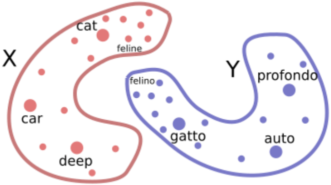
# Aligning Word Vectors on Low-Resource Languages with Wiktionary

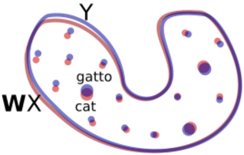
by **Mike Izbicki** (Claremont McKenna College, USA)

# Background (1): What are aligned word vectors?

First train word embeddings in multiple languages:



Then map them to a common space:



*Images from MUSE (Conneau et al., 2017).*

# Background (2): Applications of aligned word embeddings

- Transfer learning between languages
- Machine translation

# Background (2): Applications of aligned word embeddings

- Transfer learning between languages

- Machine translation

  - **Bilingual Lexicon Induction (BLI)**

    Given a word in a source language (ko):

    안녕하세요 (annyeonghaseyo)

    Find the "closest word" in the target language (en):

    hello, hi, how's it going?, wassup?

# Background (3): Lots of papers study BLI

For example:

Abdulrahim (2019), Adams et al. (2017), Ahmad et al. (2018), Alabi et al. (2020), Alaux et al. (2018), Aldarmaki and Diab (2019), Anastasopoulos and Neubig (2019), Artetxe et al. (2017b), Artetxe et al. (2017a), Artetxe et al. (2018c), Artetxe et al. (2018a), Artetxe et al. (2018b), Artetxe et al. (2020), Burdick et al. (2021), Chen and Cardie (2018), Chen and Basirat (2020), Chen and Basirat (2020), Chimalamarri et al. (2020), Chimalamarri et al. (2020), Choe et al. (2019), Conneau et al. (2017), Di Gangi and Federico (2017), Ding and Duh (2018), Dinu and Baroni (2014), Dyevre (2021), Font and Costa-Jussa (2019), Gennaro and Ash (2022), Glavaš et al. (2019), Gordon et al. (2020), Grave et al. (2018), Gupta and Jaggi (2021), Heyman and Heyman (2019), Heyman et al. (2019), Indukaev (2021), Joshi et al. (2019), Joulin et al. (2018), Kementchedjhieva et al. (2019), Kim et al. (2018), Klementiev et al. (2012), Marchisio et al. (2020), Mikolov et al. (2013), Mikolov et al. (2018), Mogadala and Rettinger (2016), Neishi et al. (2017), Ormazabal et al. (2019), Li et al. (2018), Qi et al. (2018), Rheault and Cochrane (2020), Rodriguez and Spirling (2022), Schuster et al. (2019) , Sert et al. (2021), Stringham and Izbicki (2020), Strubell et al. (2019), Vulić and Moens (2015), Vulić et al. (2019), Vulić et al. (2020), Vulić et al. (2020), Wang et al. (2020), Xia et al. (2019), Xiao and Guo (2014), Xing et al. (2015), Yang et al. (2019), Zhang et al. (2017), Zhang et al. (2019), Zhao et al. (2020)

# Problem: Existing BLI datasets are machine generated

- This results in weird artifacts:

| Thai "Word" | English "Translation" | |
|---|---|---|
| เนูเธ้เธศเธญเธฆเธตเน่ | calories | |
| เนฺเธ้เธกเธศเธญเธข | lanterns | |
| annie | annie | proper nouns |
| bdfutbol | bdfutbol | HTML/code artifacts |
| getparent | getparent | |
| roca | roca | not a word |

- Quality varies tremendously between languages (EN-ES pretty good... EN-TH pretty bad... others??? )

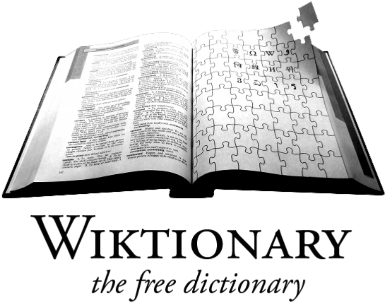
- Kementchedjhieva et al. (2019) suggest that future research *"avoids drawing conclusions from quantitative results on [the MUSE] BLI dataset."*

# Solution: Create BLI datasets with Wiktionary!

- "Wikipedia for dictionaries"
- Each entry contains:
  - word
  - language
  - part of speech
  - English-language translation
  - ...

