LAVA Corpus for WMT22 Large-Scale Machine Translation Data Track

Jun Cao[1], Yifeng Liu[2], Xingyuan Pan[3], Jiaqiang Wang[1], Xian Qian[1], Mingxuan Wang[1]
[1] ByteDance AI Lab, Shanghai, China
[2] Institute for Interdisciplinary Information Sciences, Tsinghua University
[3] Department of Computer Science and Engineering, Wuhan University
{caojun.sh, wangjiaqiang.sonian, qian.xian, wangmingxuan.89}@bytedance.com
liuyifen20@mails.tsinghua.edu.cn
panxingyuan209@gmail.com

***Abstract*— LAVA Corpus contains millions of parallel bilingual sentences, which are mined from Common Crawl. It covers 26 African languages required for WMT22 - Large Scale Multilingual Translation Task. We have proposed a method for mining parallel sentences between individual web pages or across multiple parallel web pages using the Compact Language Detector, LASER and Vecalign. For the latter case, we have also proposed a way to improve the efficiency in finding parallel web pages by inspecting urls. By using our method, we have mined over 3 million qualified sentence pairs.**

I. INTRODUCTION

Machine translation research has traditionally placed an outsized focus on a limited number of languages , mainly belonging to the Indoeuropean family. Progress in many languages, some with millions of speakers, has been hampered by the lack of data. An encouraging trend recently has been the increasing attention to low-resource languages. However, these modelling efforts have been hindered by the lack of high-quality, standardized evaluation benchmarks.

The WMT22 Data Track focuses on the contribution of novel corpora. For this task, we mined 3,225,801 bilingual sentence pairs from African languages to English/French. The corpus is named **LAVA**, because our mining method is **L**anguage **A**gnostic using **V**ector **A**lignment, and can be extended to language directions beyond African ones.

II. METHOD

Existing bi-text mining techniques can be classified as global or local according to the text range of the comparison. CCMatrix[1] is a global type aimed at mining bi-texts by comparing collections of large-scale monolingual data from web corpora, while most local mining techniques, like CCAligned[2], use a hierarchical approach: select a subset of text that may contain bi-texts from the document level, and identify parallel bi-texts in the selected text.

Global mining is considered to have the characteristics of high recall, low precision, and usually less efficient due to expensive computational costs. Meanwhile, local mining may be fast because only thousands of sentences are compared, and it is more accurate because local bilingual sentences are more likely to be parallel with a close context.

For LAVA corpus, we focused on two local scenarios, parallel sentences appearing on single web page and across multiple web pages. Specifically, we assumed that parallel sentences could appear on a single web page like a dictionary web page, which uses multiple bilingual sentences to exemplify the usage of a word. Also, it is likely that there are multiple web pages describing the same content but written in different languages, which are termed as parallel web pages, for example, Wikipedia web pages in different languages of a same concept. The Common Crawl[3] corpus contains petabytes of data collected since 2008. It contains raw web page data, extracted metadata and text extractions. The two kinds of web pages in the Common Crawl corpus are the sources for our mining work. We download more than 80 archived snapshots in WET format from Common Crawl website[1] and only reserve the web documents that comply the requirements of our scenarios. After this step, we utilized Vecalign[4], an accurate but efficient sentence alignment algorithm to mine parallel bilingual sentences. We used LASER[5] encoders released by WMT to obtain multilingual sentence embeddings and facilitate the alignment work. Since our method is local type, it can be easily deployed to a distributed cluster with low-cost compute nodes. We open source the code of our method here[2].

A. Mining in Single Web Pages

The first step is to filter out web pages that may contain parallel sentences (see Fig. 1(a)). We used Compact Language Detector 2 (cld2) to detect the top three languages found in each web page and their approximate percentages of the total text bytes. Eligible web pages should contain at least two languages, and the top two languages should account for more than a percentage threshold of 30%. We eventually filtered out 2,973,326,390 eligible web pages. The second step is to split the web page into two sentence groups assigned with the top two languages. These two groups are termed as parallel documents. The third step is to extract parallel sentences from the above parallel documents. We used Vecalign[6] tool and LASER model mentioned above to facilitate the automatic sentence alignment.

B. Mining across Multiple Parallel Web Pages

The first step is to retrieve parallel web page groups in Common Crawl corpus. Eligible web page groups should

---

[1] https://commoncrawl.org/the-data/get-started/

[2] https://github.com/volctrans/vector-align

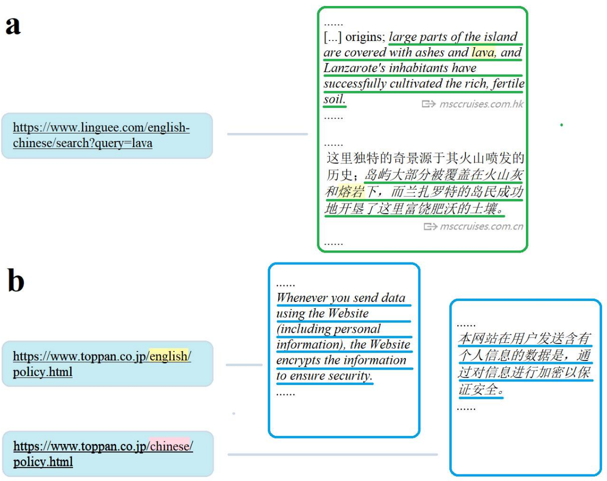
a
https://www.linguee.com/english-chinese/search?query=lava
......
[...] origins; large parts of the island are covered with ashes and lava, and Lanzarote's inhabitants have successfully cultivated the rich, fertile soil.
msccruises.com.hk
......
......
这里独特的奇景源于其火山喷发的历史；岛屿大部分被覆盖在火山灰和熔岩下，而兰扎罗特的岛民成功地开垦了这里富饶肥沃的土壤。
msccruises.com.cn
......
b
https://www.toppan.co.jp/english/policy.html
https://www.toppan.co.jp/chinese/policy.html
......
Whenever you send data using the Website (including personal information), the Website encrypts the information to ensure security.
......
......
本网站在用户发送含有个人信息的数据是，通过对信息进行加密以保证安全。
......

