Exploring Snippets as a Dataset to Overcome Challenges in CLIR

Amit Asthana[1*] and *Sanjay K. Dwivedi*[1]

[1]Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India

Abstract. Cross-lingual information retrieval (CLIR) is a challenging task that requires overcoming linguistic barriers to match user queries with relevant documents in different languages. One of the major challenges in CLIR is the lack of parallel corpora, which hinders the development of effective translation models. This challenge can be addressed using snippets as a dataset to train CLIR models. Snippets can be automatically extracted from various sources, such as search engine result pages and can provide a rich and diverse set of collections for cross-lingual information retrieval. This paper initially discusses the challenges in CLIR and then explores the use of snippets as a dataset which can lead towards the development or improvements in the techniques to improve the retrieval effectiveness and further discusses the advantages and limitations of using snippets dataset in CLIR.

1 Introduction

Cross-lingual information retrieval (CLIR) is a challenging task that has gained increasing attention in recent years due to the growth of multilingual web content and the need for accessing information in different languages. CLIR requires matching user queries expressed in one language with relevant documents in another language, which involves several challenges, including lexical and semantic differences, cultural nuances, and varying levels of language proficiency among users. To address these challenges, researchers have proposed various approaches, including machine translation, bilingual dictionaries, and cross-lingual similarity measures, but the performance of CLIR systems still falls short of human-level accuracy.

One of the major challenges in CLIR is the lack of parallel corpora, which are essential for training effective translation models. Parallel corpora consist of aligned text segments in different languages, which enable the extraction of bilingual lexicons and the estimation of translation probabilities. However, parallel corpora are expensive to create, limited in size and domain coverage, and often require manual annotation or alignment, which limits their applicability to low-resource languages and tasks.

To overcome this challenge, we propose using snippets, short text excerpts from multilingual documents, as a dataset to train CLIR models. Snippets can be automatically extracted from various sources, such as search engine result pages or online news articles, and can provide

* Corresponding author: aamitonline@gmail.com

a rich and diverse set of examples for cross-lingual retrieval. Snippets can also capture the variability of language use and the context of the query, which can be useful for disambiguating polysemous terms and resolving cross-lingual ambiguity.

2 Related work

CLIR has been a topic of research for several decades, and various approaches have been proposed to address the challenges of cross-lingual retrieval. Early approaches focused on using bilingual dictionaries and thesauri to match query terms with their translations in the document collection. Later approaches explored using statistical machine translation models to automatically translate queries and documents between languages. More recent approaches have focused on learning cross-lingual embeddings and similarity measures that can map query and document representations into a common space.

Parallel corpora have been a valuable resource for training CLIR models, and several efforts have been made to create large-scale parallel corpora for high-resource languages. However, parallel corpora are limited in their coverage of low-resource languages and domains, and their quality depends on the alignment and translation methods used. This has led researchers to explore alternative sources of training data, such as comparable corpora, parallel texts extracted from the web, and cross-lingual annotations.

Several previous studies have investigated the use of search engine snippets for information retrieval. However, most of these studies focused on monolingual information retrieval and did not consider the challenges of CLIR. Some of the major researches are shown in table 1.

TABLE 1. Major researches using snippets

Research	**Techniques used**	**Result**
Sun et al. (2006) [4]	Query expansion with snippets using local context method by applying language modelling techniques	Achieved 9.81% and 17.49% improvement with compared to local context analysis and relation-based system respectively
Karadzhov et al. [5].	Fact checking and rumour detection using deep neural network with LSTM text encoding	39.9% relative error reduction is achieved
Hearst et al. [6]	Clustered the snippets before summarising each cluster to provide a cluster label.	Achieved improvement in result

Zamir et al. [7].	Extracted key phrases from set of snippets as cluster label and created clusters based on the keys.	Achieved improvements
Ko et al. (2008) [8]	Expanded the initial query using higher relevance weight noun terms Relevance weight (TSV)	Improvement in accuracy is achieved as high as 14.8%

3 CLIR Challenges

Cross-lingual information retrieval (CLIR) faces several challenges that need to be addressed to improve the accuracy of retrieval results. One of the primary challenges is the difference in languages' architecture, making translation between languages a complex task that requires high accuracy to avoid adversely affecting retrieval efficiency. Another challenge is the small size of queries, leading to increased ambiguity in query translation and lowering retrieval result accuracy. The presence of out of vocabulary (OOV) terms in the query can negatively impact the retrieved results as these words may have existed in the input query but go unnoticed, leading to a search term miss. Multiple meanings of the same word, known as homonymy and polysemy, generally lead to translation uncertainty. Furthermore, the documents returned in response to a user's query are not always what the user expects, according to some researchers. This could be due to a lack of sufficient information present in the query for effective retrieval.

