# MuSeCLIR: A Multiple Senses and Cross-lingual Information Retrieval dataset

**Wing Yan Li** **Julie Weeds** **David Weir**
TAG laboratory
University of Sussex
Brighton, UK
{wl238, j.e.weeds, d.j.weir}@sussex.ac.uk

## Abstract

This paper addresses a deficiency in existing cross-lingual information retrieval (CLIR) datasets and provides a robust evaluation of CLIR systems' disambiguation ability. CLIR is commonly tackled by combining translation and traditional IR. Due to translation ambiguity, the problem of ambiguity is worse in CLIR than in monolingual IR. But existing auto-generated CLIR datasets are dominated by searches for named entity mentions, which does not provide a good measure for disambiguation performance, as named entity mentions can often be transliterated across languages and tend not to have multiple translations. Therefore, we introduce a new evaluation dataset (MuSeCLIR) to address this inadequacy. The dataset focusses on polysemous common nouns with multiple possible translations. MuSeCLIR is constructed from multilingual Wikipedia and supports searches on documents written in European (French, German, Italian) and Asian (Chinese, Japanese) languages. We provide baseline statistical and neural model results on MuSeCLIR which show that MuSeCLIR has a higher requirement on the ability of systems to disambiguate query terms.

| | English query | Translated query |
|---|---|---|
| (a) | Karen Carroll (judge) | KarenCarroll法官 |
| (b) | Brian Duffy (chef) | BrianDuffy主廚 |
| (c) | Larry Andersen | LarryAndersen |
| (d) | El Cacao, Veraguas | ElCacaoVeraguas |

Table 1: Example queries from BI-139 (Sun and Duh, 2020) translated to Chinese using MUSE (Conneau et al., 2018) dictionaries.

## 1 Introduction

Cross-Lingual Information Retrieval (CLIR) is a subfield of Information Retrieval (IR) where the task is to retrieve documents in language Y using queries in language X. Frequently, it is a combined process of translation and conventional IR. For efficiency, it is more common to translate queries from language X to language Y than to translate documents from language Y to language X.

Lexical ambiguity is a problem, and disambiguation found beneficial, in many natural language processing tasks, including machine translation (Raganato et al., 2019), information extraction (Delli Bovi et al., 2015) and information retrieval (Blloshmi et al., 2021). This problem is exacerbated in CLIR due to translation ambiguity (Zhou et al., 2007). For example, *letter* could mean alphabetic characters (字母) or a message (信). However, in this case, translation ambiguity compounds the problem. 信 can also refer to believing in something and this meaning has nothing to do with *letter*. Therefore, to increase the precision of the retrieval process, it is crucial to identify the correct translation of the query words in context.

Despite its importance, translation ambiguity problem has received relatively little attention by CLIR researchers, and has been overlooked by existing CLIR datasets. Many applications of CLIR, and thus many existing datasets, are dominated by searches for named entity mentions. For example, in the BI-139 English-Chinese dataset (Sun and Duh, 2020), about 73% of the queries contain at least one named entity. These do not tend to have multiple translations and, as illustrated by by the query examples from BI-139 shown in Table 1, can often be transliterated from one language to another. Looking at examples (a) and (b) of Table 1, these queries may be a mixture of named entity mentions and common nouns, but only common nouns are translated. Although named entity mentions can be ambiguous (e.g. different people share the same name), this is related to named entity ambiguity. Since we are viewing the problem from the translation perspective, we are focusing on lexical ambiguity.

In the real world, the need to search for unnamed entity mentions exists. For example, on a global topic such as conservation (Marshall et al., 2020), researchers might want to collect information about *bats* from around the globe. *Bat* is not a named entity and is named differently in different languages when referred to as an animal (Italian: Chiroptera; Chinese: 蝙蝠). Hence, retrieval systems need to distinguish between the translation of *bat* as an animal and *bat* as a piece of stick-like equipment in sports, based on the context information. However, due to a large amount of unambiguous named entity mentions, existing CLIR datasets are not adequate to train and evaluate the disambiguation ability of systems. We introduce MuSeCLIR, a new and a more fine-grained evaluation dataset.

