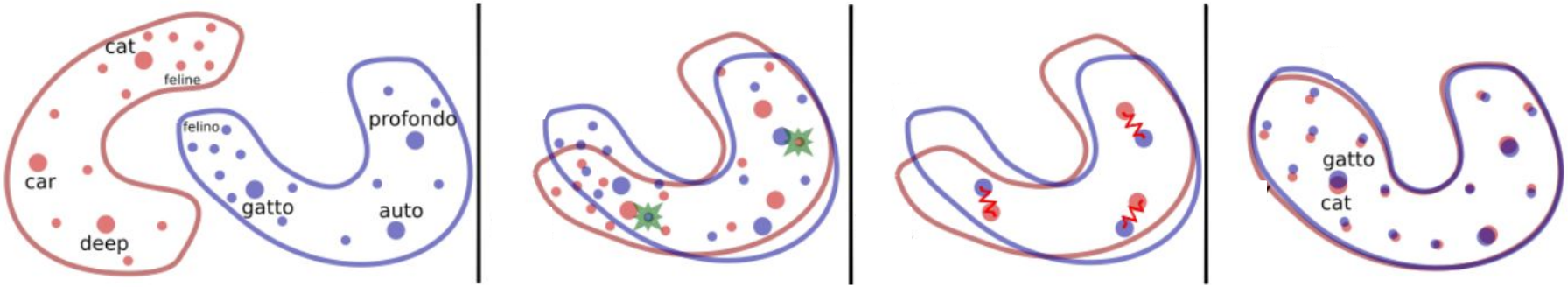
Multilingual Word Embedding Alignment for Sinhala

229405N-M.K.I.Wickramasinghe



Content

Introduction

Research Problem

Literature Survey

Progress

Introduction

Introduction

- Embeddings are the basic ingredient in many kinds of natural language processing tasks.
- In multilingual tasks unaligned embedding spaces are a huge burden. [1]
- The alignment is required for two kinds of embedding models.
    - Embedding models separately trained on monolingual data
    - Multilingual models trained on parallel multilingual data
- Multilingual model training process implicitly encourages for the alignment [2, 3, 4]
- For monolingual models, the alignment has to be done as a separate task [5, 6, 7]
- Monolingual embedding alignment is still vital since,
    - Monolingual models are lightweight
    - Can be run using simpler libraries and frameworks
    - Using multilingual models may be redundant due to supporting many languages [2, 3, 4]
    - Accuracy can be compromised due to the support of many languages in multilingual models [2]
    - The accuracy for low-resource languages can be less compared to high-resource languages due to training data imbalance in multilingual models [2]
    - Pretraining or fine-tuning a multilingual model is time and resource consuming [2,3,4]

[1] A. Kalinowski and Y. An, 'A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs', arXiv preprint arXiv:2010. 13688, 2020.
[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," arXiv preprint arXiv:2007.01852, 2020.
[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.
[5] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," arXiv preprint arXiv:1309.4168, 2013.
[6] C. Xing, D. Wang, C. Liu, and Y. Lin, "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, 2015, pp. 1006–1011.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," arXiv preprint arXiv:1804.07745, 2018.

Research Problem

Research Problem

- Monolingual word embedding models have been there for decades. [8, 9]
- Aligned word embedding models are available only for few high-resource languages[1]. [7]

The main focus of the research is to train aligned word embedding model (find the transformation matrix) between Sinhala and English languages.

  - To facilitate the above, as an intermediate goal, we shall build a Sinhala-English parallel word dataset/ dictionary
  - This will serve as an anchor dataset for Sinhala-English supervised word embedding alignment

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," arXiv preprint arXiv:1804.07745, 2018.
[1] https://fasttext.cc/docs/en/aligned-vectors.html

Literature Survey

Word Embedding Techniques

- Different vector representations for words have been there from early days and they were statistical and human crafted representations.
  - One-hot-encoding
  - Count vectorizing
  - TF-IDF [10]

- The idea of generating word embeddings without direct human interaction (complex embedding representations) was introduced in 2013 by Mikolov et al. [8] by introducing Word2Vec.