Fig. 1. The illustration of our method with each case an English-Chinese pair.
a) For single web page sources, we focused on those web pages with at least two languages with the highest 2 proportion both larger than the threshold of 30% (Example from lava - Chinese translation – Linguee, where there are lots of bilingual sentence pairs collected from other web pages).
b) For multiple parallel web pages sources, we focused on those web page pairs with urls only different on language tags.(Examples from `https://www.toppan.co.jp/english/policy.html` and `https://www.toppan.co.jp/chinese/policy.html`)

contain at least two web pages in distinct languages, and the web pages in the group should be monolingual or the language with the highest proportion accounts for more than a threshold of 80%, which can also be filtered by cld2.

Also, the most important thing is that the web pages in the same group should be parallel, which means they are of the same content but written in different languages. Considering the huge amount of data, the search efficiency, other than the recall of searching, is of great significance, and therefore, we can use some tricks, such as focusing on the url pairs with the same structure. In the real case, if the multiple web pages have parallel urls, we recognize them as parallel web pages (see Fig. 1(b)). Parallel urls refer to the urls that are same after removing language names and symbols from their suffixes, network locations or paths. We grouped web pages by their removed-language-symbols urls using regular expression. In the aggregate function, we discarded web pages with duplicate language or content.

The second step is to traverse all web page combinations in each group using LASER[5] to compute sentence embedding, and then use Vecalign[6] to mine the parallel sentences in each combination, just like mining in single web pages. We also recorded the url pairs of web pages with high alignment scores for further mining.

III. RESULTS

The statistics of LAVA Corpus for WMT22 are shown in Table I.

And one sample record from LAVA Corpus is shown as following:

```
{
  "src_text": "\"Dit is hy wat ons gemaak
              het, en nie ons self nie.\"
              - Psalm 100:3, NW.",
  "trg_text": "\"It is he that has made us,
              and not we ourselves.\"
              - Psalm 100:3.",
  "src_lang": "afr",
  "trg_lang": "eng",
  "score": 0.63,    // the higher score,
                    // the better quality
  "scope": "cross"  // single or cross
}
```

source	target	count
afr	eng	1,834,139
afr	fra	566,570
eng	kin	147,585
eng	lug	20,993
eng	nya	163,040
eng	swh	371,864
fra	kin	20,241
fra	lug	4,855
fra	nya	93,885
fra	swh	1,633
kin	lug	608
kin	swh	203
lug	swh	35
nya	swh	137
afr	swh	7
kin	nya	4
lug	nya	1
afr	kin	1
	total	3,225,801

TABLE I
STATICS OF LAVA CORPUS

IV. FURTHER ANALYSIS

A. *Further Improvement in Recall of Searching*

We only searched for web page pairs with the same structure of urls, which missed parallel web pages with more difference in urls. For example, the url of the entry of the concept ”city” in Wikipedia is `https://en.wikipedia.org/wiki/City` in English but `https://fr.wikipedia.org/wiki/Ville` in French, being different not only in language code but also in spelling of the concept. So it results in small recall of searching. For further improvement in recall, we can use more of the information on web pages in trade of efficiency. For example, [7] provided a method using unsupervised learning for embedding of web pages and indexing them by locality-sensitive hashing([8], [9]) for further exploration of parallel web page pairs.

B. *Further Improvement in Efficiency across Multiple Parallel Web Pages*

In most cases, the corresponding sentences on parallel web pages are arranged in the same order, so it could be improved to take only linear time for matching by using more aggressive algorithms other than Vecalign. However, there is a much greater probability in searching for parallel sentences, which may result in an even smaller recall.

C. Further Improvement in Language Identification and LASER

Compact Language Detector 2 (cld2) detects over 80 languages, which cover the 21 of 24 African languages for WMT22. We found the LID model released in NLLB[10] could predict 218 languages. Meanwhile, LASER3 is released with support for over 200 languages.

D. Record of Low Quality Web Pages

During our work, we found that some web pages consist of low-quality data, and it is recommended to find out a method for identifying these web pages and record them on the blacklist for improvement in the efficiency of filtering web pages.

V. SUMMARY

To conclude, our method has shown that it is practical to mine parallel corpus in a single web page and across multiple parallel web pages. And Compact Language Detector, LASER and Vecalign are effective tools which we have used to mine between parallel web pages. LAVA corpus could be downloaded from WMT22 official page[3].

REFERENCES

[1] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*, 2019.

[2] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. Ccaligned: A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*, 2019.

[3] Common crawl. `https://commoncrawl.org/`. Accessed: 2022-07-18.

[4] Brian Thompson and Philipp Koehn. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Holger Schwenk and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation. April 2017.

[6] Brian Thompson and Philipp Koehn. Exploiting sentence order in document alignment. pages 5997–6007, November 2020.

[7] Jakub Kúdela, Irena Holubová, and Ondřej Bojar. Extracting parallel paragraphs from common crawl. 2018.

[8] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, page 380–388, New York, NY, USA, 2002. Association for Computing Machinery.

[9] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, jan 2008.

[10] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 2022.

---

[3] https://www.statmt.org/wmt22/large-scale-multilingual-translation-task.html