Searching the Web for Cross-lingual Parallel Data

Ahmed El-Kishky
ahelk@fb.com
Facebook AI
Seattle, Washington

Philipp Koehn
phi@jhu.edu
Johns Hopkins University
Baltimore, Maryland

Holger Schwenk
schwenk@fb.com
Facebook AI
Paris, France

ABSTRACT

While the World Wide Web provides a large amount of text in many languages, cross-lingual parallel data is more difficult to obtain. Despite its scarcity, this parallel cross-lingual data plays a crucial role in a variety of tasks in natural language processing with applications in machine translation, cross-lingual information retrieval, and document classification, as well as learning cross-lingual representations. Here, we describe the end-to-end process of searching the web for parallel cross-lingual texts. We motivate obtaining parallel text as a retrieval problem whereby the goal is to retrieve cross-lingual parallel text from a large, multilingual web-crawled corpus. We introduce techniques for searching for cross-lingual parallel data based on language, content, and other metadata. We motivate and introduce multilingual sentence embeddings as a core tool and demonstrate techniques and models that leverage them for identifying parallel documents and sentences as well as techniques for retrieving and filtering this data. We describe several large-scale datasets curated using these techniques and show how training on sentences extracted from parallel or comparable documents mined from the Web can improve machine translation models and facilitate cross-lingual NLP.

KEYWORDS

cross-lingual document retrieval, cross-lingual sentence retrieval, machine translation, multilingual embedding, web mining

ACM Reference Format:
Ahmed El-Kishky, Philipp Koehn, and Holger Schwenk. 2020. Searching the Web for Cross-lingual Parallel Data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401417

1 MOTIVATION

Recognizing the limited availability of cross-lingual parallel data and the abundance of noisy, heterogeneous, multilingual web-data, we motivate the necessity of massively multilingual and scalable methods for automatically sifting through web-scale data and identifying the relatively scarce parallel data.

Work on mining such data goes back to the 20th century [29] and has been traditionally broken up into the stages (1) identification of multi-lingual web sites, (2) crawling, (3) text extraction, (4) document alignment, (5) sentence alignment, and (6) sentence pair filtering.

Despite its scarcity, parallel cross-lingual data plays a crucial role in a variety of tasks in natural language processing. Traditionally, machine translation approaches have leveraged parallel sentences as training data for use with sequence-to-sequence models. Because training data can be scarce, previous works have shown that training on sentences extracted from parallel or comparable documents mined from the Web can improve machine translation models [24]. On the document level, parallel cross-lingual documents can also be used for learning word-level translation lexicons [13, 25]. Other tasks that leverage parallel data include cross-lingual information retrieval as well as cross-lingual document classification. Additionally, cross-lingual data facilitates training multilingual representations such as XLM [23] which can be used as input to many downstream NLP tasks yielding language-agnostic NLP. Parallel text has been also extensively used for the projection of linguistic annotation [37] and training of paraphrasers [14].

With rapid globalization, performing NLP tasks such as machine translation for a variety of languages has become crucial. Yet the availability of parallel data for many low-resource languages makes it difficult. These cross-lingual text pairs are difficult to obtain not only due to their relative scarcity, but also the scarcity of human annotators familiar with the languages. As such, the development of language-agnostic techniques to automatically sift through large quantities of web-data in search of training data is crucial.

2 OBJECTIVES

The objective of this tutorial is to introduce the audience to the task of automatically searching for high-quality parallel text within large heterogeneous web corpora. The task is inherently an information retrieval task, and effectively mining such web corpora can have a large impact on the NLP community which faces a bottleneck in obtaining training data. We hope this tutorial spurs the information retrieval community to develop scalable and accurate techniques to tackle the interesting challenges that these tasks present.

2.1 What will be covered?

Preliminaries: We introduce the audience to the broad subject of searching the web for cross-lingual data by providing motivation in the context of obtaining training data for cross-lingual NLP. Within this context, we introduce machine translation, multi-lingual universal representations, and cross-lingual information retrieval examples validating the necessity for large-scale web-mining for parallel data to facilitate these tasks. Finally, we describe the end-to-end approach for sifting through large-quantities of crawled web-data to extract these “needle in a haystack" parallel cross-lingual texts.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07.
https://doi.org/10.1145/3397271.3401417

Multilingual Web Corpora: In this part of the presentation, we first begin by discussing source of corpora with multilingual data suitable for mining parallel text. We begin with homogeneous corpora that are the most straightforward to mine. We transition to large multilingual web corpora that have been mined for parallel data such as Wikipedia and large-scale webcrawls such as CommonCrawl snapshots.

