Aligned Embeddings or Multilingual Embeddings:

A Comprehensive Study in Word Embeddings Paradigm

Kasun Wickramasinghe

Introduction

- Multilingual processing is becoming more common in most practical use cases in the present day due to the usage of code-mixed languages, and the need to process multilingual documents in a language-agnostic manner.
- Multilingual embeddings solves this problem by proving embeddings from a shared embedding space for multiple languages
- We can consider mainly two types of multilingual embeddings
  - Multilingual embedding models that supports multiple languages (mBERT, XLM-R, LaBSE, LASER etc.)[1, 2, 3, 4, 5]
  - Aligned monolingual embeddings (Word2Vec, GloVE, FastText etc.) that explicitly mapped to a common embedding space [6, 7, 8]
- Here, we comparatively study how good each of these two embedding types in the word embeddings paradigm

[1] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. EMNLP
[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ACL
[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. ACL
[4] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of ACL
[5] Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages. In Findings of EMNLP
[6] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.
[7] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal trans form for bilingual word translation. NAACL
[8] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embed ding mappings with a multi-step framework of linear transformations. AAAI
[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. arXiv preprint arXiv:1710.04087.
[10] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. EMNLP

Methodology

- The evaluation task we are using is Bilingual Lexicon Induction (BLI) or Word Translation Precision on FastText vocabularies [11]
  - How accurate the aligned embeddings can extract word translation pairs
  - Used two retrieval criteria: Nearest Neighbour (NN) and Cross-Domain Similarity Local Scaling (CSLS)
- We have used the MUSE datasets [9] for BLI evaluation
  - A set of bilingual dictionaries for 110 language pairs
- We have experimented with the following Multilingual Models
  - mBERT [1]
  - XLM-R [2]
  - LaBSE [3]
  - LASER [4, 5]
- And, the following Monolingual Alignment Techniques
  - Procrustes Analysis [9]
  - RCSLS [10]
  - VecMap [8]
- We have selected the languages that covers the following aspects
  - Resourceness of the Languages (high, medium and low)
  - Language Family (Indo-European, Sino-Tibetan, Dravidian, Japonic and Turkik)

[1] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. EMNLP
[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ACL
[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. ACL
[4] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of ACL
[5] Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages. In Findings of EMNLP
[7] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal trans form for bilingual word translation. NAACL
[8] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embed ding mappings with a multi-step framework of linear transformations. AAAI
[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. arXiv preprint arXiv:1710.04087.
[10] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. EMNLP
[11] Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. Computational Linguistics, 43(2):273–310.

Experimental Setup

- For monolingual embedding alignment we used the first 200k FastText word embeddings [12]
- For multilingual model evaluation we used the same 200k words and generated the embeddings from the multilingual models
- The BLI has been done using the reference language as English (i.e. The language pair is always in the form of En-X)

