



# IR like a SIR

## Sense-enhanced Information Retrieval for Multiple Languages

**Rexhina Blloshmi**  
Sapienza University of Rome  
blloshmi@di.uniroma1.it

**Tommaso Pasini**  
University of Copenhagen  
tommaso.pasini@di.ku.dk

**Niccolò Campolungo**  
Sapienza University of Rome  
campolungo@di.uniroma1.it

**Somnath Banerjee**  
University of Milano-Bicocca  
somnath.banerjee@unimib.it

**Roberto Navigli**  
Sapienza University of Rome  
navigli@diag.uniroma1.it

**Gabriella Pasi**  
University of Milano-Bicocca  
gabriella.pasi@unimib.it

### Abstract

With the advent of contextualized embeddings, attention towards neural ranking approaches for Information Retrieval increased considerably. However, two aspects have remained largely neglected: i) queries usually consist of few keywords only, which increases ambiguity and makes their contextualization harder, and ii) performing neural ranking on non-English documents is still cumbersome due to shortage of labeled datasets. In this paper we present SIR (*Sense-enhanced Information Retrieval*) to mitigate both problems by leveraging word sense information. At the core of our approach lies a novel multilingual query expansion mechanism based on Word Sense Disambiguation that provides sense definitions as additional semantic information for the query. Importantly, we use senses as a bridge across languages, thus allowing our model to perform considerably better than its supervised and unsupervised alternatives across French, German, Italian and Spanish languages on several CLEF benchmarks, while being trained on English Robust04 data only. We release SIR at <https://github.com/SapienzaNLP/sir>.

### 1 Introduction

Information Retrieval (IR) is the task of retrieving from a large collection of unstructured information — generally textual documents — those items deemed relevant to users, and which are expressed by a query — typically a few keywords.

IR systems have become an integral part of our daily lives, as Web Search engines testify, by allowing us to address distinct search tasks. Relevance is the key notion in IR: indeed, the core component of an IR system is the ranking module, which estimates the relevance of a document to a given

query. This is achieved through a ranking function that complies with an underlying formal modeling such as the Vector Space Model, probabilistic models and, more recently, neural models (Guo et al., 2020). Lately, IR systems have begun taking advantage of these latter models, whose aim is learning continuous representations capable of grasping the semantics of the text, as opposed to the traditional lexical approaches comprising the bag-of-words representation. In this new line of research, following the success of neural models in several Natural Language Processing (NLP) tasks, researchers employed contextualized word representations (Devlin et al., 2019; Conneau et al., 2020) in IR to capture semantic aspects of texts (for query and documents) which prove beneficial to ranking approaches (MacAvaney et al., 2019, 2020b). Moreover, thanks to the unsupervised training strategy of contextualized language models, i.e., Masked Language Modeling, it is feasible to train multilingual models which are able to encode sentences across languages within the same semantic space.

Nonetheless, there are challenges peculiar to IR that may hinder the effectiveness of contextualized embeddings. For example, queries are typically composed of just a few keywords, which may not be sufficient to assess the relevance of documents to a query effectively. In classical IR, the technique of query expansion is employed to provide more context about users' actual needs (Rocchio, 1971), by exploiting synonymous terms to overcome the vocabulary mismatch problem. However, this is not suitable for neural language models which are trained to process well-formed sentences. This issue is even more pronounced when dealing with languages other than English, where the lack of training data hinders the use of machine learning

in the multilingual setting.

Recently, Word Sense Disambiguation (WSD) has received greatly increased attention (Bevilacqua et al., 2021), reporting large improvements not only in English (Lacerra et al., 2020; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020; Conia and Navigli, 2021; Barba et al., 2021a), but also across other languages (Scarlini et al., 2020; Procopio et al., 2021). We argue that word senses, thanks to their glosses, i.e., sentences defining word meanings, can provide valuable information to enrich the input query and to aid retrieving relevant documents that are semantically related. Moreover, multilingual sense vocabularies (where concepts are lexicalized with synonymous words in different languages) may provide a bridge across languages, leading neural models to perform better in a zero-shot setting.