  Crowd sourced data

- 1.8 million entries in 4204 languages



# Limitations of Wiktionary (1)

- Most languages have few entries

  298/4204 languages “good enough” for BLI.

  Generate test sets with the following POS splits

| Part of Speech | Number of Words |
|---|---|
| Adjective | 50 |
| Adverb | 25 |
| Noun | 125 |
| Verb | 50 |
| Total | 250 |

not proper nouns

- For languages with larger vocabulary, larger test sets are created

# Limitations of Wiktionary (2)

- Wiktionary dataset focuses on "dictionary forms" of words

  Korean dictionary word 가다 (gada = "to go") in the dataset

  Conjugated forms not in dictionary include:
  가, 가요, 가자, 가겠어, 가겠어요, 가겠습니다, 갑니다, 갑니까, 갑시다, 갔다, 갔어, 갔어요, 갔느냐, 갔습니다, 갔습니까

  Korean is agglutinative language, so MANY conjugated forms

- Some languages (like Spanish) DO include conjugated forms

# Results (1)

Despite limitations, Wiktionary data still better:

| Wiktionary Dataset | Previous Datasets |
|---|---|
| 298 Languages | $\leq$ 45 languages |
| Human Translations | Machine Translations |
| Has POS tags | No POS tags |

# Results (2)

Align the 157 word embeddings provided by Grave et al. (2018)

- largest set of aligned word embeddings to-date

15 previously unstudied languages have “high” BLI accuracy ($> 30\%$):

Armenian (39.15)
Austurian (36.92)
Azerbaijani (37.38)
Basque (36.32)
Belarusian (35.75)
Esperanto (50.00)
Galician (46.62)
Georgian (37.30)
Malayalam (33.62)
Mongolian (31.38)
Norwegian Nynorsk (32.35)
Serbian (30.76)
Serbo-Croatian (33.17)
Urdu (37.08)
Welsh (34.84)

# Results (3): Languages Presented about Today

| | | | BLI Accuracy | |
|---|---|---|---|---|
| Presenter | Language | Dataset Size | (Wiktionary) | (MUSE) |
| Everlyn | Swahili | 6134 | **18.01** | – |
| | Luhya | 35 | – | – |
| Mohaddeseh | Persian | 10907 | **39.40** | 37.39 |
| Anna | Ket | 75 | – | – |
| | Chikchi | 65 | – | – |
| | Ludic | 404 | – | – |
| | Karelian | 735 | – | – |
| | Selkup | 12 | – | – |
| | Evenki | 512 | – | – |
| | Veps | 2012 | – | – |
| Nathaniel | Jamaican | 258 | – | – |
| | Haitin Creole | 1278 | – | – |
| Vasile | Romanian | 67121 | 48.58 | **48.96** |
| Alberto | Indonesian | 15015 | **40.15** | 35.20 |
| | Malay | 5989 | **28.56** | 27.60 |
| Shivam | Bengali | 4720 | 26.68 | **28.34** |
| | Gujarati | 3284 | **16.81** | – |
| | Hindi | 14234 | **38.28** | 33.99 |
| | Marathay | 2080 | **19.82** | – |
| | Tamil | 5357 | 21.20 | **29.11** |
| Jenn | Cebuano | 13176 | **8.22** | – |
| | Tagalog | 15015 | **30.14** | 28.24 |

# Takeaways

- All code/data open source at

  `https://github.com/mikeizbicki/wiktionary_bli`

- Aligned word vectors might now be useful in YOUR language.
- Wiktionary can help you translate.
- You can help Wiktionary !?!?

# Bibliography I

Abdul Z Abdulrahim. 2019. Ideological drifts in the US constitution: Detecting areas of contention with models of semantic change. In *NeurIPS Joint Workshop on AI for Social Good, Vancouver, Canada*.

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 937–947.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. *arXiv preprint arXiv:1811.00570* (2018).

# Bibliography II

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 2754–2762.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings. *arXiv preprint arXiv:1811.01124* (2018).

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. *arXiv preprint arXiv:1903.03243* (2019).

Antonios Anastasopoulos and Graham Neubig. 2019. Should All Cross-Lingual Embeddings Speak English? *arXiv preprint arXiv:1911.03058* (2019).

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017b. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 451–462.

# Bibliography III

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. *arXiv preprint arXiv:1809.01272* (2018).

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017a. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041* (2017).

Mikel Artetxe, Gorka Labaka, Inigo Lopez-Gazpio, and Eneko Agirre. 2018c. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018)*. Association for Computational Linguistics, Brussels, Belgium.