- After that two similar models were introduced,
  - GloVe [9]
  - FastText [4]

- The beauty of these new word embeddings is that the embeddings:
  - Gives a global representation of words (gives a fixed embedding for a given word) [8, 9, 11]
  - Perform word analogy arithmetic (***Paris - France + Rome = Italy*** [8])

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[10] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.
[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the association for computational linguistics, vol. 5, pp. 135–146, 2017.

Contextual Embeddings

- The meaning of a word changes according to its context (where that word occurs in a sentence and what the other words in the sentence). This is the classical sense disambiguation problem [12].
- Therefore, having a global vector representation for a word is not a good approach in cases where a context related representations are needed
- Word2Vec [8], Glove [9] and FastText [11] embeddings are global embedding representations where earlier TF-IDF [10], one-hot etc. do have some context representations but not powerful enough/
- ELMo[13] which is a deep bi-LSTM based word embedding generator is the first context based embedding model that became popular
- After the introduction of transformers [14], a revolutionary improvement happened in the contextual embedding representations, where researches could achieve state of the art accuracies and efficiencies in embedding generation.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.
[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the association for computational linguistics, vol. 5, pp. 135–146, 2017.
[12] H. T. Ng and H. B. Lee, 'Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach', arXiv preprint cmp-lg/9606032, 1996.
[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018. [Online].Available: https://arxiv.org/abs/1802.05365
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

Transformer based Embeddings

- BERT[15] is the first transformer based embedding generator which showed state of the art results at the first place.
- Then so many BERT variations were released after that; such as RoBERTa[16], ALBERT[17], ELECTRA[18] etc. where each of them showed improved results in accuracy or efficiency.
- Other than word embedding models, sentence-embedding models were also introduced as BERT extensions of which Sentence-BERT (S-BERT)[19] were the pioneer.
- After S-BERT many S-BERT variations were released and by now there are multitudes of[1] word and sentence embedding models out there that achieve better results than the initial BERT and S-BERT. [20, 21, 22]

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.
[18] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pretraining text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.
[19] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084, 2019.
[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', arXiv preprint arXiv:1910. 01108, 2019.
[21] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, 'MPNet: Masked and Permuted Pre-training for Language Understanding', in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 16857–16867.
[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, 'Xlnet: Generalized autoregressive pretraining for language understanding', Advances in neural information processing systems, vol. 32, 2019.
[1] https://www.sbert.net/docs/pretrained_models.html

Multilingual Embeddings

- The next advancement in word and sentence embeddings is having a single model for multiple languages.
- For word embeddings there are XLM [3], XLM-R [4] etc. and for sentence embeddings there are LaBSE [2] etc.
- The beauty of multilingual models is that they have a single embedding space for all the languages it supports. Thus, we can perform mathematical operations on the embeddings across languages, providing much reprieve for multilingual tasks.

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," arXiv preprint arXiv:2007.01852, 2020.
[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

Embedding Alignment

- Aligned Embeddings are vital for multilingual tasks where embeddings of multiple languages share a single embedding space so that multilingual tasks can be performed irrespective of the language.
- Mikolov et al. [5] aligned two Word2Vec word embedding spaces assuming a simple linear mapping between the two embedding spaces
- Xing et al. [6], showed that better alignment results can be achieved by assuming an orthogonal mapping between two embedding spaces.
- Joulin et al. [7] have addressed the so called hubness issue where some words appear too frequently in the neighborhoods of other words, by introducing an improved loss function for alignment called Cross-domain similarity local scaling (CSLS).
- All the above techniques are supervised alignment techniques which need to have a parallel word dictionary to decide the alignment matrix.
- Unsupervised techniques, while not as prevalent, do exist. Some are based on traditional statistical methods [23] while others are based on adversarial approaches [24].