Based on these hypotheses, this paper makes the following contributions:

1. we introduce, for the first time, a neural approach to augment the input query with sentences defining the meanings of the words therein,
2. we present SIR, a supervised neural architecture leveraging additional semantic information for the monolingual ad-hoc Information Retrieval task, and
3. we perform an extensive evaluation in English and across several test collections on French, German, Italian and Spanish in a zero-shot setting.

Our findings show that word definitions are indeed beneficial to the task, allowing SIR to better contextualize queries and thus match more relevant documents in respect of all its baselines.

## 2 Related Work

Information Retrieval approaches have long relied on simple statistical metrics based on term frequency, such as TF-IDF and BM25 (Robertson et al., 1996), to represent texts and to match documents against a given query. These methods are still used as strong baselines nowadays (Lin, 2019), especially because they perform retrieval in an unsupervised and efficient way.

In the last decade, two different kinds of neural approaches to IR have been defined (Mitra and Craswell, 2018): the first aims at encoding queries

and documents within the same vector space (Mitra et al., 2016; Huang et al., 2013); the second, instead, focuses on learning an estimator for the relevance of a document with respect to a query (Guo et al., 2016). More recently, with the advent of transformer-based language models such as BERT (Devlin et al., 2019), contextualized representations rapidly got incorporated into retrieval models (MacAvaney et al., 2019) — which previously had relied on static embeddings only — mainly by pairing contextualized models with a binary classifier to compute a score per query-document (MacAvaney et al., 2019; Nogueira and Cho, 2019) or query-sentence pair (Akkalyoncu Yilmaz et al., 2019; Dai and Callan, 2019). Nevertheless, most of the supervised works focused on the English retrieval task, where enough labeled data are available to train a neural model. Instead, while datasets in languages other than English do exist in several tracks of TREC (Braschler et al., 2000; Oard and Gey, 2002) or CLEF (Braschler, 2003), they are rather small and not suitable for training deep neural networks. In this setting, multilingual pretrained language models came out as an effective solution, and showed themselves able to successfully leverage annotations in one language (typically English) and perform retrieval in other languages, e.g., Arabic, Mandarin, and Spanish (MacAvaney et al., 2020b) or Chinese, Arabic, French, Hindi and Bengali (Shi et al., 2020).

However, by relying on large pretrained language models, these approaches assume that queries are expressive enough to model their underlying semantics, which is not always the case. This is a long-standing issue in IR, and one which has stimulated extensive research for years. Different approaches to query expansion such as Markov chains (Metzler and Croft, 2007), term classification (Cao et al., 2008), and static word embeddings (Diaz et al., 2016; Zamani and Croft, 2016) have been applied effectively to improve query representation. More recently, researchers have tried to tackle the problem from the opposite perspective by expanding documents (Nogueira et al., 2019; Nogueira and Lin, 2019; Raffel et al., 2020) or single passages (MacAvaney et al., 2020a) with relevant queries. Nevertheless, most approaches to query expansion do not cope well with the issue of word ambiguity, which makes it hard to model and represent words when not enough context is provided.

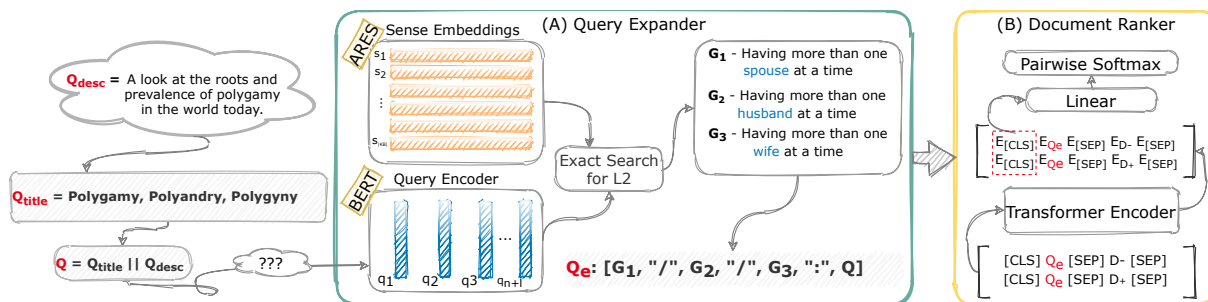


Figure 1: Illustration of SIR: The query  $Q$  composed by a title and a description is shown in the leftmost side; (A) The Query Expander module defines the potentially ambiguous query title terms by retrieving their sense glosses and composes  $Q_e$ ; (B) The Document Ranker module takes the expanded query, combines it separately with a relevant (D+) and a non relevant (D-) document and is trained to optimize pairwise cross-entropy loss.

