA

An Error Case: Google Translate

- 堂哥 “elder **son** of father’s brother” => "cousin"
- 堂姐 “elder **daughter** of father’s brother” => "cousin"

- Other powerful translators make similar errors. (DeepL, Baidu, etc.)



*Google Translate, February 15, 2024.

Sample Output of ChatGPT

You

Given a word that means [father's younger brother] in Chinese is [叔叔], and a word that means [mother's brother] in Chinese is [舅舅]. Is there a word that means [elder brother] in [English]? If yes, give me that word. If no, say no.

ChatGPT

Yes, the word in English that means "elder brother" is "brother."

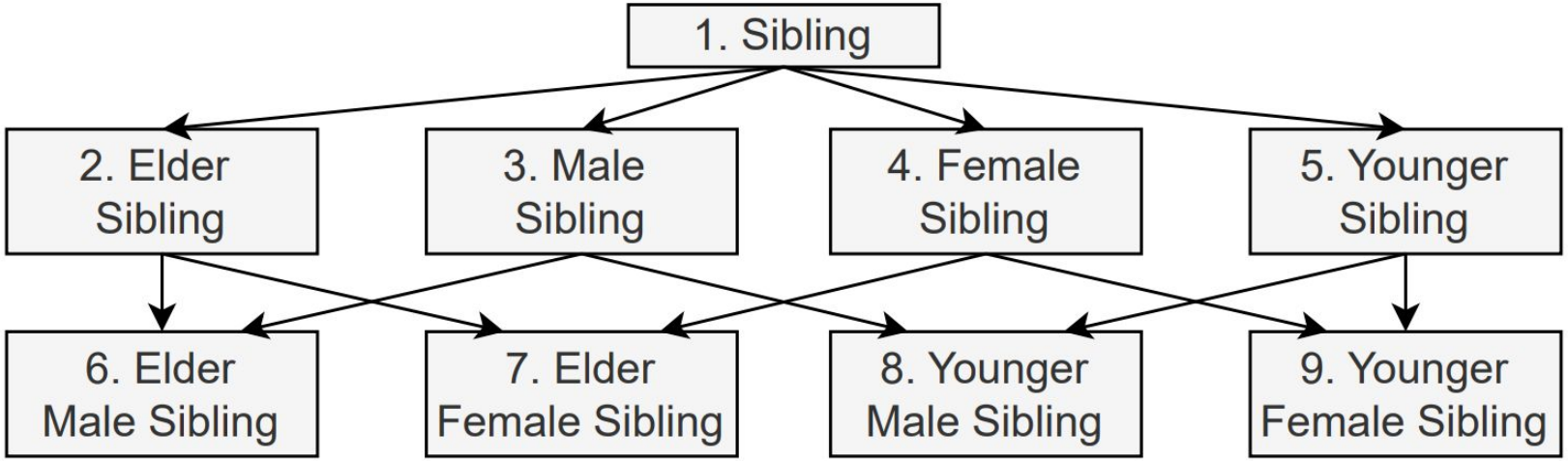
*GPT-3.5, February 15, 2024.

Outline

- **Problem:** How to identify concept lexicalizations and lexical gaps efficiently?
- **Idea:** If a concept is an exclusive disjunction of its hyponym concepts then all three concepts should have different lexicalizations.
- **Method:** Generate a candidate lexicalization for each concept by translating an unambiguous lexicalization into the target language in the context of the concept gloss. Then filter out incorrect translations using the above idea.
- **Results:** Empirical evaluations demonstrate that our approach yields higher accuracy than BabelNet and ChatGPT.

Concepts

- **Concept: discrete word meaning**
- Kinship concepts have clear definitions and hierarchical structure
  - Well-studied, good gold-standard dataset (Khishigsuren et al, 2022)



Lexicalizations and Lexical Gaps

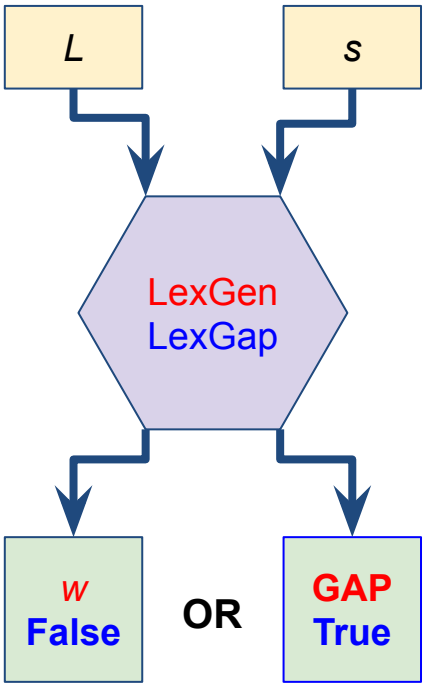
- **Lexicalization:** a single word which can express (i.e. lexicalize) a concept.
- **Lexical Gap:** a concept that has no lexicalization in a given language.