4 Snippets as a Dataset to Overcome CLIR Challenges

Snippets, as a dataset, can overcome the challenges of CLIR in several ways. They provide a source of context and additional information that can help disambiguate queries and improve retrieval accuracy. Snippets contain relevant keywords and phrases extracted from the original documents, which can help match the query with the appropriate documents. Secondly, snippets can help address the problem of out-of-vocabulary terms in queries. By including snippets containing rare or specialized terms, the search engine can better match queries with relevant documents. Traditional datasets for CLIR are often outdated and do not contain current information, which can negatively impact the accuracy of retrieval results. Snippets, on the other hand, are collected in real-time and reflect the most current information available on a topic. This enables CLIR systems to retrieve the most relevant and accurate results for user's query. Furthermore, snippets often include context and can provide additional information that can help disambiguate the query and improve the accuracy of retrieval results. This can help improve cross-lingual information retrieval and overcome the challenges of translation uncertainty.

Most of the challenges in CLIR are related with the inappropriate translation of user query. After translation other important issues remain are ambiguity, presence of out of

vocabulary (OOV) words and less informative query problem. Sometimes user's query is insufficient to retrieve the information the user seeks. Query expansion (QE), one of the popular techniques that can resolve most of the issues related to ambiguity by adding one or more term into the original query. Dataset built by using snippets can play essential role by providing the expansion term to perform the QE and provide more information about the user's intent, hence improved retrieval effectiveness is achieved.

5 Popular datasets for CLIR

One approach to query expansion involves leveraging large datasets to identify and suggest additional terms or phrases that may be relevant to the user's search query. Here are some major datasets that are commonly used for query expansion:

WordNet: WordNet is a lexical database that organizes words into sets of synonyms, called synsets, and provides semantic relationships between them. WordNet has been widely used in research on query expansion to identify synonyms or related terms that may be relevant to a user's search query.

Wikipedia: Wikipedia is a vast online encyclopaedia that covers a wide range of topics. Many search engines use Wikipedia as a source of additional information for queries and as a way to identify related terms or concepts that may be relevant to a user's search query.

Thesaurus: A thesaurus is a reference work that lists words and groups them together based on similarity in meaning. Thesauri are often used in query expansion to identify synonyms or related terms that may be relevant to a user's search query.

Corpus of documents: A corpus of documents is a collection of text documents that can be used to identify common patterns and relationships between words. Corpora are often used in query expansion to identify terms or phrases that are frequently used together in the context of a particular topic or domain.

Query logs: Query logs are records of search queries that users have submitted to a search engine. Query logs can be analysed to identify patterns in the way users search for information and to suggest additional terms or phrases that may be relevant to a user's search query.

Several publicly accessible datasets are available for use in a variety of scientific motives in the field of CLIR. The National Institute of Standards and Technology's (NIST) TREC (Text REtrieval Conference) is one of the most popular set of datasets which begins with English and progresses to include French, German, Italian, Dutch, Chinese, Arabic, and other languages.

The second set of data comes from CLEF (Cross-Language Experiment Forum), which focuses on European languages, with the first trials utilising queries in Dutch, English, French, German, Italian, Spanish, Swedish, and Finnish on documents in English, German, French, and Italian.

Another one is the NTCIR (NACSIS (National Centre for Science Information Systems) Test Collection for Information Retrieval) workshop series, which is hosted by Japan's National Institute for Informatics (NII) in which Asian languages such as Japanese, Chinese, Korean, Vietnamese, and Mongolian are highlighted.

The Forum for Information Retrieval Evaluation (FIRE) was founded in 2008 and is a South Asian alternative to TREC, CLEF, and NTCIR. Its goal is to promote research in Indian language Information Access by establishing a similar platform for Indian languages that provides data and a common forum for evaluating models and methodologies.