[5] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” arXiv preprint arXiv:1309.4168, 2013.
[6] C. Xing, D. Wang, C. Liu, and Y. Lin, “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, 2015, pp. 1006–1011.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” arXiv preprint arXiv:1804.07745, 2018.
[23] E. Grave, A. Joulin, and Q. Berthet, “Unsupervised alignment of embeddings with wasserstein procrustes,” in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 1880–1890.
[24] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” arXiv preprint arXiv:1710.04087, 2017.

Alignment Datasets for Sinhala Language

- For supervised embedding alignment we need an alignment dataset which help to identify corresponding points in the two embedding spaces. These datasets are parallel datasets [7].
- For supervised ***word embedding alignment*** what we need is a parallel word dataset or a dictionary dataset [7].
- Sinhala, being a low-resource language does not have much such resources available at the moment [25].
- At the moment a dataset suited for supervised multilingual embedding alignment for Sinhala to any other language is not publicly and freely available to the best of our knowledge.
- We came across several multilingual parallel corpora that contain Sinhala as a language, such as the works by Guzmán et al. [26, 27], Hameed et al. [28], Bañón et al. [29] and Vasantharajan and Thayasivam [30] that are comprised of sentence and paragraph level parallel entries.
- They are well suited for higher-level multilingual tasks such as Machine Translation (MT) but, not for lower-level tasks such as word embedding alignment. [26, 28].

[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," arXiv preprint arXiv:1804.07745, 2018.
[25] N. de Silva, "Survey on publicly available sinhala natural language processing tools and research," arXiv preprint arXiv:1906.02358, 2019.
[26] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," arXiv preprint arXiv:1902.01382, 2019.
[27] M. R. Costa-jussà et al., 'No language left behind: Scaling human-centered machine translation', arXiv preprint arXiv:2207. 04672, 2022.
[28] R. A. Hameed, N. Pathirennehelage, A. Ihalapathirana, M. Z. Mohamed, S. Ranathunga, S. Jayasena, G. Dias, and S. Fernando, "Automatic creation of a sentence aligned sinhala-tamil parallel corpus," in Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), 2016, pp. 124–132.
[29] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn et al., "Paracrawl: Web-scale acquisition of parallel corpora," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4555–4567.
[30] C. Vasantharajan and U. Thayasivam, "Tamizhi-net ocr: Creating a quality large scale tamil-sinhala-english parallel corpus using deep learning based printed character recognition (pcr)," arXiv preprint arXiv:2109.05952, 2021.