MuSeCLIR is designed to specifically evaluate systems' ability to carry out disambiguation in CLIR. It is derived from Wikipedia, a free and open-sourced resource. Wikipedia contains many pages that exist in multiple languages and thus makes Wikipedia a good resource for CLIR (Sun and Duh, 2020; Sasaki et al., 2018; Yu et al., 2021). Assuming every possible translation represents a word sense, common nouns with more than one possible translation are chosen. By doing so, we can minimise the number of named entity mentions appearing in queries, and test whether systems are able to rank documents more highly which contain the correct translation (in context) over other possible translations of the ambiguous words in queries.

We introduce MuSeCLIR, a new evaluation dataset that assesses the ability of systems to disambiguate ambiguous query terms. MuSeCLIR supports searches on documents written in European (French, German, Italian) and Asian (Chinese, Japanese) languages. Our codes are available on GitHub[1]. Users can reproduce and extend MuSeCLIR to other languages. In section 3, we provide the construction method of MuSeCLIR and the statistics. In section 4, MuSeCLIR is used as a benchmark to evaluate existing CLIR systems: BM25 and multilingual BERT (mBERT) ranker (Sun and Duh, 2020). The results indicate that, given similar types of queries, existing CLIR systems perform more poorly on MuSeCLIR compared to other existing datasets, showing their inadequacy and the need of MuSeCLIR to determine the most appropriate system in the real-world scenario.

## 2 Background and Related Work

**Probabilistic approach** BM25 (Robertson and Zaragoza, 2009) is a traditional bag-of-words retrieval function based on term frequency-inverse document frequency (TF-IDF). It is a statistical measure that relies on term frequency and matches a query against a document. BM25 is monolingual, so we employ the MUSE (Conneau et al., 2018) bilingual dictionaries to translate queries into the target language during experiments.

Elasticsearch[2] is an open-source search engine that implements the BM25 algorithm (Robertson and Zaragoza, 2009) and has built-in analysers that handle tokenisation and stemming. Here, we employ Elasticsearch 6.5.4 with default parameters[3]; `smartcn` and `kuromoji` analyser are used when handling the Chinese and Japanese documents, respectively.

**Neural approach** Recently, end-to-end CLIR models have attracted more attention. These systems align queries and documents into the same space and perform matching in this aligned space. Large pre-traine d language models (PLM), such as BERT (Devlin et al., 2019), is commonly adopted as the encoder (Jung et al., 2022; Nair et al., 2022; MacAvaney et al., 2019) and impressive results have been achieved. As CLIR involves multiple languages, CLIR systems usually utilise multilingual language models, like mBERT, to map queries and documents into a shared space, bypassing the translation step.

The mBERT ranker used here is a re-implementation of the vanilla BERT ranker proposed by MacAvaney et al. (2019). The vanilla BERT ranker adopted the fine-tuning paradigm with a linear layer stacked on top of BERT (Devlin et al., 2019). Following Sun and Duh (2020), the encoder is replaced with a pre-trained mBERT mode[4]. The `[CLS]` embedding at the final layer of mBERT that represents the query-document pair is used. When training, the positive sample is a query-document pair with relevance labels larger than 0; negative otherwise. The network is trained to optimise pairwise hinge loss with Adam optimiser and updates the weights in the last linear layer.

**Datasets** MuSeCLIR is compared against two existing datasets, both also developed from

[1] https://github.com/justinaL/MuSeCLIR/

[2] https://www.elastic.co/

[3] $b = 0.75$ and $k_1 = 1.2$

[4] BERT-Base; Multilingual cased; Training epochs = 20; Learning rate = 1e-5

Wikipedia: BI-139 from CLIRMatrix (Sun and Duh, 2020) and WikiClir dataset introduced by Sasaki et al. (2018). CLIRMatrix contains two subsets: BI-139 (bilingual dataset) and MULTI-8 (multilingual dataset[5]). As this paper is not studying multilingual IR, BI-139 (base version) is considered. Queries in WikiClir are the first sentences from the English Wikipedia pages with page titles removed; the average query length is 20 tokens. As page titles are kept in MuSeCLIR, we append page titles to their original queries forming *WikiClir title* for fair comparisons. Queries in BI-139 are Wikipedia page titles of 3 tokens on average. In both sets, documents are the first 200 tokens of a page that contains the main gist of the topic.