Concepts	En	Es	Fr	Ja	Fa	Zh	Pl
1	Sibling	ø	fratrie	ø	ø	同胞	ø
2	ø	ø	ø	ø	ø	ø	ø
3	Brother	hermano	frère	ø	برادر	兄弟	brat
4	Sister	hermana	sœur	ø	خواهر	姐妹	siostra
5	ø	ø	ø	ø	ø	ø	ø
6	ø	ø	ø	兄さん	ø	哥哥	ø
7	ø	ø	ø	姉ちゃん	ø	姐姐	ø
8	ø	tato	ø	おとうと	ø	弟弟	ø
9	ø	ø	ø	いもうと	ø	妹妹	ø

Data from *Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship* (Khishigsuren et al, 2022)

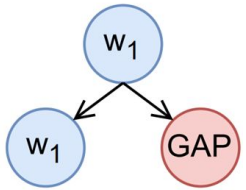
Task definition: LexGen and LexGap

- **LexGen:** Lexicalization Generation
  - Input: language *L*, concept *s*
  - Output: word *w* in *L* s.t. *w* lexicalizes *s,* **OR** a special token **GAP** indicating that no such *w* exists
- **LexGap:** Lexical Gap Detection
  - Input: language *L*, concept *s*
  - Output: **True** if no word in *L* lexicalizes *s*, **False** otherwise.
- LexGen(*L*,*s*) = **GAP** <u>if and only if</u> LexGap(*L*,*s*) = **True**



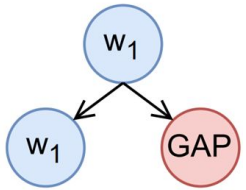
Theoretical Basis

- **Proposition 1:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing P and C1 with the same word w can result in a colloquial contradiction.

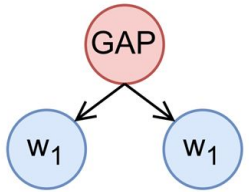


Theoretical Basis

- **Proposition 1:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing P and C1 with the same word w can result in a colloquial contradiction.

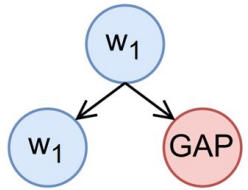


- **Proposition 2:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing C1 and C2 with the same word w can result in a colloquial contradiction.

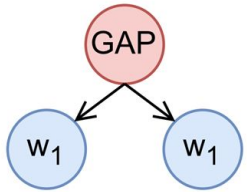


Theoretical Basis

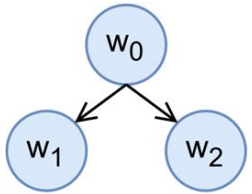
- **Proposition 1:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing P and C1 with the same word w can result in a colloquial contradiction.



- **Proposition 2:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing C1 and C2 with the same word w can result in a colloquial contradiction.



- **Corollary:** If a concept P is an exclusive disjunction of its hyponyms C1 and C2 then all their lexicalizations should be different.



Our Method

- **Generate** a candidate lexicalization *w* for each concept by **translating a seed word.**
- **Filter** using our four-step procedure:

  1. Multi-word filter: If *w* is not a single word (e.g. "male cousin"), **return GAP**

  2. Horizontal filter (Proposition 2): If *w* was also generated for a sibling node of *s*, **return GAP**

  3. Back-translation filter: If back-translating *w* does not recover the seed word, **return GAP**

  4. Vertical filter (Proposition 1): If *w* was also generated for a parent node of *s*, **and** another child of that parent node has already been tagged as a **GAP**, then **return GAP**