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics,

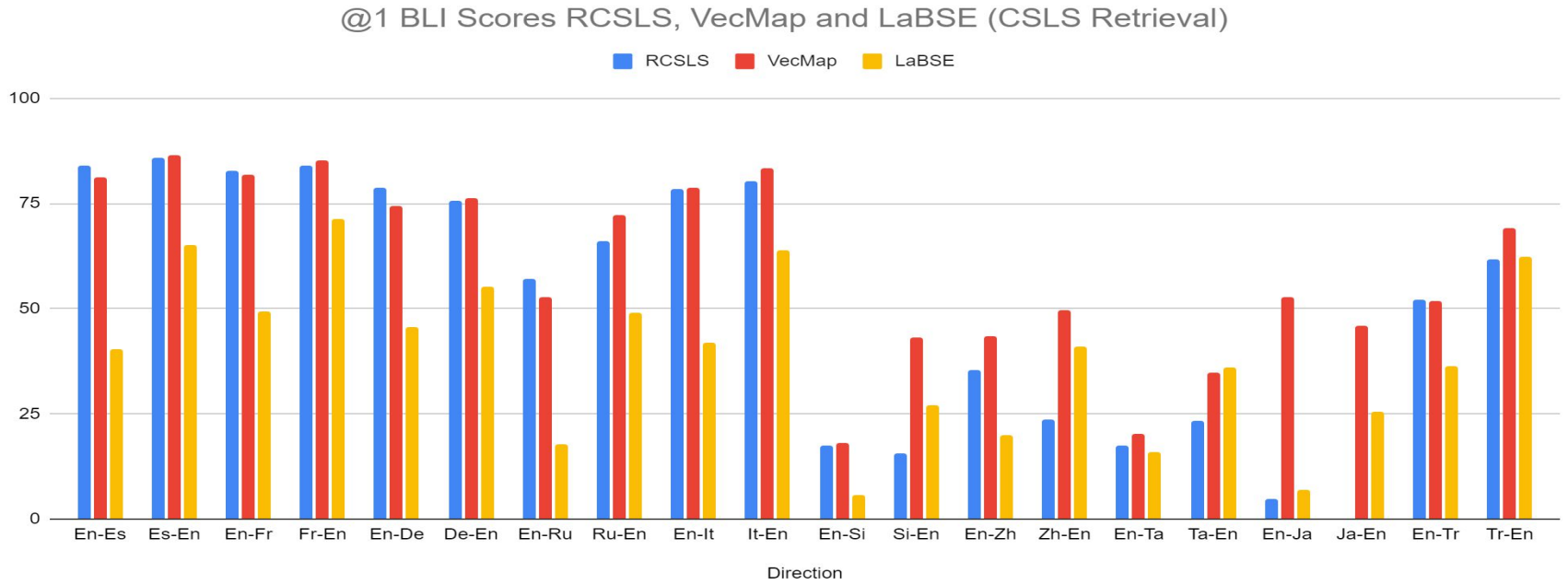
Results: Multilingual Model Comparison

- Top 1, 5 and 10 BLI scores for the selected multilingual models
- We have observed that almost all the cases LaBSE has given the best results
- The 3 languages have been selected considering resourceness and language family

Lang	Method	En-Lang						Lang-En					
		NN			CSLS			NN			CSLS		
		P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Es	mBERT	36.1	74.4	82.0	25.7	58.7	70.9	50.4	72.5	79.7	49.4	76.2	82.1
	XLM-R	37.4	78.2	83.7	28.2	62.7	73.1	50.9	71.4	76.5	61.3	78.1	80.9
	LASER2/3	37.9	75.3	82.1	39.3	75.3	82.3	57.6	78.9	85.1	56.2	78.2	84.6
	LaBSE	**40.0**	**83.1**	**89.6**	**40.5**	**84.7**	**90.2**	**65.3**	**88.9**	**92.1**	**65.0**	**87.7**	**91.5**
Zh	mBERT	23.2	49.0	56.9	20.9	46.5	54.0	12.3	29.5	36.1	28.3	43.8	49.3
	XLM-R	**30.9**	57.0	63.4	**28.8**	54.6	60.9	13.9	31.9	37.9	32.5	51.1	55.9
	LASER2/3	10.5	19.6	24.0	10.9	21.2	24.9	7.1	16.5	22.5	7.9	16.9	22.0
	LaBSE	27.1	**64.8**	**73.7**	**28.8**	**66.4**	**75.4**	**42.1**	**63.4**	**69.7**	**41.0**	**64.0**	**71.1**
Tr	mBERT	34.5	52.6	60.4	24.6	42.3	51.5	47.0	58.1	62.5	37.0	56.9	63.2
	XLM-R	35.9	62.8	69.0	28.2	50.9	59.2	50.9	62.8	66.0	45.8	60.3	64.2
	LASER2/3	35.3	57.9	64.3	32.5	50.6	57.8	56.4	67.8	70.8	45.4	65.6	70.2
	LaBSE	**36.5**	**71.1**	**78.5**	**36.3**	**74.0**	**79.9**	**64.0**	**80.9**	**84.3**	**62.4**	**80.3**	**83.6**