Multilingual Representations: In this part of our presentation, we introduce the notion of multilingual dense representations. We formulate the problem and discuss three approaches to creating these multilingual representations.

1. **Multilingual Word Embeddings:** We discuss the most basic of approaches to developing multilingual word representations. These techniques include unsupervised, weakly supervised, and supervised techniques for embedding words from different languages in the same semantic space.
2. **Multilingual Sentence Representations:** We discuss methods at directly embedding sentences into a joint multilingual space and explain why these methods are the most suitable for large-scale retrieval of cross-lingual data.
3. **Transformer-based multilingual representations:** We discuss recent work on adapting multilingual masked language models for cross-lingual understanding. We explain how self-supervised training techniques can be used to achieve state-of-the-art performance in cross-lingual understanding. We also discuss challenges in using these techniques for large-scale web tasks.

Cross-lingual Document Retrieval: In this part of the tutorial, we discuss techniques for identifying pairs of documents in different languages that are translations, near translations, or of comparable content.

1. **Metadata-Based Document Alignment:** We discuss techniques that leverage metadata for document alignment. We start with high-precision techniques such as aligning documents based on URL patterns, and explore methods that leverage document structure, temporal features.
2. **Translation-based document alignment:** We discuss methods that rely on applying machine translation models to translate documents in the target language into the source language and applying standard information retrieval techniques to align documents.
3. **Document representation Similarity Alignment:** We discuss techniques that build up multilingual document representations. Given these representations, document similarity is used to guide matching between source and target documents.
4. **Alternative Document Mining Techniques:** We discuss a variety of techniques that have been used for document alignment including techniques based on earth mover's distance and supervised document alignment techniques.

Cross-lingual Sentence Retrieval: In this part of the tutorial we discuss historical and recent methods for performing bilingual sentence alignment, the task of taking parallel documents, which have been split into sentences and for each sentence, retrieving its translation [34]. In addition, we discuss challenges to sentence alignment in low-resource languages.

Large-scale Mining of Cross-lingual Sentences: In this section, we discuss recent techniques that attempt to identify bilingual sentence pairs directly from large web-corpora, without first identifying parallel document pairs. Such techniques can be described as a large-scale cross-lingual sentence retrieval task and recent advances in dense retrieval have made this possible. We discuss efficient and scalable techniques for nearest neighbor search in large dense representations, scoring functions for identifying high-quality pairs from candidate retrievals, and efforts to directly retrieve pairs from Wikipedia and CommonCrawl web snapshots [31, 32].

Sentence Pair Filtering: In this section, we discuss the necessity of performing more-costly filtering of any mined parallel data to a stricter degree. We explain how filtering parallel data, especially filtering noisy data obtained from mined web-data is crucial, especially for low-resource language pairs, drawing from findings of two shared tasks that we organized on this subject [20, 21]. We will discuss how alignment errors have been noted to have a small effect on statistical machine translation performance, however misaligned sentences have been shown to be much more detrimental to neural machine translation [17].

3 WHY A TUTORIAL AT SIGIR 2020

With the globalization of information and proliferation of social media and online communication, people are exposed to an explosion of information in potentially hundreds of different languages. Effective semantic understanding of this data at scale and facilitating cross-lingual communication requires language-agnostic NLP as well as high-quality machine translation. While efforts in cross-lingual NLP and machine translation can potentially solve these issues, their major bottleneck is the scarcity of cross-lingual parallel data. Cross-lingual information retrieval offers an opportunity for automatically obtaining this crucial training data from large unstructured multilingual corpora. This tutorial will introduce open source tools and present an organized picture of recent research on cross-lingual text retrieval with motivation for automatically obtaining training data for NLP. We demonstrate not only several key retrieval tasks and state-of-the-art solutions, but also introduce several downstream NLP tasks that can be used to evaluate retrieval performance.

3.1 Audience and Prerequisites

Researchers and practitioners in the field of information retrieval, data mining, natural language processing, web search, and information systems. While an audience with a good background in these areas would benefit most from this tutorial, we believe the material to be presented would give a general audience and newcomers an introductory pointer to the current work and important research topics in this subfield, and inspire them to learn more. Only preliminary knowledge about machine learning, information retrieval, and natural language processing and their applications are needed.

4 TUTORIAL OUTLINE

This tutorial presents a comprehensive overview of the techniques developed for automatically mining parallel data from the web developed in recent years. We will discuss the following key issues and papers.