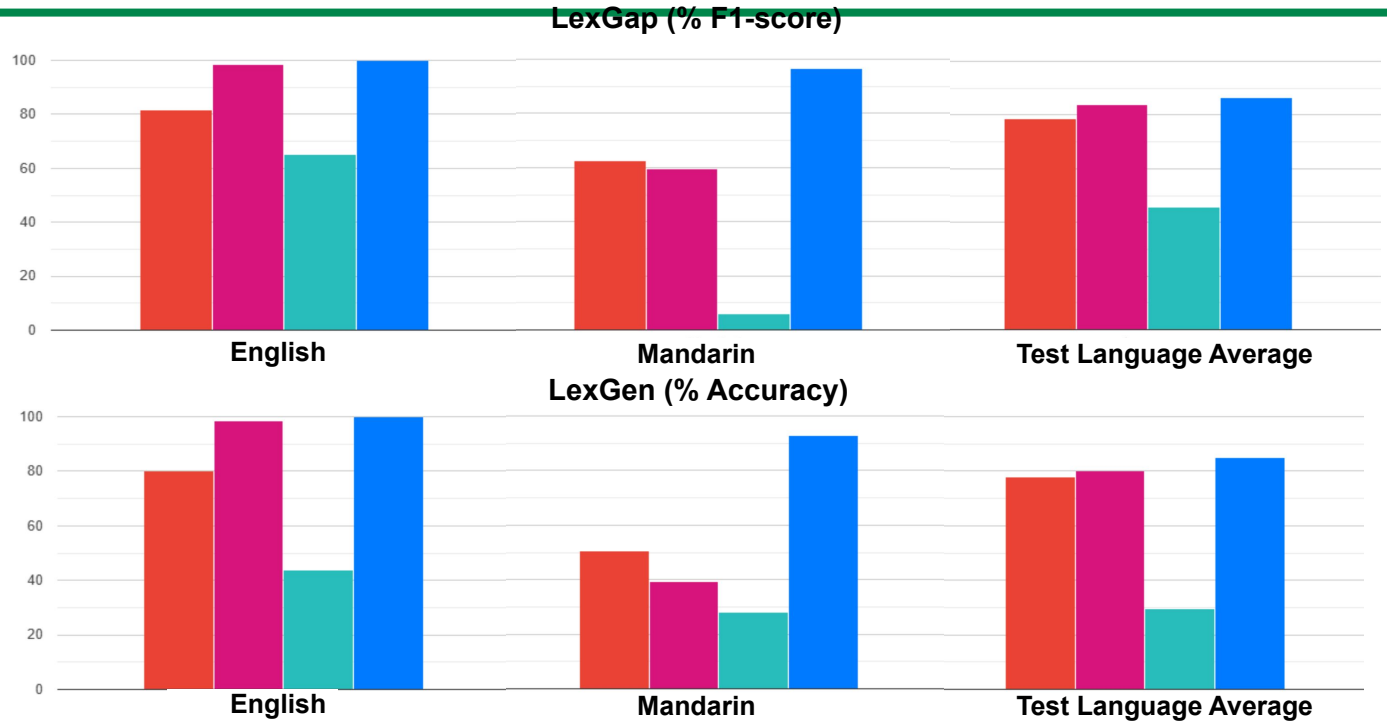
- If *w* makes it past the filters, return *w* for LexGen, **False** for LexGap

Experimental Setup

- **Data:** *Database of Lexical Diversity in Kinship* by Khishigsuren et al. (2022)
- **Translator:** Google Translate
- **Metrics:** Accuracy for LexGen, F1 score for LexGap
- **Comparison:** All-Gaps, BabelNet 5.1, and ChatGPT w/ GPT-3.5 Turbo
- **Languages**
  - Development languages: English, Mandarin, and Persian.
  - Test languages: Spanish, Russian, French, German, Polish, Arabic, Italian, Mongolian, Hungarian, and Hindi.

*GPT-3.5 Turbo and Google Translate were accessed on February 15, 2024

Results


LexGap (% F1-score)
100
80
60
40
20
0
English
Mandarin
Test Language Average
LexGen (% Accuracy)
100
80
60
40
20
0
English
Mandarin
Test Language Average
All-Gaps
BabelNet
ChatGPT
Ours