# 3 Dataset construction

MuSeCLIR is an English-centric dataset where all queries are in English that makes use of WikipediaAPI[6]. Common nouns with multiple translations are chosen from MUSE (Conneau et al., 2018) bilingual dictionaries.

Wikipedia provides a `disambiguation` page for potentially ambiguous article titles. This page contains links to other possible subtopics in addition to the main topic (selected nouns in our case). For example, *window* can refer to architecture, rectangle display on computers, etc. This page will provide links to the corresponding articles. Here, we assume each link within the `disambiguation` page is associated with a sense of the noun. In order to conserve the overall level of ambiguity, we remove nouns with less than 2 links on the `disambiguation` page. Administration pages like `Help`, `disambiguation`, etc. are then removed. Before adding a linked page to the final collection, we also check that i) the page contains an inter-language link to the desired target language; ii) the noun exists in the page title; and iii) the summary of the page in both English and the target language is not empty. We initialise the construction based on common nouns to minimise the number of named entity mentions. This method is easily adapted to other parts of speech, but we focus on polysemous nouns here.

| Language | # ambiguous nouns | Total # sentence queries | # documents |
|---|---|---|---|
| FR | 2,045 (0.52) | 41,958 | 9,884 |
| DE | 2,389 (0.51) | 49,698 | 10,740 |
| IT | 1,565 (0.52) | 30,675 | 7,640 |
| ZH | 1,137 (0.50) | 26,344 | 5,080 |
| JA | 882 (0.52) | 20,675 | 4,168 |

Table 2: MuSeCLIR dataset statistics. Corresponding entropy of sense distribution are in brackets.

## 3.1 Design

MuSeCLIR queries are sentences. They are from the selected English Wikipedia page summaries that include the chosen noun; they do not necessarily appear in the target language pages[7]. To demonstrate the necessity of context, we also experiment with just the ambiguous noun as the query (*MuSeCLIR noun*). Documents are page summaries in the target language. Each query is paired with 1 relevant document, and there are two judgement labels, 1 for relevant and 0 for not relevant.

The design of train, validation and test sets are different. In train and validation sets, each query pairs with document candidates as determined by the word sense. The set of documents to rank for a given query contains all of the document for the correct translation, together with a random selection of irrelevant documents. It is harder to spot irrelevant documents in MuSeCLIR during test time. For each test set, the set of documents to rank is generated separately for each noun. Thus, queries of the same noun will be ranking the same set of documents. This set includes documents of both correct and incorrect translations, together with a random selection of irrelevant documents. This design is a characteristic of MuSeCLIR that preexisting datasets do not share.

## 3.2 Statistics

In the following experiments, the query language $X$ is English and the document language $Y$ is either an European language (French (FR), German (DE), Italian (IT)) or an Asian language (Chinese (ZH), Japanese (JA)). The scripts of Chinese documents are unified into traditional Chinese characters.

The mean entropy of sense distribution for all languages is around 0.5, meaning they are moderately ambiguous datasets (Jin et al., 2009). The

[5] Multilingual IR is a task where queries need to retrieve documents from a multilingual pool of documents

[6] https://github.com/martin-majlis/Wikipedia-API

[7] Target language pages are a mixture of pages written individually and translated from the English page.

entropy of a word is measured using the probability distribution over the senses of that word. The higher the entropy, the more ambiguous the dataset. The graphs of sense distribution and entropy distribution are given in Appendix A. For Asian languages, each sense has around 5 sentences on average and about 4 sentences on average for European languages. The average sentence length is 24 tokens; documents have 300 words on average.

Following CLIRMatrix (Sun and Duh, 2020), we aimed to select 10,000 queries for training and 1,000 queries for validation and testing. When selecting training data, we first randomly sampled $10,000/(4 \times 5) = 500$ nouns, where 4 is the average number of senses per noun and 5 is the average number of sentences per sense. We then selected as training data all of the senses and sentences for each noun, resulting in the exact numbers of queries shown in Table 2.

