

Translation-based Lexicalization Generation and Lexical Gap Detection:



Application to Kinship Terms

Senyu Li, Bradley Hauer, Ning Shi, Grzegorz Kondrak
{senyu, bmhauer, ning.shi, gkondrak}@ualberta.ca



Task Definition: LexGen and LexGap

Lexicalization Generation (LexGen):

Given a language L and a concept s , $\text{LexGen}(L, s)$ returns a word in L which lexicalizes s , or a special GAP token indicating that no such word exists.

Lexical Gap Detection (LexGap):

Given a language L and a concept s , $\text{LexGap}(L, s)$ returns TRUE if L has no word that lexicalizes s , or FALSE otherwise.

LexGap returns TRUE if and only if LexGen returns a GAP.

ChatGPT

S 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."

Google Translate

Detect language English Chinese (Simplified) Spanish

我有一个堂哥，但是没有堂姐。

I have a cousin, but no cousin.

Wǒ yǒu yīgè táng gē, dànshì méiyǒu táng jiě.

14 / 5,000 拼

Concepts, Lexicalizations and Lexical Gaps

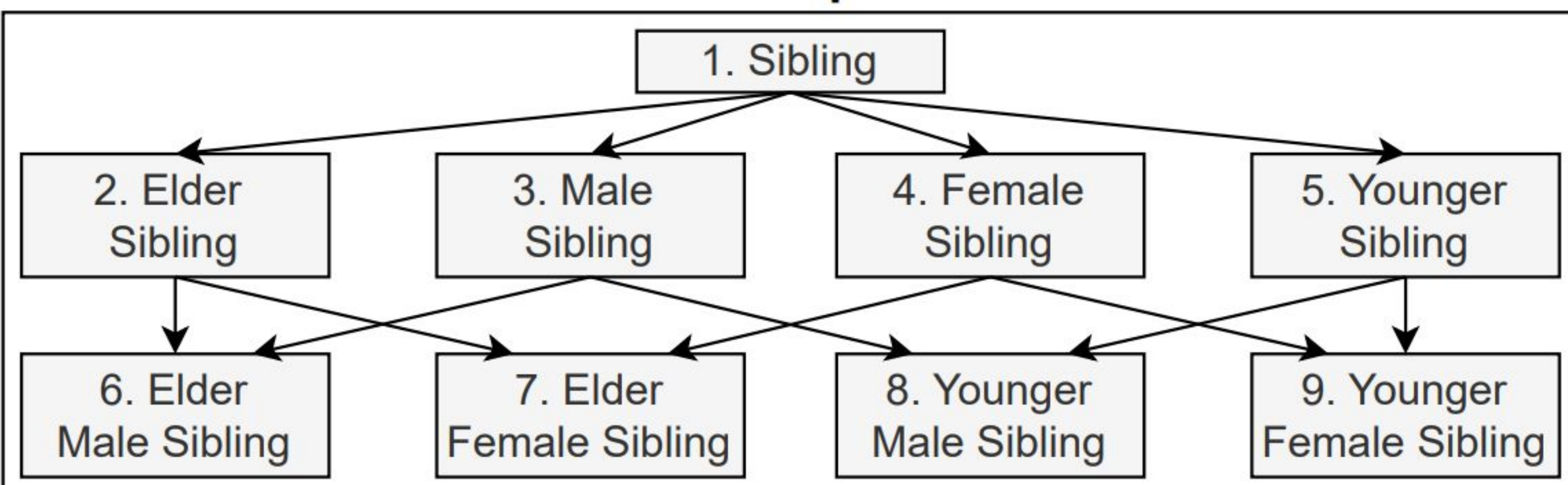
Concept: discrete word meaning

Kinship concepts have clear definitions and hierarchical structure

Lexicalization: a single word that can express (i.e. lexicalize) a concept.

Lexical Gap: a concept that has no lexicalization in a given language.

