Textual Representations for Crosslingual Information Retrieval

Bryan Zhang
Amazon.com
bryzhang@amazon.com

Liling Tan
Amazon.com
lilingt@amazon.com

Abstract

In this paper, we explored different levels of textual representations for cross-lingual information retrieval. Beyond the traditional token level representation, we adopted the subword and character level representations for information retrieval that had shown to improve neural machine translation by reducing the out-of-vocabulary issues in machine translation. Additionally, we improved the search performance by combining and re-ranking the result sets from the different text representations for German, French and Japanese.

1 Introduction

Cross-lingual information retrieval (CLIR) systems commonly use machine translation (MT) systems to translate the user query to the language of the search index before retrieving the search results (Fujii and Ishikawa, 2000; Pecina et al., 2014; Saleh and Pecina, 2020; Bi et al., 2020).

Traditionally, information retrieval and machine translation systems convert search queries to tokens and n-grams level textual representation (Jiang and Zhai, 2007; McNamee and Mayfield, 2004; Leveling and Jones, 2010; Yarmohammadi et al., 2019). Modern neural machine translation (NMT) systems have shown that subwords and character representations with flexible vocabularies outperform fixed vocabulary token-level translations (Sennrich et al., 2016; Lee et al., 2017; Kudo and Richardson, 2018; Wang et al., 2019). This study explores the shared granularity of textual representations between machine translation and cross-lingual information retrieval.

Textual representations of varying granularity encode queries differently, resulting in more diverse and robust search retrieval. Potentially, subwords and character-level representations are less sensitive to irregularities in noisy user-generated queries, e.g. misspellings and dialectal variants.

Tokens:	*americium ist ein chemisches element ...*
Subwords:	*_am er ic ium _ist _ein _chemische s _ element ...*
Characters:	*a m e r c i u m _ i s t _ e i n _ c h e m i s c h e s _ e l e m e n t*

Table 1: Example of a Pre-processed Document with Different Text Representations

2 Related Work

Neural machine translation had shown to outperform older paradigm of statistical machine translation models significantly and even "*achieved human parity in specific machine translation tasks*" (Hassan et al., 2018; Läubli et al., 2018; Toral, 2020). Moving from fixed token-level vocabulary to a subword representation unlocks open vocabulary capabilities to minimize out-of-vocabulary (OOV) issues[1].

Byte-Pair Encoding (BPE) is a popular subword algorithm that splits tokens into smaller units (Sennrich et al., 2016). This is based on the intuition that smaller units of character sequences can be translated easily across languages.

For instance, these smaller units appear when handling compound words via compositional translations, such as

For instance, subword units can better handle compound words via compositional German to English translations, ***schoko****lade* → *chocolate* and ***schoko****-creme* → *chocolate cream*. Suwbords can also cope with translations where we can easily copy or translate part of the source tokens or translate cognates and loanwords via phonological or morphological transformations, e.g. *posi****tiv*** →

[1] Although subwords allow more flexibility than tokens in creating unseen words, most NMT systems cannot support a genuinely open vocabulary thus a backoff token <unk> is often used during inference to represent subwords that is not seen in the training data.

Proceedings of the 4th Workshop on e-Commerce and NLP (ECNLP 4), pages 116–122

positive and *negativ* (German) → *negative*.

While BPE reduces the OOV instances, it requires the input to be pre-tokenized before applying the subword compression. Alternatively, Kudo and Richardson (2018) proposed a more language-agnostic approach to subword tokenization directly from raw string inputs using unigram language models.

Completing the whole gamut of granular text representations, Lee et al. (2017) explored character-level neural machine translations that do not require any form of pre-processing or subword or token-level tokenization. They found that multilingual many-to-one character-level NMT models are more efficient and can be as competitive as or sometimes better than subwords NMT models. Moreover, character-level NMT can naturally handle intra-sentence code-switching. In the context of CLIR, they will be able to handle mixed language queries. Following this, Wang et al. (2019) found that using byte-level BPE vocabulary is 1/8 the size of a full subword BPE model. A multilingual NMT (many-to-one) setting achieves the best translation quality, outperforming subwords models and character-level models.

While finer granularity of text representations was exploited for machine translation, to our best knowledge, information retrieval studies have yet to study the impact of using these subword representations on traditional information retrieval systems (Robertson, 2004; Robertson and Zaragoza, 2009; Aly et al., 2014). However, many previous works have leapfrogged to using fully neural information retrieval systems representing text with underlying various subword representations and neural dense text representation.

Often, these neural representations are available in multilingual settings in which the same neural language model can encode texts in multiple languages. Jiang et al. (2020) explored using the popular multilingual Bidirectional Encoder Representations from Transformers (BERT) model to learn the relevance between English queries and foreign language documents in a CLIR setup. They showed that the model outperforms competitive non-neural traditional IR systems on a few of the sub-tasks.