## 4 Evaluation

Two methods are considered in the following experiments: BM25, an unsupervised probabilistic approach, and an mBERT ranker, a supervised neural network (Sun and Duh, 2020). Typically, there are two ranking stages in IR systems. At the initial stage, each query will search over all documents and then rerank on a subset of documents returned from the first stage. However, the central challenge of MuSeCLIR is a cross-lingual and cross-sense problem, not the conventional IR task. Baseline models are two individual ranking models. BM25 ranks the complete document collection, and following (Sun and Duh, 2020), mBERT ranker ranks 100 documents at test time.

Results are reported using MAP@10 (mean average precision), calculated using `pytrec_eval`[8] (Van Gysel and de Rijke, 2018). As the metric considered will cut off at 10, we limit BM25 to return 10 documents. Queries are translated using MUSE (Conneau et al., 2018) bilingual dictionaries per token, and the first possible translation is returned, disregarding other possible translations. Out-of-vocabulary tokens will use the original form (i.e. English). Table 3 and 4 presents results across five datasets. BM25 (trans) refers to BM25 results using the translated queries. To investigate the effect of mixed language tokens in documents, we experiment on single language documents. They are created by matching tokens within the corresponding language Unicode range.

### 4.1 Results

**Elasticsearch - BM25** BM25 is a monolingual IR system, so performance after translation should be better. For both European (Table 3) and Asian (Table 4) languages, datasets with short queries like *MuSeCLIR noun* and *BI-139* are not improving after translation. Since short queries have fewer words for the matching process, they have more sparse representations. On the other hand, queries of *BI-139* are dominated by named entity mentions and without context. As MUSE dictionaries are not translating named entity mentions adequately, the performance of *BI-139* dropped after translation.

The number of foreign language tokens in queries and documents could be another factor affecting performances. The more foreign token found in documents, the better the models perform after translation, especially for datasets with longer search queries. *MuSeCLIR* might be an easier task for BM25 as the number of target language tokens in the document collections is higher and smaller in size than the existing datasets. For example, in the Chinese collection, as of MuSeCLIR, 75% of the tokens in the documents are Chinese characters; it is only 59% for BI-139. Hence, BM25 (trans) might have been rewarded more from token matching on *MuSeCLIR* than *WikiClir* and *BI-139*.

When ranking single language documents, overall performance decreased; thus, results reported in the previous setting have taken advantage of the English content. Similarly, MAP@10 increased after translation but is lower than in the mixed language setting. Potentially, this is caused by the decrease in document lengths, leading to shorter irrelevant documents ranked higher in the list more frequently.

**Multilingual BERT ranker** Unsurprisingly, mBERT ranker achieves better MAP@10. Since a multilingual language model is employed, languages are mapped into the same space, and no preceding translation is required. Across the board, *WikiClir* and *BI-139* have higher MAP@10 than the *MuSeCLIR*s. Possibly this is due to the list of documents to rank being more confusing in *MuSeCLIR* than the existing datasets.

Both CLIRMatrix and *WikiClir* label documents with more than one relevance level, but we found that it is seldom the case where a document is rel-

[8] A tool written in Python that builds on top of the standard TREC comunity evaluation tool https://github.com/usnistgov/trec_eval