1. **Preliminaries of Mining the Web for Parallel Data**
   - (a) Scarcity of parallel text but abundance of large multilingual heterogeneous corpora
   - (b) Parallel sentences for training machine translation
   - (c) Aligned multilingual documents for training language agnostic document classification.
2. **Web Crawling and Multilingual Corpora**
   - (a) ParaCrawl: Web-scale parallel corpora for the languages of the EU .
   - (b) Wikipedia web data
   - (c) CommonCrawl: open repository of web crawl data .
3. **Cross-lingual Representations**
   - (a) MUSE: Multilingual Unsupervised and Supervised Embeddings.
   - (b) LASER: Language Agnostic Sentence Representations
   - (c) XLM & XLM-R: Transformer-based Cross-lingual Representations
4. **Parallel Document Retrieval**
   - (a) Metadata-based Approaches and URL-Aligned CommonCrawl
   - (b) Translation with td/idf approaches
   - (c) XLSMD and SAE: Dense-representation Similarity Approaches .
5. **Document-Level Parallel Sentence Retrieval**
   - (a) HunAlign: Dictionary-based Sentence Alignment
   - (b) Gargantua: unsupervised sentencealignment for symmetrical and asymmetrical parallel corpora
   - (c) BleuAlign: MT-based sentence alignment
   - (d) VecAlign: Order-aware Sentence Alignment in Linear-time and Space
6. **Global Parallel Sentence Retrieval**
   - (a) FAISS: Billion-scale Simiarity Search with GPUs .
   - (b) WikiMatrix: LASER with Margin retrieval on Wikipedia .
   - (c) CCMatrix: LASER with Margin retrieval on CommonCrawl
7. **Parallel Sentence Filtering**
   - (a) Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora .
   - (b) Bicleaner: Feature-based filtering .
   - (c) Low-Resource Corpus Filtering using Multilingual Sentence Embeddings .
   - (d) QE-Clean: Quality Estimation for Sentence Filtering .

5 SUPPORT MATERIALS

To accompany the lecture-style tutorial, the instructors will provide resources in the form of example Jupyter notebooks, freely available datasets, and open-source libraries and code. These include:

- **Open-source Code & Libraries**
  - Large-scale Dense Representation Similarity Search (GPU and CPU-based)
  - Unsupervised & Supervised Multilingual Word Embeddings
  - Unsupervised & Supervised Multilingual Sentence Embeddings
  - Massively Multilingual Sentence-segmentation and Tokenization
  - CommonCrawl Preprocessing Code
- **Open-source Code for Generating Cross-lingual Datasets**
  - Code for Generating High-quality Monolingual Data from CommonCrawl
  - Code for Generating Parallel documents from CommonCrawl corpus
  - Code for Generating Parallel sentences from Wikipedia corpus
  - Code for Generating Parallel sentences from CommonCrawl corpus
  - Code for Generating Parallel documents from CommonCrawl corpus
- **Jupyter Notebooks Demonstrating**
  - Language-specific Sentence segmentation
  - Cross-lingual Sentence Representation
  - Large-scale Text Similarity and Nearest-neighbor Retrieval with Dense Representations

6 ABOUT THE INSTRUCTORS

The tutorial will be presented by Ahmed El-Kishky, Philipp Koehn, and Holger Schwenk:

- **Ahmed El-Kishky** is a Research Scientist at Facebook AI where he works on developing automated methods for obtaining machine translation training data. Before that, El-Kishky received his PhD from the University of Illinois at Urbana-Champaign where he was supported by the National Science Foundation Graduate Research Fellowship (NSF-GRF) and the National Defense Science and Engineering Graduate (NDSEG) Fellowship. In his career, El-Kishky has published papers and given tutorials in venues such as KDD, VLDB, SIGMOD, WWW, WSDM, ICDM, and BIGDATA.
- **Philipp Koehn** is a professor in the Department of Computer Science, is recognized worldwide for his leading research in and applications for developing and understanding data-driven methods to solve long-standing, real-world challenges of machine translation and machine learning. Koehn authored the textbook “Statistical Machine Translaton" and the forthcoming “Neural Machine Translation". Koehn serves on the editorial boards for multiple journals, among them: Transaction of the Association of Computational Linguistics; Machine Translation Journal; Artificial Intelligence Review; Computation, Corpora, Cognition, and ACM Transactions on Asian and Low-Resource Language Information Processing. Koehn is president of the ACL Special Interest Group on Machine Translation which organizes a series of ACL workshops and conferences on Machine Translation since 2005 (WMT).