Word Sense Disambiguation (WSD) is specifically tailored to resolve this issue, and several attempts were made in the past to include word senses within IR pipelines. These early attempts, unfortunately, did not produce encouraging results (Krovetz and Croft, 1992; Voorhees, 1993; Sanderson, 1994). Indeed, Sanderson (2000) emphasised that the effectiveness of WSD integration was diminished by the inaccuracies in disambiguation. A little over a decade later, instead, Zhong and Ng (2012) presented a successful application of WSD in IR by incorporating word senses and synonym relations into a language modeling approach. In addition, further developments over the years led to the remarkable performance attained by modern WSD models, which now perform close to the inter-annotator agreement upper bound (Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020; Barba et al., 2021a,b). This makes us optimistic that these models are finally suitable to be used within downstream tasks.

Differently from previous works, in this paper, we explore this possibility and focus on enriching the query context by devising a neural approach to first retrieve word senses for the input query terms, and then encode their definitions together with query and documents to perform end-to-end document ranking. To the best of our knowledge, this is the first time a Word Sense Disambiguation approach has been employed to expand the query with sense definitions and we show that this is not only beneficial in the monolingual setting but also cross-lingual zero-shot settings.

### 3 Preliminaries

In this Section, we describe the task we are tackling and the resources we exploit.

#### 3.1 Task

We focus on the task of ranking documents given a query, i.e., a topic composed of a title and a description. More formally, let  $Q_{title} = [t_1, \dots, t_n]$  be the sequence of  $n$  terms of the topic title,<sup>1</sup>  $Q_{desc} = [d_1, \dots, d_l]$  the sequence of  $l$  words describing the topic, and  $\mathcal{C}$  a collection of documents. The retrieval task we focus on consists of learning a scoring function  $S_\theta(Q, D) \forall D \in \mathcal{C}$ , to rank documents in the collection according to their relevance to the query  $Q$ , where  $Q = Q_{title} || Q_{desc} = [q_1, \dots, q_{n+l}]$ , i.e., the concatenation of  $Q_{title}$  and  $Q_{desc}$  and  $\theta$  denotes model parameters.

#### 3.2 Resources

In our approach we make use of BabelNet<sup>2</sup> (Navigli and Ponzetto, 2010; Navigli et al., 2021) as vocabulary of senses. BabelNet is a multilingual knowledge base, which organizes word meanings — namely senses — into synsets, i.e., sets of synonyms that express a common concept in different languages (up to 500). Each synset within BabelNet is associated with different glosses in multiple languages<sup>3</sup> that describe its meaning.

## 4 SIR

### 4.1 Motivation

While previous works have focused on expanding the query with related terms in a two-pass re-ranking procedure, we argue that providing sense definitions related to the input query would be more effective for injecting semantics into neural models. Consider the example in Figure 1.

<sup>1</sup>Usually  $n \leq 3$ .

<sup>2</sup>Version 4.0.

<sup>3</sup>Glosses may come from different sources, such as WordNet (Miller, 1992) and Wikipedia.

The query  $Q_{title}$  consists of three terms only, i.e., *Polygamy*, *Polyandry*, and *Polygyny*, and it is not a well-formed sentence. The query description, i.e.,  $Q_{desc} = A\ look\ at\ the\ roots\ and\ prevalence\ of\ polygamy\ in\ the\ world\ today$  in the example, has proved to be useful in enabling neural models to better represent the input query (Dai and Callan, 2019), as it describes the kind of documents to be retrieved. Therefore, we further leverage this information to also retrieve sense definitions related to the terms within the title through a system for Word Sense Disambiguation. For example, given the title and its description, we can add the following sense definitions: i) Having more than one spouse at a time, ii) Having more than one husband at a time, and iii) Having more than one wife at a time, which explicitly define the meaning of each query term.