Alternatively, previous researches have also used a cascading approach to machine translation and traditional IR where (i) the documents are translated to the foreign languages with neural machine translation and/or (ii) the foreign queries are translated before retrieval from the source document index (Saleh and Pecina, 2020; Oard, 1998; McCarley, 1999).

Saleh and Pecina (2020) compared the effects of statistical machine translation (SMT) and NMT in a cascaded traditional CLIR setting. They found that the better quality translations from NMT outperforms SMT and translating queries to the source document language that achieved better IR results than using foreign language queries on an index of translated documents.

Although fully neural IR systems are changing the paradigm of information retrieval, traditional IR (e.g. TF-IDF or BM25) approaches remain very competitive and can still outperform neural IR systems for some tasks (Boytsov, 2020; Jiang et al., 2020). In this regard, we follow up on the cascading approach to machine translation and information retrieval on traditional IR systems. This study fills the knowledge gap of understanding the effects of subword representation in traditional IR indices.

3 Experiments

We report the experiments on different textual representations on traditional IR in a cross-lingual setting using a large-scale dataset derived from Wikipedia Sasaki et al. (2018).

Sasaki et al. (2018) focused their work on a supervised re-ranking task using relevance annotations. We use those annotations from the same Wikipedia dataset to perform the typical retrieval task. The dataset was designed so that the English queries are expected to retrieve the Wikipedia documents in the foreign languages, and the foreign documents with the highest relevance are annotated with three levels of relevance. Formally, the ground truth data is a set of tuples: (English query, q, foreign document, d and relevance judgement r, where $r \in \{0, 1, 2\}$).[2]

Lang	#Docs	#Tokens	#Subwords	#Chars
DE	2.08	344	580	2,086
FR	1.89	289	405	1,508
JA	1.07	510	475	734

Table 2: Corpus statistics on Wikipedia documents in dataset from Sasaki et al. (2018). (All numbers are in units of one million)

We note that the Wikipedia documents in the dataset are not parallel (i.e. not translations of

[2]Note that a single English query can be mapped to multiple documents with varying relevance judgements

each other) but they are comparable in nature depending on the varying amounts of contributions available on the official Wikipedia dumps across different languages. For our study, we use the German, French and Japanese document collections and report retrieval performance of English queries translated to these languages.[3]

The Wikipedia corpus came pre-tokenized, so we had to detokenize the documents[4](Tan, 2018) before putting them through the subword tokenizer. We used pre-trained SentencePiece subword tokenizers used by the OPUS machine translation models(Tiedemann and Thottingal, 2020)[5]. Additionally, we emulated the typical pre-processing steps for character-level machine translation and split all individual characters by space, replacing the whitespaces with an underscore character.

Table 2 shows the corpus statistics of the number of documents, tokens, subwords, and characters for the respective languages. Although Latin alphabetic languages benefit from the extra information produced by splitting the tokens into subwords, Japanese presents an opposite condition. Japanese became more compact when represented by the subwords in place of the tokens. The examples in Table 1 show an instance of a sentence pre-processed in different levels of granularity. The underscore in the subword sequence represents a symbolic space and is usually attached to the following subword unit, whereas the whitespace represents the unit boundary between the subwords.

The English queries were translated using the same OPUS machine translation models.[6] Although these machine translation models are open source and free to use under a permissive CC-BY license, it takes a significant amount of GPU computation and major changes to the HuggingFace API (Wolf et al., 2020) to efficiently translate the query samples parallelized inference. We will release the modified code for parallel GPU inference and translation outputs for the data used in this experiment for future convenience to improve the replicability of this paper.

[3]We use the raw dataset from http://www.cs.jhu.edu/~kevinduh/a/wikiclir2018/ for the document indices.

[4]https://github.com/alvations/sacremoses

[5]https://huggingface.co/Helsinki-NLP

[6]We use the opus-mt-en-de, opus-mt-en-fr, and opus-mt-en-jap models, their BLEU and ChrF scores (Papineni et al., 2002; Popović, 2015) can be found on https://huggingface.co/Helsinki-NLP (Tiedemann and Thottingal, 2020; Tiedemann, 2020)

3.1 Information Retrieval System

We use the Okapi BM25 implementation in PyLucene as the retrieval framework with hyperparameter setting ($k_1 = 1.2, b = 0.75$) (Manning et al., 2008). We consider the top 100 documents ($top_k = 100$) in the search ranking as search results for each query.

3.1.1 Building index for the documents

For each foreign language, we created an index for the documents with 5 *TextField* as follows:

- **id**: the unique index of the document
- **surface**: the raw text of the document
- **tokens**: the document after tokenization
- **subword**: the document in SentencePiece subwords
- **char**: the document in characters

3.1.2 Querying the document index

During retrieval, each translated query is first processed into its respective text representations (tokens, subwords or characters) and parsed using Lucene's built-in query parser and analyzer. Additionally, we tried to improve the search results by combining and re-ranking the result sets from the different text representations.