- **Holger Schwenk** is a research scientist at Facebook Artificial Intelligence Research Paris. He received his PhD in Computer Science from the University Paris 6 in 1996, and prior to joining Facebook in 2015, he was professor for computer science at the University of Le Mans where he led a large group on statistical machine translation. During Schwenk's career, he has authored papers in top machine learning and natural language processing venues such as ACL, NAACL, EMNLP, and NeurIPS.

Previous Tutorials The authors have presented many tutorials at machine learning, natural language processing, data mining and database conferences. The following is a set of recent tutorials given by the authors.

- **Tutorial at ACL, 2016:**"Computer Aided Translation: Advances and Challenges".
- **Tutorial at SIGMOD 2016:** "Automatic Entity Recognition and Typing in Massive Text Data" [26].
- **Tutorial at WWW 2016:** "Automatic Entity Recognition and Typing in Massive Text Corpora" [28].
- **Tutorial at KDD 2015:** "Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach" [27].
- **Tutorial at MT Summit XV 2015:** "Computer Aided Translation: Advances and Challenges" [19].
- **Tutorial at KDD 2014:** "Bringing Structure to Text: Mining Phrases, Entities, Topics, and Hierarchies" [15]
- **Tutorial at EACL 2006:** "Statistical Machine Translation: the basic, the novel, and the speculative", [22].
- **Tutorial at HLT/NAACL 2004:** "What's New in Statistical Machine Translation", [18].

REFERENCES

[1] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL* 7 (2019), 597–610.

[2] Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *COLING*. ACL, 81–89.

[3] Christian Buck and Philipp Koehn. 2016. Quick and reliable document alignment via tf/idf-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. 672–678.

[4] Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-Resource Corpus Filtering using Multilingual Sentence Embeddings. *arXiv preprint arXiv:1906.08885* (2019).

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

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

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

[8] Common Crawl. 2018. Common crawl. *URl: http://http://commoncrawl. org* (2018).

[9] Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The cmu-avenue french-english translation system. In *WMT*. 261–266.

[10] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2019. A Massive Collection of Cross-Lingual Web-Document Pairs. *arXiv preprint arXiv:1911.06154* (2019).

[11] Ahmed El-Kishky and Francisco Guzmán. 2020. Massively Multilingual Document Alignment with Cross-lingual Sentence-Mover's Distance. *arXiv preprint arXiv:2002.00761* (2020).

[12] Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. 118–119.

[13] Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *COLING*.

[14] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *NAACL*.

[15] Jiawei Han, Chi Wang, and Ahmed El-Kishky. 2014. Bringing structure to text: mining phrases, entities, topics, and hierarchies. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1968–1968.

[16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).

[17] Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *WNGT* (Melbourne, Australia). ACL, 74–83. http://aclweb.org/anthology/W18-2709

[18] Kevin Knight and Philipp Koehn. 2003. What's New in Statistical Machine Translation.. In *HLT-NAACL*. 5–5.

[19] Philipp Koehn. 2015. Computer Aided Translation Advances and Challenges. (2015).

[20] Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *WMT*. ACL, Florence, Italy, 54–72. https://doi.org/10.18653/v1/W19-5404

[21] Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *WMT*. ACL, Belgium, Brussels, 726–739. https://www.aclweb.org/anthology/W18-6453

[22] Philipp Koehn and EM Training. 2006. Statistical machine translation: the basic, the novel, and the speculative. *Tutorial at EACL* (2006).

[23] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* (2019).

[24] Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31, 4 (2005), 477–504.

[25] Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *ACL*. ACL, 519–526.

[26] Xiang Ren, Ahmed El-Kishky, Heng Ji, and Jiawei Han. 2016. Automatic entity recognition and typing in massive text data. In *Proceedings of the 2016 International Conference on Management of Data*. 2235–2239.

[27] Xiang Ren, Ahmed El-Kishky, Chi Wang, and Jiawei Han. 2015. Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2319–2320.

[28] Xiang Ren, Ahmed El-Kishky, Chi Wang, and Jiawei Han. 2016. Automatic entity recognition and typing in massive text corpora. In *Proceedings of the 25th International Conference Companion on World Wide Web*. 1025–1028.

[29] Philip Resnik. 1999. Mining the Web for Bilingual Text. In *ACL*. http://acl.ldc.upenn.edu/P/P99/P99-1068.pdf

[30] Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 955–962. https://doi.org/10.18653/v1/W18-6488

[31] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791* (2019).

[32] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. *arXiv preprint arXiv:1911.04944* (2019).

[33] Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. (2010).

[34] Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. In *EMNLP-IJCNLP*. 1342–1348.

[35] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4* 292 (2007), 247.

[36] Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *EMNLP*. 2945–2950.

[37] David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT* (San Francisco, CA). 161–168.