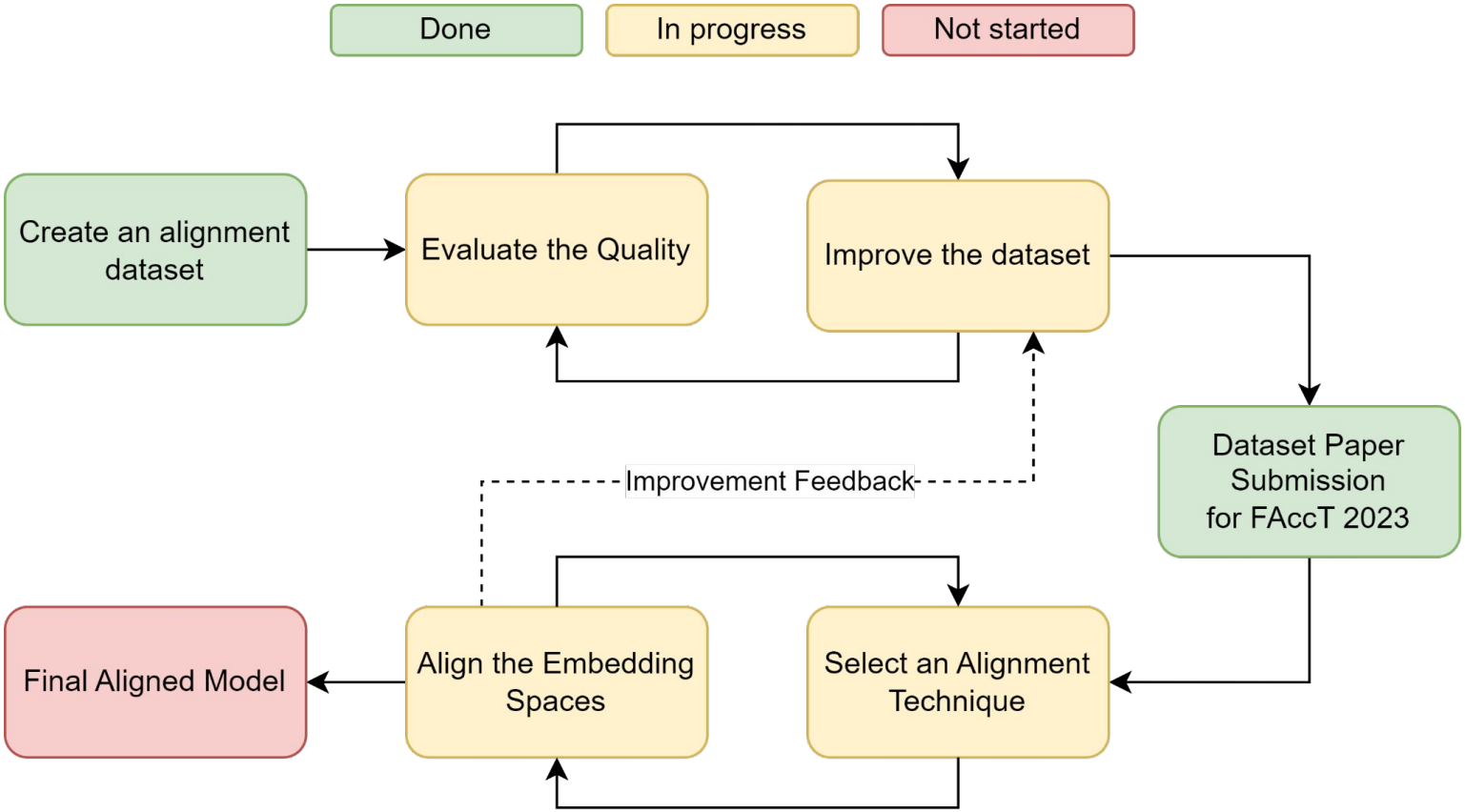
Aligned Embeddings in multilingual models

- Unlike in monolingual models, in multilingual models the embeddings get aligned in the embedding process itself.
- This is achieved by using alignment supportive training objective in the training process.
- LaBSE[2], XLM[3] and XLM-R[4] uses Translation Language Modeling (TLM) task for embedding alignment following the initial pre training tasks such as Masked Language modeling, Next Sentence Prediction etc.

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," arXiv preprint arXiv:2007.01852, 2020.
[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

Progress

Progress


Done
In progress
Not started
Create an alignment dataset
Evaluate the Quality
Improve the dataset
Dataset Paper Submission for FAccT 2023
Improvement Feedback
Final Aligned Model
Align the Embedding Spaces
Select an Alignment Technique

FAccT Paper Submission

- Conference: ACM FAccT Conference 2023
- A dataset paper
- Presents three Sinhala-English parallel word datasets
- Auxiliary task of the main research - Creating an alignment dataset for supervised word embedding alignment
- Notification of outcome: 6 April 2023

References

References

[1] A. Kalinowski and Y. An, 'A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs', arXiv preprint arXiv:2010. 13688, 2020.
[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," arXiv preprint arXiv:2007.01852, 2020.
[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.
[5] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," arXiv preprint arXiv:1309.4168, 2013.
[6] C. Xing, D. Wang, C. Liu, and Y. Lin, "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, 2015, pp. 1006–1011.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," arXiv preprint arXiv:1804.07745, 2018.
[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.
[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the association for computational linguistics, vol. 5, pp. 135–146, 2017.
[12] H. T. Ng and H. B. Lee, 'Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach', arXiv preprint cmp-lg/9606032, 1996.
[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018. [Online].Available: https://arxiv.org/abs/1802.05365
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.
[18] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pretraining text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.
[19] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084, 2019.
[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', arXiv preprint arXiv:1910. 01108, 2019.
[21] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, 'MPNet: Masked and Permuted Pre-training for Language Understanding', in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 16857–16867.