Conclusion

- Novel translate-and-filter method for:
  - Generating **lexicalizations**
  - Detecting **lexical gaps**
- Grounded in linguistic theory, with clear definitions and and propositions
- Leverages translation and hypernym/hyponym relations
- Future work: Beyond kinship to other domains

github.com/UAlberta-NLP/KinshipAutoLex

Proposition 1

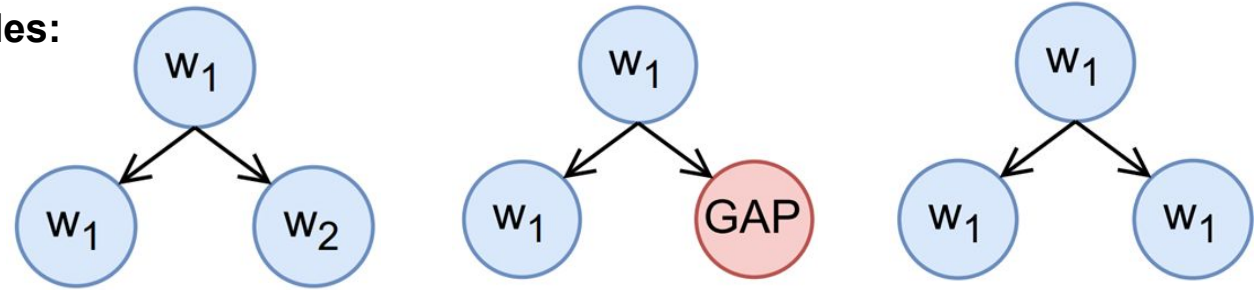
If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing P and C1 with the same word w can result in a colloquial contradiction.

Proof: C2 could be expressed by a phrase “w but not w”, This phrase intuitively corresponds to a logical contradiction: $w(x) \wedge \neg w(x)$.

Example:

Robin is my parent but not my father => Robin es mi padre pero no mi padre

Excluded triples:



*This Example was obtained from Google Translate accessed on February 15, 2024

Proposition 2

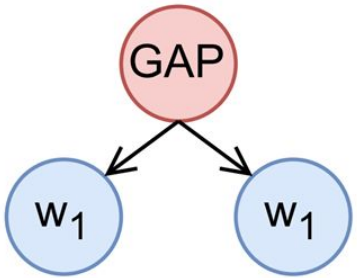
If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing C1 and C2 with the same word w can result in a colloquial contradiction.

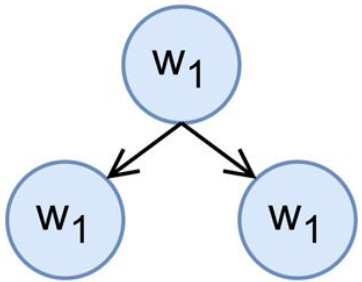
Proof: P could be expressed by a phrase “either w or w”, this phrase intuitively corresponds to a logical contradiction: $w(x) \oplus w(x)$.

Example:

Tengo una prima pero no tengo ningún primo => I have a cousin but I have no cousin

Excluded triples:



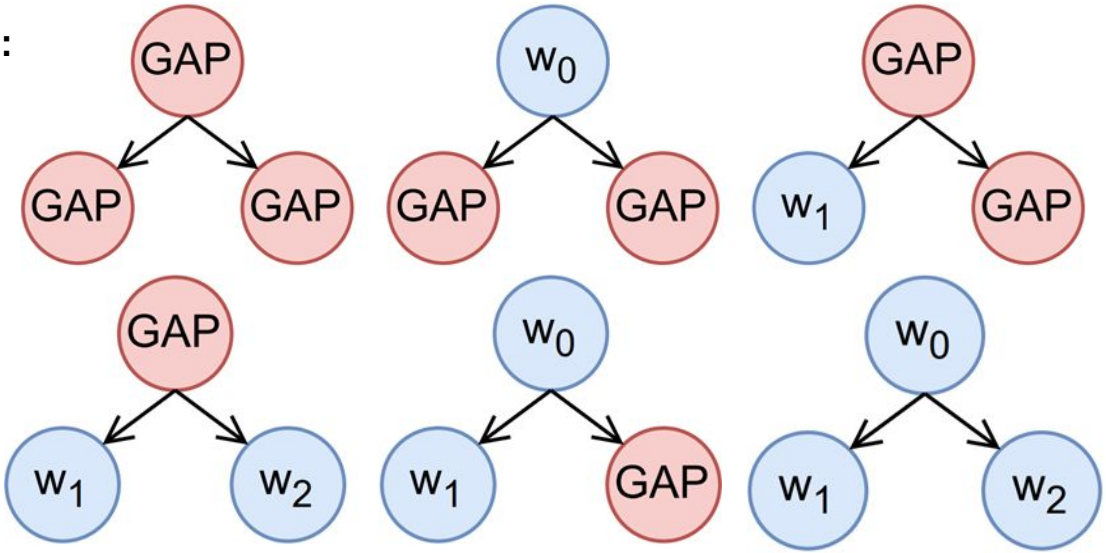


*This Example was obtained from Google Translate accessed on February 15, 2024

Corollary

If a concept P is an exclusive disjunction of its hyponyms C1 and C2 then all their lexicalizations should be different.