# Bibliography IV

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. *arXiv preprint arXiv:2004.14958* (2020).

Laura Burdick, Jonathan K Kummerfeld, and Rada Mihalcea. 2021. Analyzing the Surprising Variability in Word Embedding Stability Across Languages. *EMNLP* (2021).

Shifei Chen and Ali Basirat. 2020. Cross-lingual Word Embeddings beyond Zero-shot Machine Translation. *Swedish Language Technology Conference (SLTC-2020)* (2020).

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. *arXiv preprint arXiv:1808.08933* (2018).

Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological segmentation to improve crosslingual word embeddings for low resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 5 (2020), 1–15.

# Bibliography V

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2019. word2word: A collection of bilingual lexicons for 3,564 language pairs. *arXiv preprint arXiv:1911.12019* (2019).

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).

Mattia Antonino Di Gangi and Marcello Federico. 2017. Monolingual embeddings for low resourced neural machine translation. In *Proceedings of the 14th International Conference on Spoken Language Translation*. 97–104.

Shuoyang Ding and Kevin Duh. 2018. How Do Source-side Monolingual Word Embeddings Impact Neural Machine Translation? *arXiv preprint arXiv:1806.01515* (2018).

# Bibliography VI

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 624–633.

Arthur Dyevre. 2021. The promise and pitfall of automated text-scaling techniques for the analysis of jurisprudential change. *Artificial Intelligence and Law* 29, 2 (2021), 239–269.

Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).

Gloria Gennaro and Elliott Ash. 2022. Emotion and reason in political language. *The Economic Journal* 132, 643 (2022), 1037–1059.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508* (2019).

# Bibliography VII

Joshua Gordon, Marzieh Babaeianjelodar, and Jeanna Matthews. 2020. Studying political bias via word embeddings. In *Companion Proceedings of the Web Conference 2020*. 760–764.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. *arXiv preprint arXiv:2106.04302* (2021).

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. *NAACL-HLT* (2019).

Tom Heyman and Geert Heyman. 2019. Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology* 72, 8 (2019), 2084–2109.

# Bibliography VIII

Andrey Indukaev. 2021. Studying ideational change in Russian politics with topic models and word embeddings. In *The Palgrave Handbook of Digital Russia Studies*. Palgrave Macmillan, Cham, 443–464.

Ishani Joshi, Purvi Koringa, and Suman Mitra. 2019. Word embeddings in low resource Gujarati language. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 5. IEEE, 110–115.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. *EMNLP/IJCNLP* (2019).

# Bibliography IX

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. *Empirical Methods in Natural Language Processing (EMNLP)* (2018).

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. 1459–1474.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 138–143. http://aclweb.org/anthology/P18-2023

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *WMT* (2020).

# Bibliography X

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013).

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 692–702.

# Bibliography XI

Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. 99–109.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. *arXiv preprint arXiv:1906.05407* (2019).

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323* (2018).

Ludovic Rheault and Christopher Cochrane. 2020. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis* 28, 1 (2020), 112–133.

# Bibliography XII

Pedro L Rodriguez and Arthur Spirling. 2022. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics* 84, 1 (2022), 101–115.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492* (2019).

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2021. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review* (2021), 08944393211010398.

Nathan Stringham and Mike Izbicki. 2020. Evaluating word embeddings on low-resource languages. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 176–186.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

# Bibliography XIII

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? *EMNLP-IJCNLP* (2019).

Ivan Vulić, Anna Korhonen, Goran Glavaš, and others. 2020. Improving bilingual lexicon induction with unsupervised post-processing of monolingual word vector spaces. *Workshop on Representation Learning for NLP (RepL4NLP)* (2020).

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 363–372.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070* (2020).

Shirui Wang, Wenan Zhou, and Chao Jiang. 2020. A survey of word embeddings based on deep learning. *Computing* 102, 3 (2020), 717–740.

# Bibliography XIV

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785* (2019).

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. 119–129.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1006–1011.

Wei Yang, Wei Lu, and Vincent W Zheng. 2019. A simple regularization-based algorithm for learning cross-domain word embeddings. *arXiv preprint arXiv:1902.00184* (2019).

# Bibliography XV

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1959–1970.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are Girls Neko or Sh\= ojo? Cross-Lingual Alignment of Non-Isomorphic Embeddings with Iterative Normalization. *arXiv preprint arXiv:1906.01622* (2019).

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699* (2020).