Text Information Retrieval in Tetun

Gabriel de Jesus(✉)

INESC TEC and Faculty Engineering of the University of Porto (FEUP),
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
gabriel.jesus@inesctec.pt

Abstract. Tetun is one of Timor-Leste's official languages alongside Portuguese. It is a low-resource language with over 932,000 speakers that started developing when Timor-Leste restored its independence in 2002. Newspapers mainly use Tetun and more than ten national online news websites actively broadcast news in Tetun every day. However, since information retrieval-based solutions for Tetun do not exist, finding Tetun information on the internet and digital platforms is challenging. This work aims to investigate and develop solutions that can enable the application of information retrieval techniques to develop search solutions for Tetun using *Tetun INL* and focus on the ad-hoc text retrieval task. As a result, we expect to have effective search solutions for Tetun and contribute to the innovation in information retrieval for low-resource languages, including making Tetun datasets available for future researchers.

Keywords: Information retrieval · Tetun · Search · Ad-hoc retrieval · Low-resource language

1 Introduction and Motivation

Information retrieval (IR) deals with finding documents of unstructured nature that satisfy an information need from within large collections [1]. In the ad-hoc retrieval task, users typically formulate information needs as a query using natural language text and execute it through search. Then, the retrieval system retrieves documents that are relevant to the given query from large textual document collections and returns them to the user in an ordered list. One important IR application is keyword-based web search, as showcased by the Google search engine. Since specific information retrieval-based approaches for Tetun do not exist, finding documents that satisfy an information need written in Tetun is challenging.

Tetun is the language spoken in Timor-Leste. Timor-Leste is a multilingual country with two official languages, Tetun and Portuguese, two working languages, Indonesian and English [5], and more than 30 dialects [6]. Tetun is a dialect that was previously used as a trade language before Timor-Leste restored

This PhD research is financed by the Portuguese Foundation for Science and Technology (FCT) under the scholarship grant number SFRH/BD/151437/2021.

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 429–435, 2023.
https://doi.org/10.1007/978-3-031-28241-6_48

its independence and became a new sovereign state on 20 May 2002. In 2002, the Government of Timor-Leste designated Tetun as one of its official languages and since then, it has become a dominant language in public life.

There are two major varieties of Tetun: Tetun Dili (referred to as Tetun) and Tetun Terik [2]. Tetun Dili comprises Tetun INL and Tetun DIT. Tetun Terik is one of Timor-Leste's dialects. Tetun INL is Tetun for which the government of Timor-Leste produced standard orthography through the *Instituto Nacional de Linguísticas* (INL) [3]. Tetun INL has become the official Tetun dialect being used in the education system, official publications, and media [4]. Tetun DIT is produced by the linguists at Dili Institute of Technology (DIT) with a few standard difference with the Tetun INL in writing structures [2], such as Tetun INL uses "ñ" and "ll", while Tetun DIT uses "nh" and "lh", e.g., millaun (Tetun INL), milhaun (Tetun DIT).

By the 2015 census, the population of Timor-Leste was 1.17 million and the proportion of Tetun speakers was 79.04%, alongside Portuguese (2.56%), Indonesian (2.02%), and English (1.04%) [6]. Most of the Tetun verbs, nouns, and adjectives are Portuguese loanwords and Tetun shows much more influence from Portuguese noted in media [7–9]. Consequently, Tetun is still widely varied in writing mainly influenced by Portuguese and Indonesian.

To tackle the aforementioned problem, this work aims to develop search solutions for Tetun using *Tetun INL* and focus on the ad-hoc text retrieval task, including developing a text corpus and a test collection for evaluation. Since online news has played a crucial role in promoting Tetun in the last five years, from 2017 to 2022, we will use Timor News [10] as our case study. Timor News will be used to showcase Tetun search solutions to further evaluate the algorithms' performance and retrieval effectiveness. Timor News broadcasts national and international news only in Tetun INL and has over 4,000 news articles registered in the database. The portal is accessed by visitors from around the globe, with a total of 1,500 visitors on average per day, and as it was founded by the author of this paper [11], we have full access to the platform. The remainder of this paper is organized as follows. Section 2 presents the related work. We describe our main research questions in Sect. 3. Then, Sects. 4 and 5 present the research methodology and the specific research issues. Finally, Sect. 6 closes with a final remark.