6 Snippets and its importance as a dataset

Snippets are a common and crucial component of contemporary search engines like Google. Snippets are designed to make it easier for users to evaluate the usefulness of linked documents without having to click through each one individually [8]. Snippets are short summaries of linked documents that display below links in search engine result page (SERP) and typically contain one to three lines of text. The sentences or pieces of sentences that make up a given snippet may be taken from various parts of the source document, including visible text and invisible meta-data, including the HTML meta description tag, the alt-text on HTML img tags, and various microformat and microdata structured meta-data languages [9, 10]. Search engines generally generate dynamic snippets to choose document text that is pertinent to the user's current query when user submits a search query [11]. SERPs contain a variety of snippets that can be categorized into five types [12] i.e. regular snippets, rich snippets, Google News, entity types, and featured snippets as shown in table 2. It has been explored by focusing on how people read and understand the news that people usually read headlines rather than reading detailed news stories [13]. Studies show that the same behaviour also holds with respect to snippet [14], as it engages a significant percentage of users' attention when they explore SERPs. We also found that even if users read the full articles, news headlines have ability to build the perceptions of a user about a news [15].

TABLE 2. Types of Snippets

Type of Snippet	Description
Regular Snippet	A basic snippet that includes the page title, URL, and a brief description of the webpage's content.
Rich Snippet	A snippet that provides additional information such as ratings, reviews, and pricing, displayed with a preview.
Google News	A snippet that displays news articles related to a user's search query with a thumbnail image and a link to the original source.
Entity Types	A snippet that displays information related to a specific entity such as a person, organization, or place.
Featured Snippet	A snippet that appears at the top of the search results page, displaying a brief summary of the most relevant information for a user's search query.

Effective query expansion requires the use of a dataset to select suitable words or phrases. The Internet's vast expansion and frequent events, such as deals, discussions, wars, and summits, necessitate an accurately updated dataset that includes almost all necessary information to address user query. To achieve this objective, it is critical to create a dataset that contains real-time information collected from search engines in the form of snippets. This dataset can be tailored to users' needs by utilizing snippets and provides accurate, comprehensive, and up-to-date information to enhance the effectiveness of query expansion.

7 Extraction of Snippets

Extracting relevant snippets from search engine results pages can be a powerful technique for building a dataset for various research applications. Snippets are the brief descriptions that appear under the title of a search result, and they provide a preview of the content on the web page. By extracting snippets related to specific queries or topics, researchers can build a dataset that provides a representative sample of relevant information from the web.

The choice of technique for extracting snippets depends on several factors. Web scraping is a common technique for automating the extraction of snippets from SERPs. This technique involves using software tools like BeautifulSoup or Scrapy to extract snippets from web pages. APIs provided by search engines like Google and Bing allow developers to programmatically access search results and extract snippets. This technique is useful for extracting snippets at scale, and it is often more efficient than web scraping.

Crowdsourcing is another technique for extracting snippets from search engine results pages. This technique involves outsourcing tasks to a large group of people, usually through online platforms like Amazon Mechanical Turk. Workers can be asked to visit SERPs and extract relevant snippets based on specific criteria. Crowdsourcing can be useful for building datasets with high levels of precision and accuracy, but it can also be costly and time-consuming.

Manual extraction is a straightforward technique that involves manually visiting SERPs and extracting relevant snippets. This technique is best suited for small-scale datasets or for specific queries that are difficult to automate. However, it can be time-consuming and may not be feasible for large-scale datasets.

The choice of technique for extracting relevant snippets from search engine results pages depends on several factors, including the scale of the dataset, the complexity of the queries, and the available resources. Regardless of the technique used, it is essential to ensure that the dataset is representative, unbiased, and of high quality. Each method has its pros and cons for snippet extraction as shown in table 3.

TABLE 3. Pros and Cons of snippet extraction methods

Method	**Pros**	**Cons**
Web Scraping	Web scraping is a flexible method that can be used to extract search engine snippets.	Web scraping can be technically challenging, especially for beginners.
	It allows for customization of the scraper to fit specific requirements.	Search engines may have measures to prevent web scraping, making it difficult to access the data.
API Integration	APIs provide a simple and standardized way to access search engine data.	APIs may have usage limits or require payment, which can limit the amount of data that can be extracted.
	APIs often have clear usage terms and conditions, making it easier to avoid legal issues.	APIs may not provide access to all search engine data, leading to incomplete datasets.
Crowdsourcing	allows for large-scale data collection from multiple sources.	may introduce noise or bias into the data due to the variability in the crowd's expertise and quality of work.
	useful for validating automated methods or collecting data from multiple search engines.	Quality control can be challenging, and it may require additional resources to monitor and manage the crowd.
Manual data collection	can be done without technical expertise or specialized software.	can be time-consuming and expensive, especially for large datasets.
	useful for small datasets or validation of automated methods.	The collected data may be prone to errors or biases introduced by the human collector.