3.1.3 Search result expansion

Our intuition is that queries of more granular text representation can improve the robustness of the retrieval and potentially override the textual noise (e.g., misspellings are handled better for some languages). Hence, we attempt to expand the list of possible candidate documents by combining the search results from the token and the subword representations.

Given a query q and its token q_{token} and subword $q_{subword}$ representations, we obtained two sets of search results from their respective indices R_{tokens} and $R_{subword}$. We concatenated R_{tokens} and $R_{subword}$, and remove the repeated candidates that appear in both sets from $R_{subword}$ as illustrated in Figure 1.

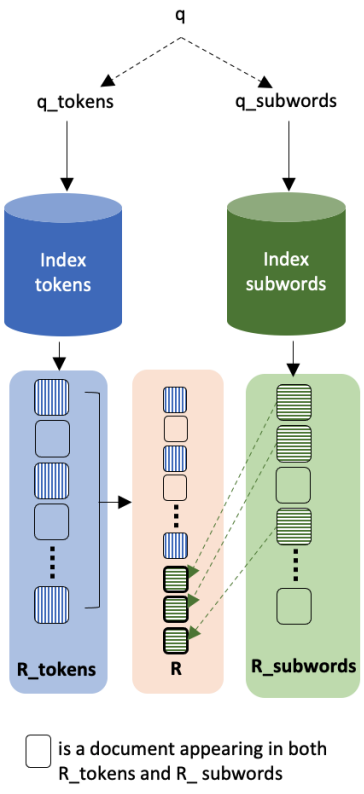

Figure 1: Search Results Expansion

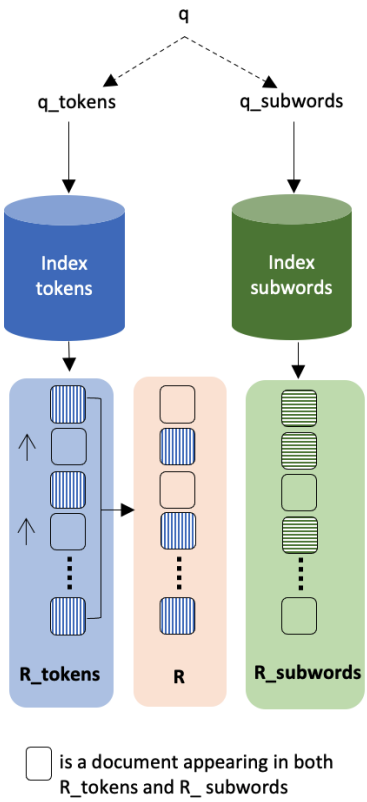

Figure 2: Search Results Re-ranking

3.1.4 Search result re-ranking

Aside from expanding the search results, we tried a re-ranking technique. We presumed that if different representations retrieve a document from a single query, it is more relevant than the documents that appear solely from one representation. Thus, we boosted the rank of the documents (D_{shared}) that are retrieved both in R_{tokens} and $R_{subword}$ from the same query. After boosting the rank of such documents (D_{shared}) by 1: $d \in D_{shared}$, $rank_{new}(d) = rank_{original}(d) - 2$, we re-rank the token-based search result, as illustrated in Figure 2 to get the final search result R.

3.2 Evaluation Metrics

We choose the following ranking metrics to evaluate the retrieval performance of the different text representations of query translation. Those ranking metrics are Mean Reciprocal Ranking **(MRR)**, Mean Average Precision **(MAP)**, normalized Discounted Cumulative Gain **(nDCG)**;

- MRR measures the ranking of the first document that is relevant to a given query in the search result.
- MAP evaluates the rankings of top 100 documents that are relevant to a given query in the search result.
- nDCG calibrates the ranking and relevance score of all the documents that are relevant to a given query in the search result. We compute nDCG@16 for the top-16 search results respectively.

4 Results

Table 3, 4 and 5 show the result for the CLIR experiments on the translated English queries and the German, French, and Japanese documents of different textual representations. For all the German and French setups, the token level representation achieved the best MAP, MMR, and NDCG scores, followed by subwords at significantly lower performance. Character-level representation performs the words at a magnitude $10\hat{4}$ times worse than token-level results.