|  | MuSeCLIR |  |  | MuSeCLIR noun |  |  | WikiClir |  |  | WikiClir title |  |  | BI-139* |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Models | FR | DE | IT | FR | DE | IT | FR | DE | IT | FR | DE | IT | FR | DE | IT |
| Results on original documents (mixed language) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| BM25 | 0.19 | 0.15 | 0.21 | 0.21 | 0.22 | 0.15 | 0.07 | 0.11 | 0.08 | 0.19 | 0.28 | 0.22 | 0.13 | 0.12 | 0.15 |
| BM25 (trans) | 0.34 | 0.18 | 0.37 | 0.17 | 0.11 | 0.18 | 0.17 | 0.14 | 0.14 | 0.32 | 0.30 | 0.30 | 0.12 | 0.11 | 0.14 |
| mBERT ranker | 0.79 | **0.80** | **0.81** | **0.43** | **0.45** | **0.44** | **0.87** | **0.89** | **0.85** | 0.90 | **0.95** | **0.91** | **0.59** | **0.61** | **0.67** |
| Results on clean documents (single language) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| BM25 | 0.22 | 0.09 | 0.10 | 0.17 | 0.14 | 0.08 | 0.07 | 0.08 | 0.04 | 0.19 | 0.20 | 0.10 | 0.11 | 0.09 | 0.09 |
| BM25 (trans) | 0.27 | 0.11 | 0.22 | 0.15 | 0.12 | 0.12 | 0.17 | 0.09 | 0.08 | 0.32 | 0.21 | 0.17 | 0.10 | 0.10 | 0.08 |
| mBERT ranker | **0.81** | 0.76 | 0.77 | 0.42 | 0.41 | 0.39 | 0.85 | **0.89** | 0.81 | **0.91** | 0.94 | 0.89 | 0.58 | 0.59 | 0.59 |

Table 3: **MAP@10 results on retrieving documents written in European languages using English queries.** (*)Authors reported 0.84, 0.88 and 0.84 nDCG@10 for FR, DE and IT respectively, we obtained 0.84, 0.85 and 0.82.

|  | MuSeCLIR |  | MuSeCLIR noun |  | WikiClir |  | WikiClir title |  | BI-139* |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Models | ZH | JA | ZH | JA | ZH | JA | ZH | JA | ZH | JA |
| Results on original documents (mixed language) |  |  |  |  |  |  |  |  |  |  |
| BM25 | 0.11 | 0.14 | 0.22 | 0.31 | 0.01 | 0.02 | 0.02 | 0.05 | 0.09 | 0.11 |
| BM25 (trans) | 0.42 | 0.22 | 0.21 | 0.19 | 0.02 | 0.05 | 0.05 | 0.09 | 0.03 | 0.05 |
| mBERT ranker | **0.77** | **0.81** | 0.42 | **0.46** | **0.88** | **0.81** | **0.94** | **0.84** | **0.81** | **0.79** |
| Results on clean documents (single language) |  |  |  |  |  |  |  |  |  |  |
| BM25 | 0.01 | 0.01 | 0 | 0 | 0.01 | 0 | 0.01 | 0 | 0 | 0 |
| BM25 (trans) | 0.40 | 0.20 | 0.20 | 0.14 | 0.03 | 0.03 | 0.05 | 0.06 | 0.01 | 0.02 |
| mBERT ranker | **0.77** | 0.76 | **0.44** | 0.43 | 0.86 | 0.77 | 0.92 | 0.81 | 0.66 | 0.71 |

Table 4: **MAP@10 results on retrieving documents written in Asian languages using English queries.** (*)Authors reported 0.84 nDCG@10 for both ZH and JA, we obtained 0.87 and 0.85 respectively.

evant to more than one query. This implies that these datasets focus more on evaluating systems' ability to position documents in the "right" order, which is a less challenging task. Moreover, existing datasets define linked documents of a page as less relevant documents. Less relevant documents do not necessarily relate to other senses of the queries and thus lower sense distribution entropy. Results demonstrated that mBERT ranker struggles more on MuSeCLIR than existing lower sense distribution entropy datasets. Observations are similar between mixed language and single language documents, and European and Asian language pairs.

Finally, we note that the contextual information in queries is crucial. There is a consistent drop in performance of approximately 50% from *MuSeCLIR* to *MuSeCLIR noun*. Without contextual information, it is impossible for systems to always choose the relevant document pertaining to the correct sense of the word. The mBERT ranker performs well on *MuSeCLIR*, indicating that it is doing well at disambiguating the nouns in the queries using the context. However, the MAP@10 of *MuSeCLIR* is not as high as an evaluation performed on *WikiClir* and *WikiClir title*, even when their queries are sentences. This would suggest there is scope to further improve CLIR systems.