8 Advantages and Limitations of Snippets as a Dataset

Using search engine snippets as a dataset in cross-lingual information retrieval (CLIR) offers various benefits. Snippets provide a rich source of data that can be used to improve the performance of cross-lingual models. Snippets also provide a way to overcome the challenges of working with multilingual data, such as the scarcity of parallel corpora and the difficulty in obtaining translations. By leveraging the large volume of snippets available, it is possible to create a comprehensive and up-to-date dataset that can improve the accuracy of CLIR systems. Snippets provide a more natural language representation of the information needs of users as snippets are created based on user queries, which can help to improve the effectiveness of cross-lingual embeddings and other machine learning techniques used in CLIR.

Despite the benefits of using snippets as a dataset, there are some limitations and challenges to consider. Snippets may not always provide a complete or unbiased representation of the topic or event in question, as they are limited in length and scope. Additionally, there may be issues with the quality or relevance of some snippets, as they are extracted from a variety of sources.

9 Conclusion

Using snippets as a dataset can be a valuable approach to address the challenge of accessing accurate and up-to-date information for query expansion and other techniques for improving retrieval effectiveness. By leveraging the existing literature and discussing the benefits and potential applications of using snippets as a dataset, we found collection of snippets can be an alternative of conventional datasets and also hope to inspire further research and development in this area.

References

1. Sharma, Vijay Kumar, and Namita Mittal. "Cross lingual information retrieval (CLIR): Review of tools, challenges and translation approaches." Information systems design and intelligent applications (2016): 699-708.
2. Zhou, Dong, et al. “Query expansion for personalized cross-language information retrieval”, Semantic and Social Media Adaptation and Personalization (SMAP), 2015 10th International Workshop on. IEEE, 2015.
3. A. Seetha, S. Das and M. Kumar, "Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method," 10th International Conference on Information Technology (ICIT 2007), 2007, pp. 56-61, doi: 10.1109/ICIT.2007.53.
4. Sun, Renxu, Chai-Huat Ong, and Tat-Seng Chua. "Mining dependency relations for query expansion in passage retrieval." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006.
5. Karadzhov, Georgi, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. "Fully automated fact checking using external sources." arXiv preprint arXiv:1710.00341 (2017).
6. M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In Proceedings of the 19th ACM SIGIR, pages 76–84, 1996.

7. O. Zamir and O. Etzioni. Grouper: a dynamic clustering interface to Web search results. Computer Networks-the International Journal of Computer and Telecommunications Networking, 31(11):1361–1374, 1999.
8. Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E. Williams. 2007. Fast Generation of Result Snippets in Web Search. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07). ACM, New York, NY, USA, 127–134. https://doi.org/10.1145/1277741.1277766
9. Google. 2018. Customizing Results Snippets. Google Custom Search Developer Documentation. https://developers.google.com/custom-search/docs/snippets.
10. Google. 2018. Providing Structured Data. Google Custom Search Developer Documentation. https://developers.google.com/custom-search/docs/structured_data.
11. Lorena Leal Bando, Falk Scholer, and Andrew Turpin. 2010. Constructing Querybiased Summaries: A Comparison of Human and System Generated Snippets. In Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10). ACM, New York, NY, USA, 195–204. https://doi.org/10.1145/1840784.1840813
12. Strzelecki, Artur, and Paulina Rutecka. "Featured snippets results in Google web search: an exploratory study." Marketing and Smart Technologies: Proceedings of ICMarkTech 2019. Springer Singapore, 2020.
13. M. Glenski, C. Pennycuff, and T. Weninger. 2017. Consumers and Curators: Browsing and Voting Patterns on Reddit. IEEE Transactions on Computational Social Systems 4, 4 (Dec 2017), 196–206.
14. Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking Analysis of User Behavior in WWW Search. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04). ACM, New York, NY, USA, 478–479. https://doi.org/10.1145/1008992.1009079
15. Ullrich K. H. Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. Journal of Experimental Psychology: Applied 20, 4 (2014), 323–33.