2 Related Work

Low-resource languages (LRLs) can be understood as the languages that are less studied, less computerized, low density [12,13], that lack text corpora, or reduced accessibility [14]. Developing information retrieval tools for a LRL requires corpora, specific techniques and algorithms, and test collections. For the text corpora construction, Artetxe et al. [15] built a Basque corpus comprising 12.5 M documents and 423 M tokens employing tailored crawling and reported that it could be an effective alternative to obtain high-quality data for LRLs. Linder et al. [16] developed SwissCrawl, a Swiss German corpus composed of over half one million sentences, by crawling the web using a tool they developed.

Dovbnia and Wróblewska [17] developed a language identification (LID) model experimented with three Celtic language variations comprising 9,969 sentences corpus and reported that using n-gram characters as input enables building a robust LID model for LRLs in supervised and semi-supervised settings. Ferilli [18] described that term and document frequency (TDF) could be used to identify stopwords from a small number of corpora and stated that it outperformed the classic term frequency (TF) [19–21] and the normalized inverse document frequency (NIDF) from Lo et al. [22]. Tukeyev et al. [23] developed a Turkic stemmer using stopwords, affixes, and root lists and reported a ratio of 97% in the experiment tested with 1,000 Turkic words.

Chavula and Suleman [24] built a test collection for Chichewa, Citumbuka, and Cinyanja, where the topics formulation and relevance assessment of the document-query pairs were conducted by five external assessors. Esmaili et al. [25] built a test collection for Sorani Kurdish and Aleahmad et al. [26] created the Hamshahri test collection for Persian. The relevance of the query-document pairs for the first test collection was judged by the native Sorani speakers, and Persian students collaborated in the latter.

There are several search engines for low-resource languages, as showcased in the works for Bantu languages. Holy et al. [27] developed a search engine for nine Bantu languages that enable users to search using multilingual queries. Malumba et al. [28] developed a search engine for isiZulu and a search engine for IsiXhosa was developed by Kyeyune [29]. All these search engines were developed using a similar infrastructure - crawling the documents from the web using focused crawling [30], identifying documents using a language identification model, and indexing, retrieval, and ranking using the Solr platform.

3 Research Questions

This research will be focused on developing solutions for ad-hoc text retrieval for Tetun and our main research questions (RQs) are the following.

RQ1—*What retrieval strategies provide the most effective solutions for Tetun text-based search?* Different retrieval strategies for the ad-hoc retrieval task have been studied, from the classical probabilistic-based [31] to neural-based approaches [32,33]. Considering the language diversity and the fact that Tetun is a low-resource language, this question extends our knowledge to investigate the works in ad-hoc information retrieval on low-resource languages which can be used to support in developing text retrieval solutions for Tetun search.

RQ2—*Do query processing operations improve the text retrieval effectiveness in Tetun INL?* Tetun INL words contain accented letters (á, é, í, ó, ú, ñ), apostrophes (‘), and hyphens (for mono-semantic). Moreover, a multilingual environment influences Tetun in writing. This research question intends to investigate how applying query and document preprocessing operations, such as normalizing the accented letters, removing stopwords, and stemming, can improve text retrieval effectiveness in Tetun text search.

RQ3—*What impact does the Tetun variant used in Tetun INL has on the text retrieval effectiveness in Tetun text search?* Since Tetun is still widely varied in writing, this research question aims to investigate whether properly written Tetun INL documents improve text retrieval effectiveness.

4 Research Methodology

The proposed steps for the research are described as follows. We will start by crawling the World Wide Web to collect documents and store them in a repository. Documents are preprocessed to remove unnecessary elements, apply the LID model to extract Tetun texts from the collections, and then build a text corpus for Tetun. Three individuals with proficiency in Tetun will be hired to evaluate the corpus quality using a similar approach to Artetxe et al. [15].

Documents and topics will be characterized according to TREC guidelines, where some topics will be extracted from search query logs of Timor News, and the team members will formulate the others. A TREC-style approach, based on the Cranfield approach [34], will be employed to build a test collection for evaluation. The open-source platforms for IR research, such as Solr, Elasticsearch, and Terrier IR, will be used to index and rank documents for each query and then create a pool of documents to judge their relevance. Five Tetun native speakers will be hired to conduct the relevance judgments.

After a corpus and a test collection are built, we will conduct experiments using the topics and documents developed to respond to the research questions outlined in Sect. 3. The most effective retrieval strategy for Tetun will be showcased in a search prototype using Timor News as a case study to further evaluate the algorithms' performance.