Concepts



Lexicalizations

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

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

Method

for each concept s do

$L_0(s) \leftarrow \text{Translate}(\text{seed}(s), \text{gloss}(s))$

for each concept s do

▷ **Multi-Word Filter #1**

$L_1(s) \leftarrow \text{GAP}$ if $L_0(s)$ is not a word

for each triple (s_0, s_1, s_2) do

▷ **Horizontal Filter #2**

$L_2(s_1) \leftarrow \text{GAP}$; $L_2(s_2) \leftarrow \text{GAP}$ if $L_1(s_1) = L_1(s_2)$

for each concept s do

▷ **Back-Translation Filter #3**

$L_3(s) \leftarrow \text{GAP}$ if $\text{BackTrans}(L_2(s), \text{gloss}(s)) \neq \text{seed}(s)$

for each triple (s_0, s_1, s_2) do

▷ **Vertical Filter #4**

if $L_3(s_0) = L_3(s_1)$ then

if $L_3(s_2) = \text{GAP}$ then $L_4(s_1) \leftarrow \text{GAP}$ else $L_4(s_0) \leftarrow \text{GAP}$

Propositions and Corollary

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 \Rightarrow Robin es mi padre pero no mi padre

Excluded triples: 7, 8, 10 (See figure below.)

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:

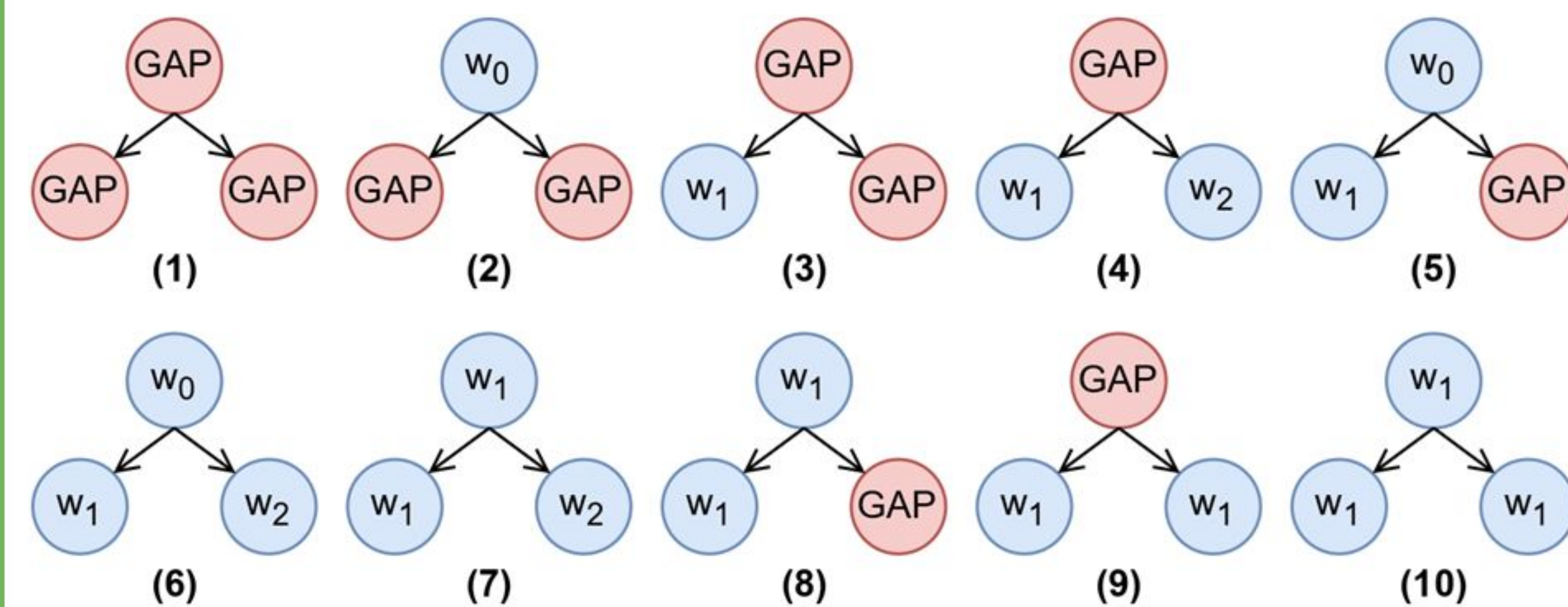
Tengo una prima pero no tengo ningún primo \Rightarrow I have a cousin but I have no cousin

Excluded triples: 9, 10

Corollary

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

Remaining triples: 1, 2, 3, 4, 5, 6



Experiment Setup and Results

Data: Database of Lexical Diversity in Kinship by Khishigsuren et al. (2022).