## 5 Conclusion and Further Work

To address a deficiency in existing CLIR datasets, we introduce MuSeCLIR, a CLIR dataset that has been designed to challenge the ability of models to deal with ambiguous query terms. This dataset focused on polysemous common nouns with more than one possible translation, and which are regarded as ambiguous on Wikipedia. We argue that MuSeCLIR is a more suitable evaluation dataset for CLIR than pre-existing datasets.

Our method is replicable and extendable to other language pairs and other parts of speech. In the future, we also intend to test models that are trained with MuSeCLIR on real-world data and standard CLIR test sets.

## References

Rexhina Blloshmi, Tommaso Pasini, Niccolò Campolungo, Somnath Banerjee, Roberto Navigli, and Gabriella Pasi. 2021. IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1041, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Jin, Diana McCarthy, Rob Koeling, and John Carroll. 2009. Estimating and exploiting the entropy of sense distributions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 233–236, Boulder, Colorado. Association for Computational Linguistics.

Euna Jung, Jaekeol Choi, and Wonjong Rhee. 2022. Semi-siamese bi-encoder neural ranking model using lightweight fine-tuning. In *Proceedings of the ACM Web Conference 2022*, pages 502–511.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1101–1104, New York, NY, USA. Association for Computing Machinery.

Benjamin M. Marshall, Colin Strine, and Alice C. Hughes. 2020. Thousands of reptile species threatened by under-regulated global trade. *Nature Communications*, (1):4738.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396, Cham. Springer International Publishing.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *SIGIR*. ACM.

Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, WWW '21, page 1029–1039, New York, NY, USA. Association for Computing Machinery.

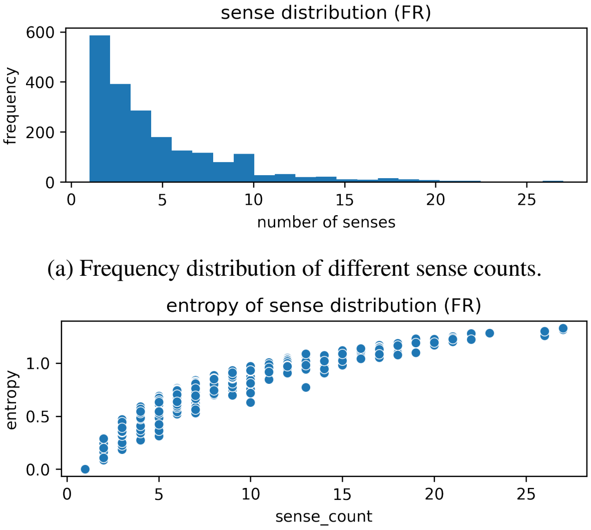
Dong Zhou, Mark Truran, and Tim J Brailsford. 2007. Ambiguity and unknown term translation in clir.

# A Sense Distribution Plots

Here, we provide plots of sense distribution and plots of the entropy of sense distribution across our language collection. The entropy of a word ($w$) is measured using the probability distribution ($p$) over the senses of the word, following equation 1.

$$H(\text{senses}(w)) = -\sum_{ws_i \in \text{senses}(w)} p(ws_i) \log\left(p(ws_i)\right) \quad (1)$$

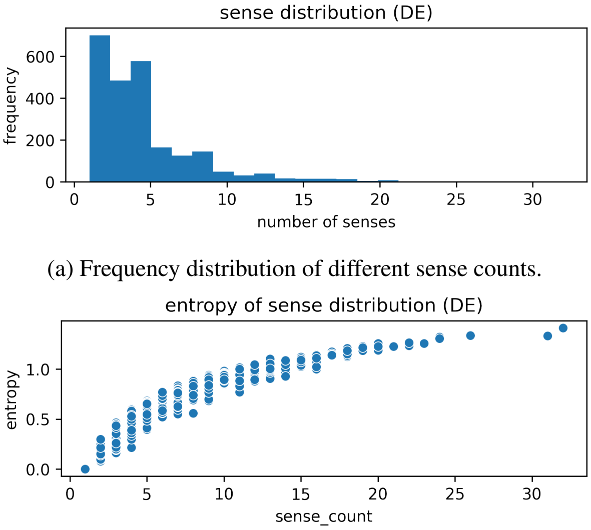
Each dot on the sense entropy distribution represents a noun. Essentially, the entropy distribution plot is the reverse version of the corresponding sense distribution plot.