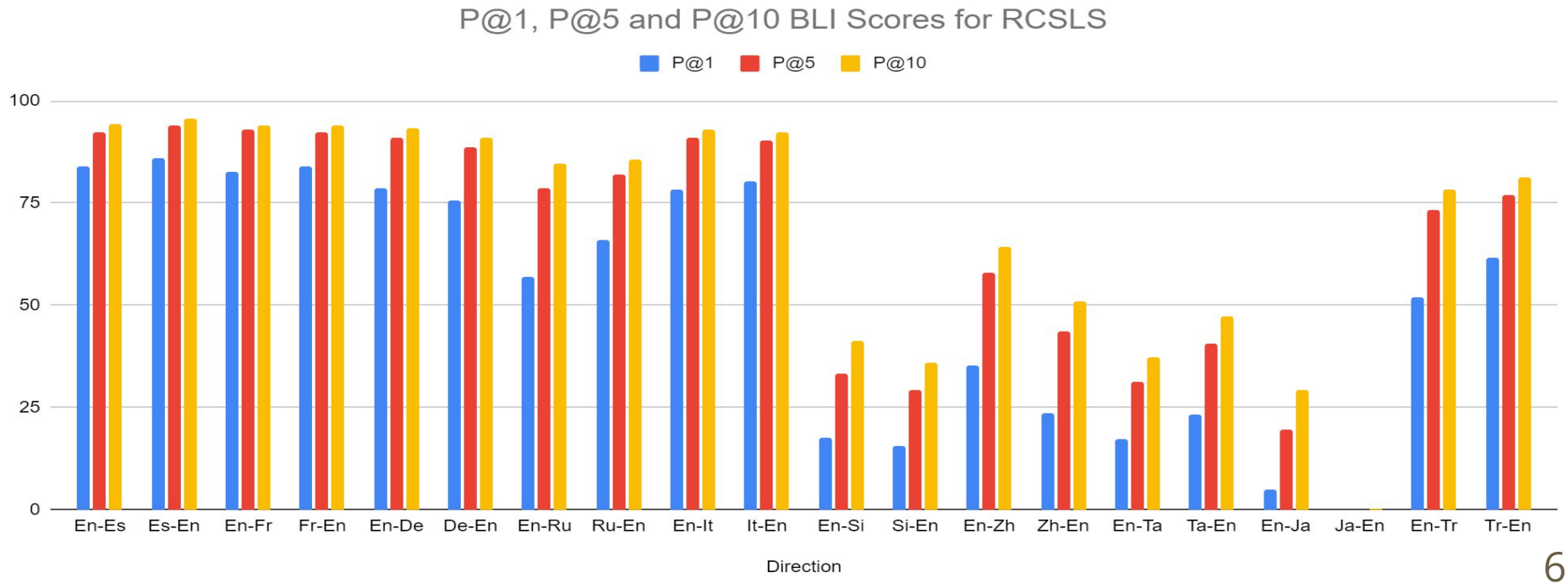
Results: Alignment Techniques Comparison

- Top-1 BLI scores for RCSLS vs VecMap vs LaBSE
- VecMap wins most of the cases and RCSLS has competitive results
- RCSLS and VecMap exceed LaBSE by a considerable margin in most of the cases



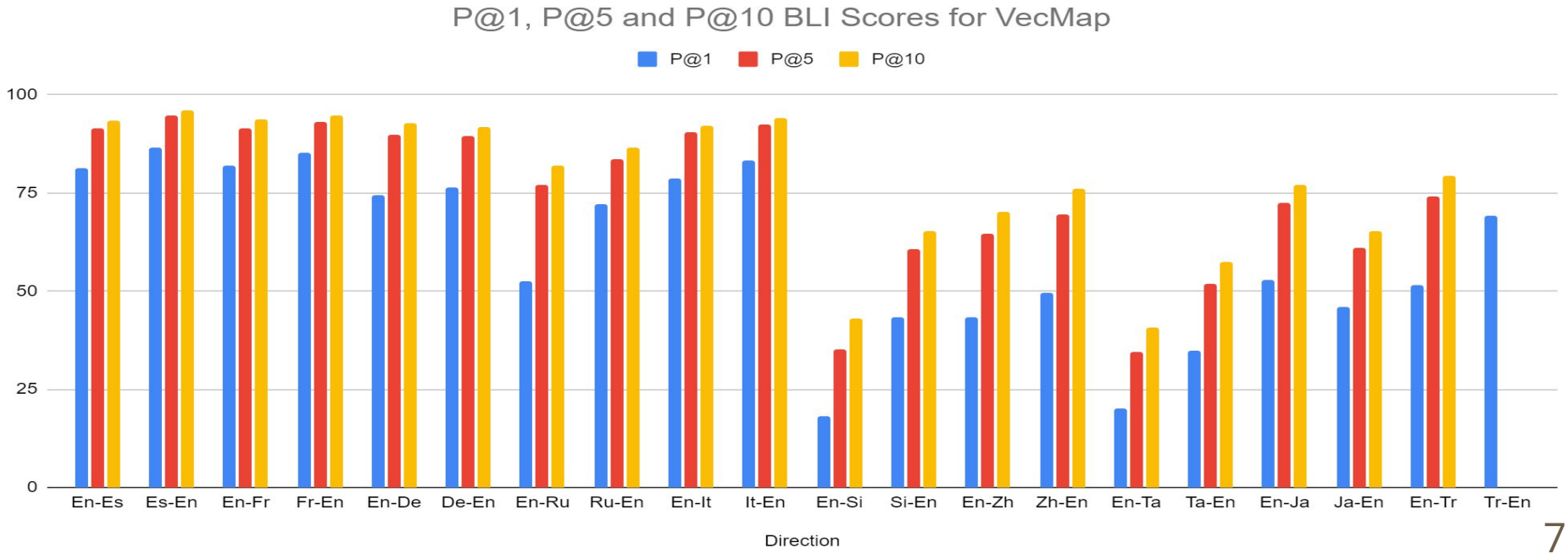
Results: RCSLS Alignment Top-N Retrieval Comparison