We expected a margin between the token and subword level performance but the stark difference was surprising. Although machine translation can exploit the sequential nature of the open vocabulary with the subwords representation, traditional information retrieval methods disregard the other textual representation to a lesser extent. However, for

Metric	Token	Subword	Characters	Expansion	Re-ranking
MAP	**0.31299**	0.10072	0.00031	0.30432	0.30688
MRR	0.39938	0.12783	0.00033	0.39956	**0.40368**
nDCG	**0.40410**	0.13461	0.00021	0.13461	0.00021

Table 3: Results of CLIR Experiments on Translated English Queries on ***German*** Wikipedia

Metric	Token	Subword	Characters	Expansion	Re-ranking
MAP	**0.30330**	0.06931	0.00035	0.29859	0.29898
MRR	0.37866	0.08492	0.00039	**0.37872**	0.37830
nDCG	**0.36810**	0.09153	0.00060	0.36397	0.36537

Table 4: Results of CLIR Experiments on Translated English Queries on ***French*** Wikipedia

Metric	Token	Subword	Characters	Expansion	Re-ranking
MAP	**0.00039**	0.00036	0.00024	0.00036	0.00024
MRR	**0.00038**	0.00037	0.00025	0.00037	0.00025
nDCG	**0.00076**	0.00054	0.00022	0.00074	0.00075

Table 5: Results of CLIR Experiments on Translated English Queries on ***Japanese*** Wikipedia

Japanese, we see that the subword representation performs very similarly to the tokens counterparts.

For German and French documents, the intuition behind the poor performance of the character-level representation can be attributed to the meaningless and arbitrary nature of the unordered bag of characters. Whereas in Japanese, with its mix of syllabic and logographic orthography, the individual characters can potentially encode crucial semantic information.

We can see that both search result expansion and re-ranking techniques can improve the final search results for some languages. Table 3, 4 and 5 show that the search result expansion technique improves MRR for all three languages compared with the token-based retrieval baseline, and it improves both MRR and MAP for Japanese. The re-ranking technique achieves the highest MRR for both German and Japanese. Improvement in the MRR indicates that those two techniques can improve the ranking of the first relevant document appearing in the search results, which can be beneficial for cross-lingual e-commerce search systems. Neither the expansion nor the re-ranking technique achieves a better nDCG score, which is consistent with our expectation of improving the accuracy and robustness of retrieval with minimal changes to the relevance score that affects nDCG.

5 Conclusion

We explored the different granularity of textual representations in a traditional IR system within the CLIR task by re-using the subword representation from the neural machine translation systems. Our experiments in this paper provide empirical evidence for the underwhelming impact of subwords in traditional IR systems for Latin-based languages as opposed to the advancements that subword representation has made in machine translation.[7] In some scenarios, it is possible to achieve better CLIR performance by combining and expanding retrieval results of token and subword representations.

We conducted the experiments in this study using well-formed queries and documents. Our intuition is that a combination of the different textual representations can improve the robustness of the indexing and retrieval systems in realistic situations with noisier data (e.g. queries spelling or translations errors). For future work, we want to explore similar experiments with noisy e-commerce search datasets.[8]

[7]The processed datasets, code to generate the translations and evaluations will be made available under an open source license upon paper acceptance.

[8]We note that many open-source CLIR experiments are constrained to Wikipedia document searches. Although the lesson learned from these experiments can impact industrial e-commerce applications, the lack of open source e-commerce IR datasets limited the results reported in this study.

References

Robin Aly, Thomas Demeester, and Stephen Robertson. 2014. Probabilistic models in IR and their relationships. *Inf. Retr.*, 17(2):177–201.

Tianchi Bi, Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Constraint translation candidates: A bridge between neural query translation and cross-lingual information retrieval.

Leonid Boytsov. 2020. Traditional ir rivals neural models on the ms marco document ranking leaderboard.

Atsushi Fujii and Tetsuya Ishikawa. 2000. Applying machine translation to two-stage cross-language information retrieval. In *Envisioning Machine Translation in the Information Future*, pages 13–24, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation.

Jing Jiang and ChengXiang Zhai. 2007. An empirical study of tokenization strategies for biomedical information retrieval. *Inf. Retr.*, 10(4-5):341–363.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France. European Language Resources Association.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Johannes Leveling and Gareth J. F. Jones. 2010. Subword indexing and blind relevance feedback for english, bengali, hindi, and marathi ir. *ACM Transactions on Asian Language Information Processing*, 9(3).

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

J Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214.

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97.

Douglas W Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Conference of the Association for Machine Translation in the Americas*, pages 472–483. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, 61(3):165 – 185. Text Mining and Information Analysis of Health Documents.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Stephen Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:2004.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Shadi Saleh and Pavel Pecina. 2020. Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6849–6860, Online. Association for Computational Linguistics.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Liling Tan. 2018. Sacremoses: Python implementations of moses statistical machine translation pre-processing tools. https://github.com/alvations/sacremoses.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at wmt 2019.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords. *arXiv preprint arXiv:1909.03341*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. 2019. Robust document representations for cross-lingual information retrieval in low-resource settings. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 12–20, Dublin, Ireland. European Association for Machine Translation.