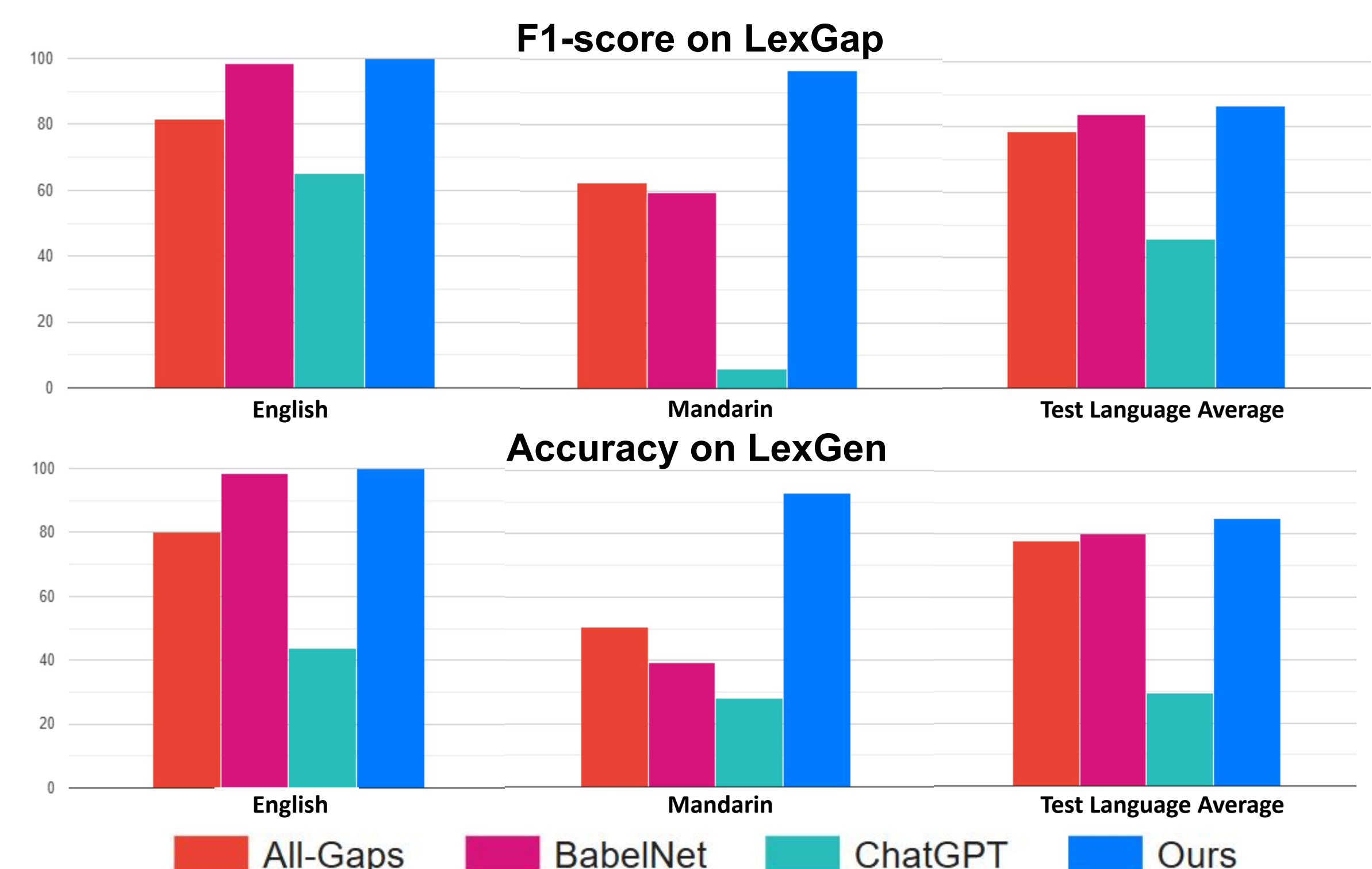
Translator: Google Translate, accessed February 15, 2024.

Metrics: Accuracy for LexGen and F1 score for LexGap.

Comparison: All-Gaps, BabelNet (v. 5.1), and ChatGPT (GPT-3.5 Turbo).

Development languages: English, Mandarin, and Persian.

Test languages: Spanish, Russian, French, German, Polish, Arabic, Italian, Mongolian, Hungarian, and Hindi.



Conclusion

- A novel translation-based method that generates concept lexicalizations and detects lexical gaps.
- Our method is grounded in formal definitions and propositions, and leverages translation and hypernym/hyponym taxonomy relations.
- Future work:
 - Apply our method to other domains
 - Employ large language models

github.com/UAlberta-NLP/KinshipAutoLex