- RCSLS Alignment BLI scores for Top-1, 5 and 10



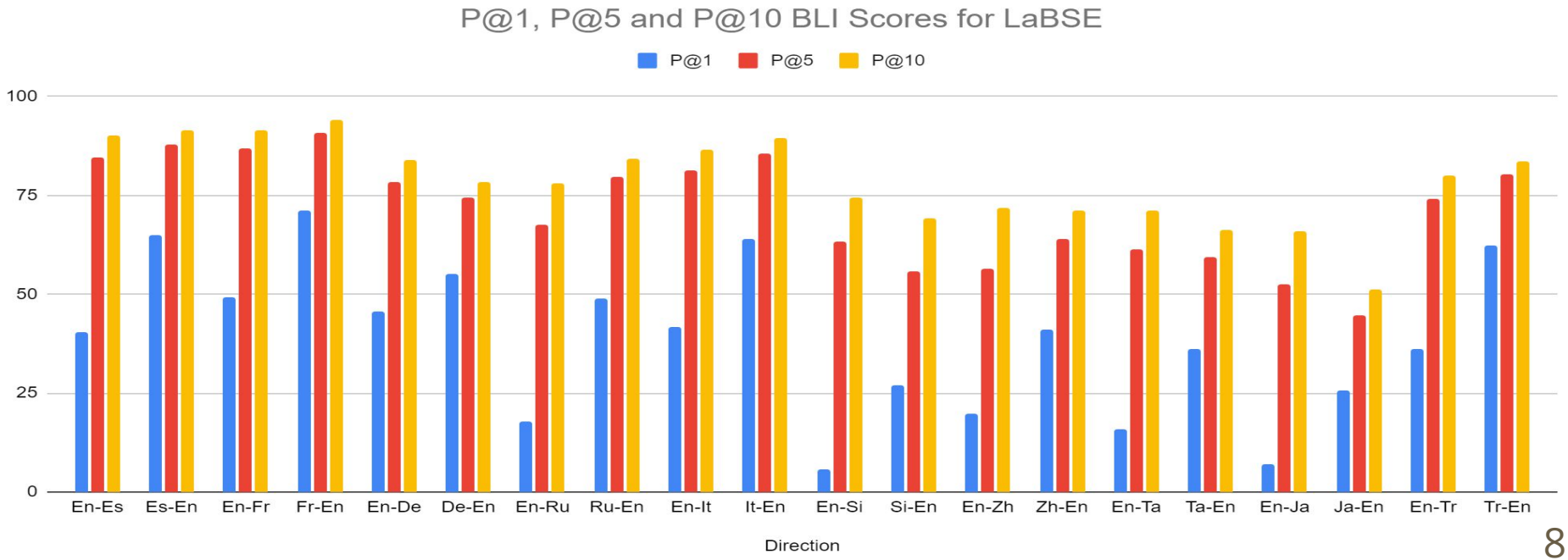
Results: VecMap Alignment Top-N Retrieval Comparison