5 Specific Research Issue

Lin [35] stated that neural-based models had shown substantial improvements over traditional methods, even in the low-resource languages, referring to the work of Yang et al. [36]. However, the result reported by Yang et al. [36] was still quite far (0.3152 AP) from the best-known result on Robust04 (0.3686 AP). Therefore, since Tetun is a newly developed and LRL, we intend to discuss the gap between traditional and neural-based approaches in ad-hoc retrieval in our research context to get insightful feedback. So the question would be: is it possible to do transfer learning or create a multilingual aligned model from high-resource languages like English or partially similar languages like Portuguese?

6 Final Remark

This paper presents the idea of developing search solutions for Tetun which includes main research questions, methodologies to be adopted, and the specific research issues to be discussed. As a continuation of the work, we will follow the steps outlined in Sect. 4 to respond the research questions in Sect. 3.

References

1. Manning, C.-D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press, Cambridg (2009)
2. van-Klinken, C.-W., Hajek, J., Nordlinger R.: Tetun Dili: a grammar of an East Timorese language, Pacific Linguistics, Canberra, Australia (2002)
3. The standard orthography of the tetum language. https://archive.org/details/the-standard-orthography-of-the-tetum-language. Accessed 31 Oct 2022
4. Government decree-law No. 1/2004 of 14 April 2004 - the standard orthography of the tetun language. https://mj.gov.tl/jornal/lawsTL/RDTL-Law/RDTL-Gov-Decrees/Gov-Decree-2004-01.pdf. Accessed 31 Oct 2022
5. Constitution of the democratic republic of timor-leste. https://timor-leste.gov.tl/wp-content/uploads/2010/03/Constitution_RDTL_ENG.pdf/. Accessed 31 Oct 2022
6. Timor-leste population and housing Census 2015. General directorate of statistics, ministry of finance, democratic republic of timor-leste. https://www.statistics.gov.tl/category/publications/census-publications Accessed 31 Oct 2022
7. Hajek, J., van-Klinken., C.-W.: language contact and gender in Tetun Dili: what happens when Austronesian meets romance?. University of Hawai'i Press **58**, 59–91 (2019). https://doi.org/10.1353/ol.2019.0003
8. Zuzana, G.: Tetun in Timor-Leste: The role of language contact in its development. PhD thesis, Universidade de Coimbra, Portugal (2018). https://hdl.handle.net/10316/80665
9. van-Klinken, C. W., Hajek, J.: Language contact and functional expansion in Tetun Dili: the evolution of a new press register. Multilingual **37**, 613–647 (2018)
10. Timor news: an online news agency based in Dili, Timor-Leste, https://www.timornews.tl
11. The registered and licensed social communication agencies in press council of timor-Leste. https://conselhoimprensa.tl/baze-de-dadus/registu-media. Accessed 31 Oct 2022
12. Magueresse, A., Carles, V., Heetderks, E.: Low-resource languages: a review of past work and future challenges. CoRR, abs/2006.07264 (2020). https://arxiv.org/abs/2006.07264
13. Cieri, C., Maxwell, M., Strassel, M.-S., Tracey, J.: Selection criteria for low resource language programs. In: Calzolari, N., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, 23–28 May 2016. European Language Resources Association (ELRA) (2016). https://www.lrec-conf.org/proceedings/lrec2016/summaries/1254.html'
14. Hoenen, A., Koc, C., Rahn, M.-D.: A manual for web corpus crawling of low resource languages. Umanistica Digitale **4**(8), 342–344 (2020). https://doi.org/10.6092/issn.2532-8816/9931
15. Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O., Soroa, A.: Does corpus quality really matter for low-resource languages?. CoRR abs/2203.08111 (2022). https://doi.org/10.48550/arXiv.2203.08111
16. Linder, L., Jungo, M., Hennebert, J., Musat, C.-C., Fischer, A.: Automatic creation of text corpora for low-resource languages from the internet: the case of swiss German. In Béchet, F., et al. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, 11–16 May 2020, pp. 2706–2711, European Language Resources Association (2020). https://aclanthology.org/2020.lrec-1.329/