(a) Frequency distribution of different sense counts.

(b) Entropy of sense distribution across different number of sense.

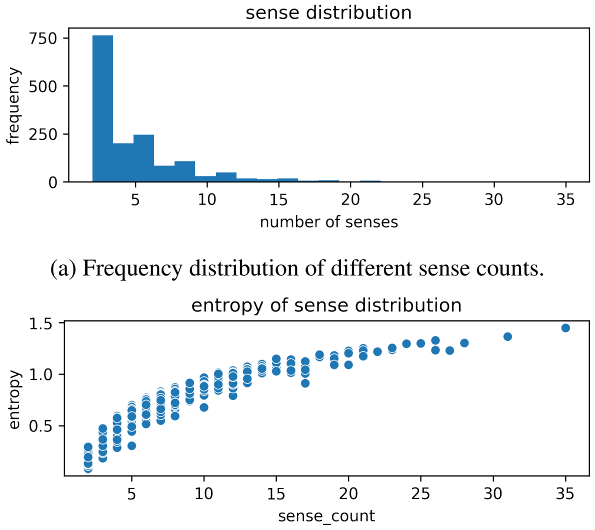
Figure 1: Plots of the French collection in MuSeCLIR.



(a) Frequency distribution of different sense counts.

(b) Entropy of sense distribution across different number of sense.

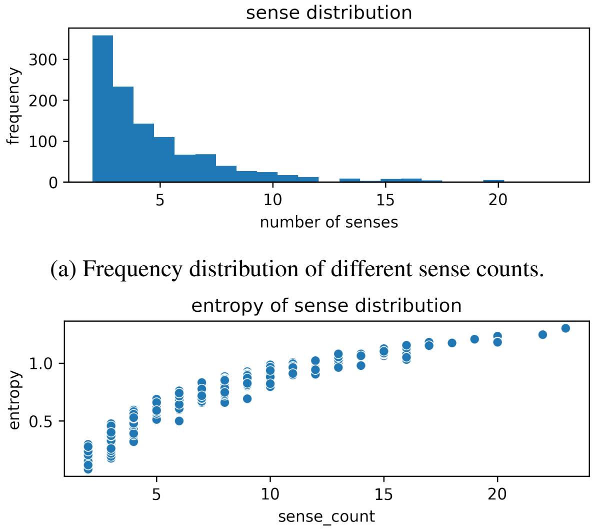
Figure 2: Plots of the German collection in MuSeCLIR.



(a) Frequency distribution of different sense counts.

(b) Entropy of sense distribution across different number of sense.

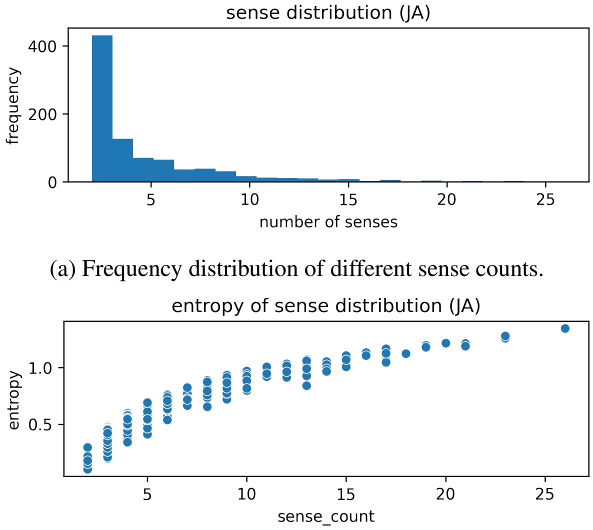
Figure 3: Plots of the Italian collection in MuSeCLIR.



(a) Frequency distribution of different sense counts.

(b) Entropy of sense distribution across different number of sense.

Figure 4: Plots of the Chinese collection in MuSeCLIR.



(a) Frequency distribution of different sense counts.

(b) Entropy of sense distribution across different number of sense.

Figure 5: Plots of the Japanese collection in MuSeCLIR.