- VecMap Alignment BLI scores for Top-1, 5 and 10



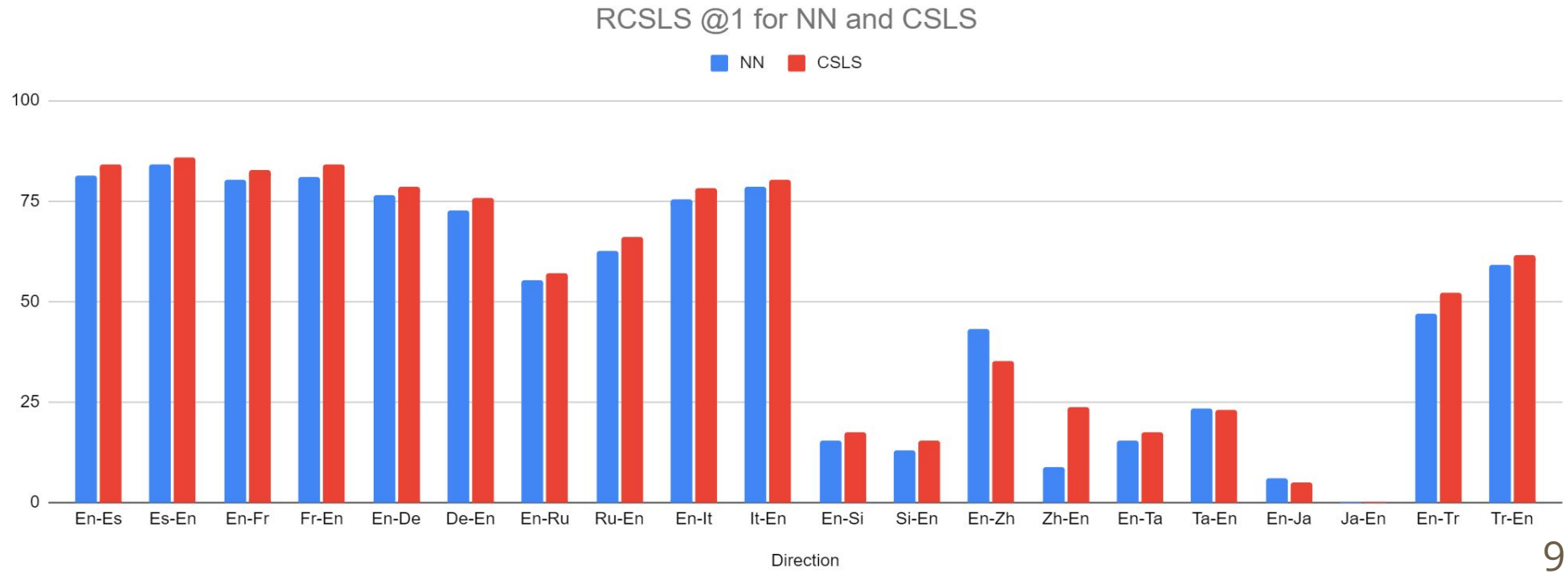
Results: LaBSE Alignment Top-N Retrieval Comparison

- LaBSE Alignment BLI scores for Top-1, 5 and 10



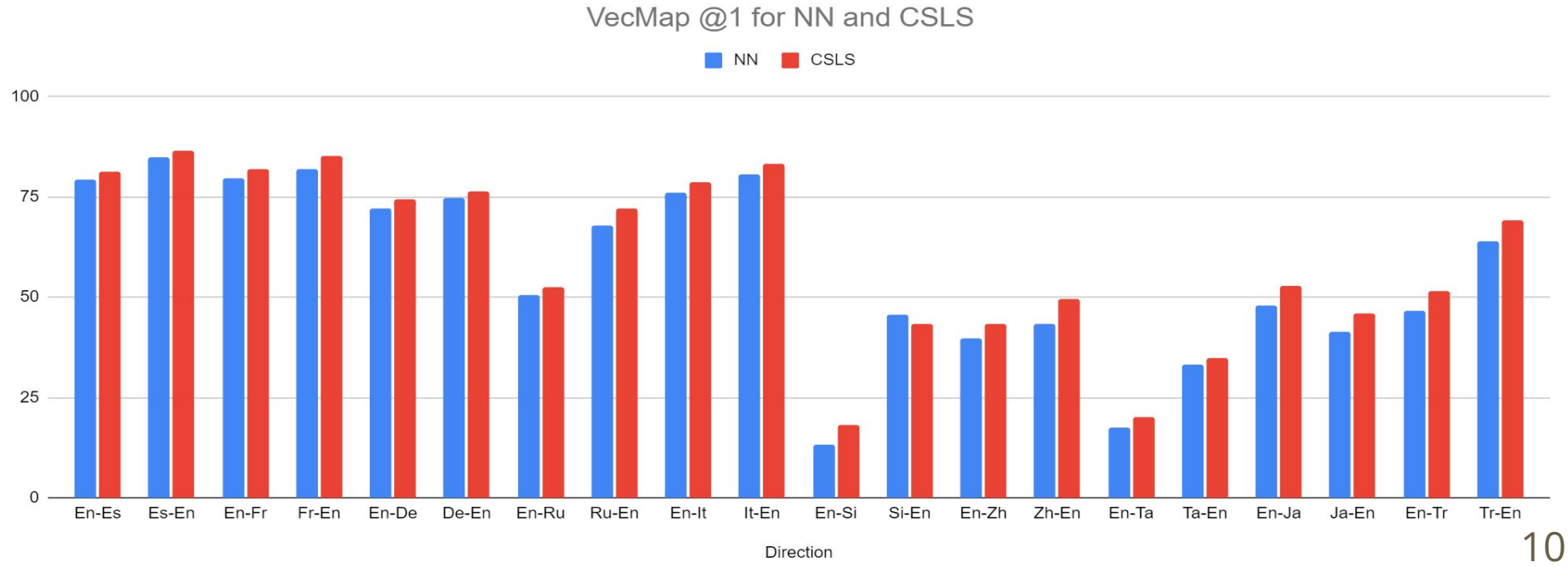
Results: RCSLS Alignment Retrieval Criteria Comparison