With this in mind, in this Section we introduce SIR, our approach to *Sense-enhanced Information Retrieval*. SIR is divided into two steps: i) *expand* (§ 4.2), where we employ a multilingual neural model to expand the input query (Figure 1, A), and ii) *rank* (§ 4.3), where the actual document scoring takes place (Figure 1, B).

## 4.2 Query Expander

Inspired by multiple retrieval-augmented approaches for NLP (Guu et al., 2020; Lewis et al., 2020), we enrich the query with the definitions of the senses that are most closely related to its terms, which we collect by means of a learned *sense gloss retriever* component. To this end, we leverage a simple yet effective 1-Nearest-Neighbours (1-NN) approach between the query contextualized word embeddings and sense vectors for BabelNet concepts. As representations for word senses, we use ARES (Scarlini et al., 2020), which provides English and multilingual sense embeddings for all BabelNet synsets containing a WordNet sense.<sup>4</sup> This choice is motivated by three reasons:

- ARES embeddings have been successfully applied to English and multilingual WSD with a simple 1-NN algorithm, achieving state-of-the-art performances;<sup>5</sup>
- the ARES embedding space is comparable to that of BERT (Devlin et al., 2019);

<sup>4</sup>Embeddings can be downloaded at [sensebert.org/#ares](https://sensebert.org/#ares).

<sup>5</sup>ARES results in multilingual all-words WSD are reported in Appendix B.

- the linkage of ARES with BabelNet allows us to easily collect sense definitions in different languages.

To represent query terms  $q_i \in Q$ , instead, we use BERT as its representations are comparable to those of ARES, thus making the retrieval easy and without any need for training. Indeed, in order to retrieve the senses — and thus the definitions — that are closely related to a query  $Q$ , we first feed it through BERT and extract the representations for each word  $q_i$  therein. Then, for each term of the query title, i.e.,  $q_i$ ,  $i \leq n$ , we retrieve the sense with the closest vector in terms of L2 distance.<sup>6</sup>

To avoid the query becoming excessively long, we retain only the top- $k$  closest senses according to their L2 distance, where  $k = \min(m, n)$  and  $m$  is a hyperparameter of the system. For each sense  $s_i \in [s_1, \dots, s_k]$ <sup>7</sup> that we retain, we collect its gloss  $G^i$  in the language of interest from BabelNet.

Finally, we build our expanded query by prepending every gloss  $G^i$  to  $Q$ , i.e.,  $Q_e = [g_1^1, \dots, g_{|G^1|}^1, \dots, g_1^k, \dots, g_{|G^k|}^k, q_1, \dots, q_{n+l}]$ , where  $g_i^j$  represents the  $i$ -th token of the gloss associated with the  $j$ -th closest sense.

## 4.3 Document Ranker

After the query expansion step, we use the enriched query in a Document Ranker module. While our approach can be used in combination with any document ranker, in this paper we employ a popular neural ranking model from the literature based on BERT, i.e., VanillaBERT (MacAvaney et al., 2019), which has been applied to both English and multilingual zero-shot IR settings (MacAvaney et al., 2020b). In Figure 1 (B) we schematize the Document Ranker architecture. Following VanillaBERT, we finetune a pretrained BERT Transformer model for learning the query-document scoring function. The input to the model is formatted following the standard practice, i.e.,  $[CLS] Q_e [SEP] D [SEP]$ , while the ranking score is produced by projecting the vector of the  $[CLS]$  token through a dense layer. The model is trained using a pairwise cross-entropy loss between a relevant and a non-relevant document for the query, which leads the model to rank the relevant document always higher than the non-relevant one. More formally, given a triple  $(Q_e, D+, D-)$ , where document  $D+$  is ranked

<sup>6</sup>Refer to Appendix A for further details on sense retrieval.

<sup>7</sup> $s_1$  denotes the closest sense while  $s_k$  the farthest.

	CLEF 2000-2003				CLEF 2004-2008		
	2000	2001	2002	2003	2004	2005	2006
French	34	49	50	52	49	50	49
German	37	49	50	56	-	-	-
Italian	34	47	49	51	-	-	-
Spanish	-	49	50	57	-	-	-

Table 1: Number of queries for each non-English test benchmark from CLEF collections.

higher than document  $D-$ , the model is trained to optimize the loss function:

$$\mathcal{L}_\theta(Q_e, D+, D-) = \max(0, 1 - \mathbf{S}_\theta(Q_e, D+) + \mathbf{S}_\theta(Q_e, D-))$$

where  $\theta$  denotes the parameters of the model and  $S_\theta(\cdot, \cdot)$  is the ranking function that we are learning. At inference time, given a query, we score all documents in the collection and rank them accordingly.

