Cross-Lingual Word Embeddings for Extremely Low-Resource Languages: Improving Bilingual Lexicon Induction for Hiligaynon

Mary Ann Tan
Dario Stojanovski
Alexander Fraser

Center for Information and Language Processing (CIS)
Ludwig Maximilian University of Munich (LMU)

BACKGROUND

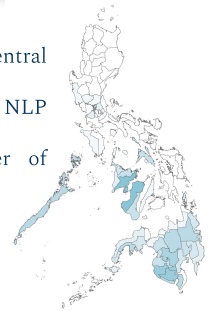
Cross-Lingual Word Embeddings (CLWEs)
- represents words from two (2) or more languages in a shared embedding space.
- 👇 allows retrieval of translation pairs through **Bilingual Lexicon Induction (BLI)**

Why CLWEs/BLI 💡👉 Machine Translation?
✓ low supervision
✓ no need for sentence-aligned parallel corpora
✓ data requirements:
- monolingual corpora
- small seed dictionary

However, most studies on CLWEs focus on:
- homogeneous language pairs (i.e. European Languages)
- moderate to high resource languages (>100M tokens)

HILIGAYNON

Why Hiligaynon?
- Austronesian language spoken in Central and Southern Philippines.
- ~10M native speakers, but dearth of NLP resources.
- Top 95 in the world by number of speakers.
- Only one (1) published corpus with ~250k tokens.



Challenges:
⏬ Extremely low-resource.
🖊 No officially recognized writing convention.
🤔 Complex morphological processes.
👎 Sociopolitical and cultural sentiments hinder proliferation of available resources.
❓ Previous study achieved 0% Precision @1.

METHODS

STRATEGY

CORPUS
Increase the size of the existing monolingual corpus.

SEED LEXICON
Determine the optimal seed lexicon, in terms of content and size.

CLWE METHODS
Find a CLWE approach applicable to a low-resource setting.

TRANSFER LEARNING
Leverage linguistically similar and/or high-resource languages.

Ultimate Goal:
Improve the BLI performance of EN-HIL CLWEs.

1 Increased corpus size.

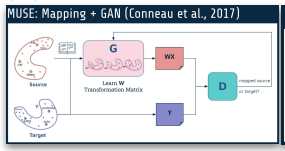
Source	Sentences	Tokens
PALITO	4,895	168,391
Bible (APSD)	56,270	818,209
Blogs	3,393	61,851
KVED	17,282	72,364
Total	**81,840**	**1,120,815**

2 Well-curated seed lexicon.

✓ most frequent target words.
✓ many2many lexical pairs.

3 Adapted applicable methods.

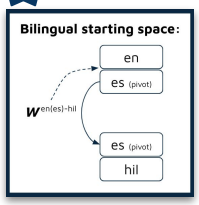
MUSE: Mapping + GAN (Conneau et al., 2017)



Pseudo-Multilingual + Joint (Duong et al., 2016)

Input: Monolingual corpora + bilingual lexicon with polysemy
1. Concatenate monolingual corpora.
2. With CBOW:
   a. Replace the center source word with its target translation from the bilingual lexicon.
   b. Select the translation that maximizes the cosine similarity between the target word embedding, and the sum of the embeddings of the source word vector and its context words.

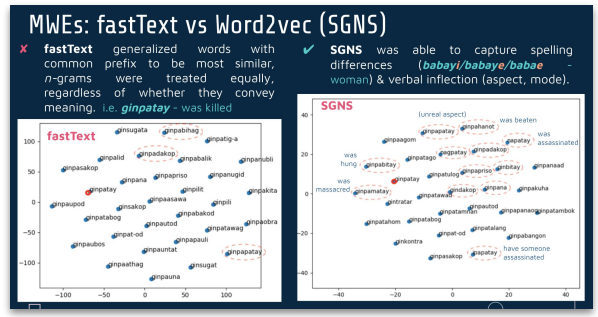
4 Transfer learning: leverage high-resource languages.



	Linguistic Similarity	Resources[1]	Remarks
Cebuano (CEB)	✓	✓	Closest language to HIL with NLP resources.
Filipino (TL)	✓	✓	PH language with the most number of NLP resources
Spanish (ES)	Loan Words	✓	Numerous loan words in HIL
Indonesian (ID)	?	✓	Austronesian language with the largest number of fastText word embeddings

EXPERIMENTS & RESULTS

1. MWEs: fastText vs SGNS



2. Seed Lexicon

Description	P@1 (%)	P@5 (%)	P@10 (%)
Baseline[2]	4.00	8.50	11.00
High Coverage	-2.30	-4.06	-4.85
Common Target Words*	+4.38	+9.10	+12.46
MUSE + HIL-Common	+2.23	+4.29	+6.00
Best	**8.38**	**17.60**	**23.46**

3. Transfer learning with pivot languages.

Src-Pivot-Tgt	P@1 (%)	P@5 (%)	P@10 (%)
EN-HIL[1]	8.38	17.60	23.46
Pivot on ES[2]	+0.88	+1.42%	+1.62
Pivot on TL[2]	-4.68	-5.98	-7.97
Pivot on ID[2]	-3.50	-4.13	-2.75

CONCLUSION

✅ Success of transfer learning is dependent on the quality of the source and target monolingual word embeddings.

✅ Well-curated seed lexicon for training improves retrieval.

Target Languages	Corpus	Word Vectors	P@1
EN-ES-HIL	1.2M tokens	9.7k (SGNS)	9.26%
EN-TL	Wikipedia Dump	66.6k (fastText)	15.93%