References

[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, 'Xlnet: Generalized autoregressive pretraining for language understanding', Advances in neural information processing systems, vol. 32, 2019.\

[23] E. Grave, A. Joulin, and Q. Berthet, "Unsupervised alignment of embeddings with wasserstein procrustes," in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 1880–1890.

[24] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," arXiv preprint arXiv:1710.04087, 2017.

[25] N. de Silva, "Survey on publicly available sinhala natural language processing tools and research," arXiv preprint arXiv:1906.02358, 2019.

[26] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," arXiv preprint arXiv:1902.01382, 2019.

[27] M. R. Costa-jussà et al., 'No language left behind: Scaling human-centered machine translation', arXiv preprint arXiv:2207. 04672, 2022.

[28] R. A. Hameed, N. Pathirennehelage, A. Ihalapathirana, M. Z. Mohamed, S. Ranathunga, S. Jayasena, G. Dias, and S. Fernando, "Automatic creation of a sentence aligned sinhala-tamil parallel corpus," in Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), 2016, pp. 124–132.

[29] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn et al., "Paracrawl: Web-scale acquisition of parallel corpora," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4555–4567.

[30] C. Vasantharajan and U. Thayasivam, "Tamizhi-net ocr: Creating a quality large scale tamil-sinhala-english parallel corpus using deep learning based printed character recognition (pcr)," arXiv preprint arXiv:2109.05952, 2021.

[31] A. Fernando, S. Ranathunga, D. Sachintha, L. Piyarathna, and C. Rajitha, "Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages," Knowledge and Information Systems, pp. 1–42, 2022.

[32] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote et al., "Quality at a glance:An audit of web-crawled multilingual datasets," Transactions of the Association for Computational Linguistics, vol. 10, pp. 50–72, 2022.

[33] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," arXiv preprint arXiv:1607.01759, 2016.

[34] S. Thillainathan, S. Ranathunga, and S. Jayasena, "Fine-Tuning Self-Supervised Multilingual Sequence-To-Sequence Models for Extremely Low-Resource NMT," in 2021 Moratuwa Engineering Research Conference (MERCon). IEEE, 2021, pp. 432–437.

[35] A. Fernando, G. Dias, and S. Ranathunga, "Data augmentation and list integration for improving domain-specific sinhala-english-tamil statistical machine translation," 2021.

[36] A. Fernando and S. Ranathunga, "Data augmentation to address out-of-vocabulary problem in low-resource sinhala-english neural machine translation," arXiv preprint arXiv:2205.08722, 2022.

[37] R. Perera, T. Fonseka, R. Naranpanawa, and U. Thayasivam, "Improving english to sinhala neural machine translation using part-of-speech tag," arXiv preprint arXiv:2202.08882, 2022.

[38] N. de Silva, "Sinhala text classification: observations from the perspective of a resource poor language," ResearchGate, 2015.

[39] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. De Silva, and G. Dias, "Implementing a corpus for sinhala language," in Symposium on Language Technology for South Asia 2015, 2015.

[40] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. de Silva, and G. Dias, "Comparison between performance of various database systems for implementing a language corpus," in International Conference: Beyond Databases, Architectures and Structures. Springer, 2015, pp. 82–91.

[41] T. Mikolov, W.-T. Yih, and G. Zweig, 'Linguistic regularities in continuous space word representations', in Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.

Thank You

Questions

?