17. Dovbnia, O., Wróblewska, A.: Automatic language identification for celtic texts. CoRR abs/2203.04831 (2022). https://doi.org/10.48550/arXiv.2203.04831
18. Ferilli, S.: Automatic multilingual stopwords identification from very small corpora. Electron. **10**(17) (2021). https://doi.org/10.3390/electronics10172169
19. Ferilli, S., Izzi, G.L., Franza, T.: Automatic stopwords identification from very small corpora. In: Stettinger, M., Leitner, G., Felfernig, A., Ras, Z.W. (eds.) ISMIS 2020. SCI, vol. 949, pp. 31–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67148-8_3
20. Baeza-Yates, R., Ribeiro-Neto, B.-A.: Modern Information Retrieval - the Concepts and Technology Behind Search, 2nd edn. Pearson Education Ltd., Harlow (2011)
21. Croft, W.-B., Metzler, D., Strohman. T.: Search Engines - Information Retrieval in Practice. Pearson Education, London (2009). https://www.search-engines-book.com
22. Lo, R.-T., He, B., Ounis, T.: Automatically building a stopword list for an information retrieval system. J. Digital Inf. Manage. **3**(1), 3–8 (2005)
23. Tukeyev, U., Karibayeva, A., Turganbayeva, A., Amirova, D.: Universal programs for stemming, segmentation, morphological analysis of Turkic words. In: Nguyen, N.T., Iliadis, L., Maglogiannis, I., Trawiński, B. (eds.) ICCCI 2021. LNCS (LNAI), vol. 12876, pp. 643–654. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88081-1_48
24. Chavula, C., Suleman, H.: Ranking by language similarity for resource scarce southern bantu languages. In: International Conference on the Theory of Information Retrieval (ICTIR), Virtual Event, Canada, 2021, pp. 137–147. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3471158.3472251
25. Esmaili, K.-S., et al.: Building a test collection for Sorani Kurdish. In: ACS International Conference on Computer Systems and Applications, AICCSA 2013, Ifrane, Morocco, 27–30 May 2013, pp. 1–7, IEEE Computer Society (2013). https://doi.org/10.1109/AICCSA.2013.6616470
26. Aleahmad., A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F.: Hamshahri: A standard Persian text collection. Knowl. Based Syst. **22**(5), 382–387 (2009). https://doi.org/10.1016/j.knosys.2009.05.002
27. von-Holy, A., Bresler, A., Shuman, O., Chavula, C., Suleman, H.: Bantuweb: a digital library for resource scarce South African languages. In: Masinde, M. Proceedings of the South African Institute of Computer Scientists and Information Technologists, SAICSIT 2017, Thaba Nchu, South Africa, 26–28 September 2017, pp. 36:1–36:10, Association for Computing Machinery (2017). https://doi.org/10.1145/3129416.3129446
28. Malumba, N., Moukangwe, K., Suleman, H.: AfriWeb: a web search engine for a marginalized language. In: Allen, R.B., Hunter, J., Zeng, M.L. (eds.) ICADL 2015. LNCS, vol. 9469, pp. 180–189. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27974-9_18
29. Kyeyune, M.-J.: IsiXhosa search engine development report. Department of Computer Science, University of Cape Town (2015). https://pubs.cs.uct.ac.za/id/eprint/1035/1/report.pdf. Accessed 10 Jul 2022
30. Chakrabarti, S., van-den-Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource. Comput. Netw. **31**(11–16), 1623–1640 (1999). https://doi.org/10.1016/S1389-1286(99)00052-3
31. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, April 2009. vol. 3, pp.

333–389, Now Publishers Inc., Hanover, MA, USA (2009). https://doi.org/10.1561/1500000019
32. Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document Ranking with a Pretrained Sequence-to-Sequence Model. In: Chon, T., He, Y, Liu, Y. Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020, ACL, vol. EMNLP 2020, pp. 708–718. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.63
33. Yang, W., Zhang, H., Lin, J.: Simple Applications of BERT for Ad hoc document retrieval. CoRR abs/1903.10972, (2019). https://arxiv.org/abs/1903.10972
34. Clough, P.-D., Sanderson, M.: Evaluating the performance of information retrieval systems using test collections. Inf. Res. **18**(2) (2013). https://www.informationr.net/ir/18-2/paper582.html'
35. Lin, J.: The neural hype, justified!: a recantation. ACM SIGIR Forum **53**(2), 88–93 (2019). https://doi.org/10.1145/3458553.3458563
36. Yang, W., Lu, K., Yang, P., Lin, J.: Critically examining the "Neural Hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In: Piwowarski, B., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, 21–25 July 2019, pp. 1129–1132, ACM (2019). https://doi.org/10.1145/3331184.3331340