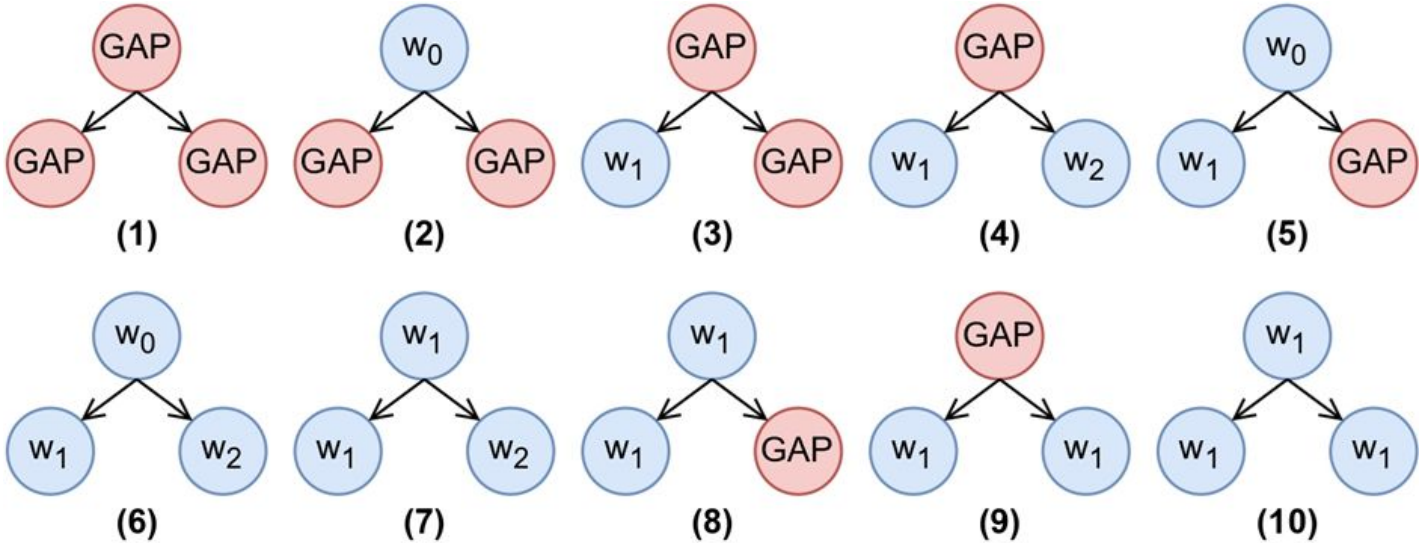
Remaining triples:



Disjunctive Triples

Kinship concepts can often be arranged into triples

Concept sp is an exclusive disjunction of hyponym concepts s1 and s2.



Our Method

Generate a candidate lexicalization for each concept by translating a seed word into the target language in the context of the concept gloss. Then Apply 4 filters sequentially to the obtained translations.

- **Multi-word filter**
  - **for** each concept s **do** $L_1(s) \leftarrow$ GAP **if** $L_0(s)$ is not a word
- **Horizontal filter (backboned by proposition 2)**
  - **for** each triple (s0, s1, s2) **do** $L_2(s1) \leftarrow$ GAP; $L_2(s2) \leftarrow$ GAP **if** $L_1(s1) = L_1(s2)$
- **Back-translation filter**
  - **for** each concept s **do** $L_3(s) \leftarrow$ GAP **if** BackTrans($L_2(s)$, gloss(s)) $\neq$ seed(s)
- **Vertical filter (backboned by proposition 1)**
  - **for** each triple (s0, s1, s2) **if** $L_3(s0) = L_3(s1)$ **then**
    if $L_3(s2)$ = GAP **then** $L_4(s1) \leftarrow$ GAP **else** $L_4(s0) \leftarrow$ GAP