- Retrieval criteria evaluation for RCSLS alignment
- Almost all the cases CSLS retrieval outperforms NN retrieval



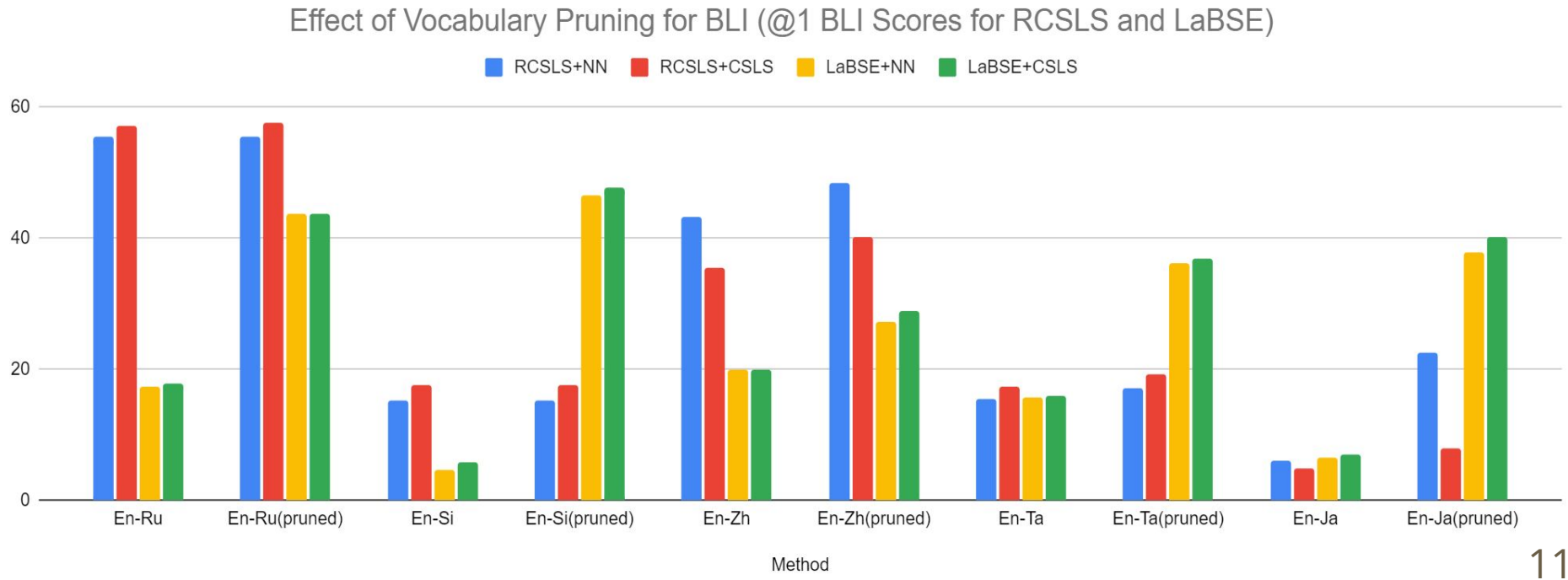
Results: VecMap Alignment Retrieval Criteria Comparison

- Retrieval criteria evaluation for VecMap alignment
- Almost all the cases CSLS retrieval outperforms NN retrieval



Results: Effect of Vocabulary Pruning for BLI

- Pruning the vocabularies by removing the entries that do not belong to that particular language when evaluating BLI



Thank You!