## 5 Experiments

In this Section, we describe the baselines we compare our approach with, as well as the tasks and datasets used for training and evaluating them.

### 5.1 Experimental Setup

We focus on the monolingual English and non-English Information Retrieval tasks. However, due to the lack of large non-English labeled datasets suitable for training neural ranking models, we follow the zero-shot setting proposed by MacAvaney et al. (2020b), i.e., zero-shot cross-lingual ranking. In this setting, the training of the model is done in a language for which there exists enough relevance-labeled data, i.e., English, and it is tested on queries and documents written in other languages.<sup>8</sup>

**Datasets.** We evaluate SIR using two different datasets: i) TREC Robust 2004 for English retrieval, and ii) the standard collections of CLEF 2000-2003<sup>9</sup> ad-hoc and CLEF 2004-2008<sup>10</sup> ad-hoc News retrieval Test Suites, from which we consider the French, German, Italian and Spanish monolingual tasks. These sum up to 18 non-English evaluation benchmarks. In addition, we report the results in the aggregation of the queries for each language in the respective CLEF campaigns, henceforth denoted as ALL. For English retrieval experiments we use the TREC Robust 2004 dataset

(Voorhees, 2004, Robust04) and create (query, document) pairs for training by following MacAvaney et al. (2019), i.e., by considering a document as relevant only if it is among the top- $k$  retrieved by BM25 and non-relevant otherwise. Robust04 consists of 249 queries on which we perform five-fold cross-validation, using 3 folds for training, 1 for validation and the remaining 1 for testing. We use the splits reported in Huston and Croft (2014).<sup>11</sup> For non-English retrieval experiments, we follow MacAvaney et al. (2020b) and use 4 folds of TREC Robust 2004 for training and the remaining 1 fold for validation. The evaluation, instead, is performed on the CLEF test sets listed above.<sup>12</sup> We report the number of queries for each test collection in Table 1.

**Comparison systems.** We compare SIR with BM25 and BM25+RM3 query expansion as implemented in the Anserini toolkit (Yang et al., 2018), using the default parameters. Our main competitor is VanillaBERT,<sup>13</sup> which has the same underlying neural ranking model as SIR, with the exception of our Query Expander module. This comparison allows us to clearly measure the impact of sense glosses on the document ranking task. As for the non-English setting, we evaluate two versions of SIR: i) SIR<sub>EN</sub> which augments the query with the English glosses of the retrieved senses, and ii) SIR<sub>TL</sub> which concatenates to the non-English query the glosses of the retrieved senses in the target language, when applicable.<sup>14</sup> Interestingly enough, in this setting, switching from SIR<sub>EN</sub> to SIR<sub>TL</sub> comes at no cost, since we rely on a multilingual knowledge base, i.e., BabelNet. To remain consistent with the non-English setting, we consider only English glosses during training (since query language is always in English), and feed SIR<sub>TL</sub> with glosses in other languages at inference time only.

**Training and hyperparameters.** The SIR model relies on two BERT Transformer models, one for the Query Expander, to encode the query, and another one for the Document Ranker component, to encode the query-document pair. We use BERT as query encoder so as to create

<sup>11</sup>We use the folds in Table 1 of Huston and Croft (2014).

<sup>12</sup>We do not evaluate in the multilingual TREC benchmarks as in MacAvaney et al. (2020b) due to unavailability of the data. Instead we run their released code in CLEF Test Suite for comparison.

<sup>13</sup>[github.com/Georgetown-IR-Lab/cedr](https://github.com/Georgetown-IR-Lab/cedr)

<sup>14</sup>When there is no available gloss in the language of the query, we fallback to the English gloss.

<sup>8</sup>Note that queries and documents are in the same language.

<sup>9</sup>[www.islrn.org/resources/317-005-302-361-6/](http://www.islrn.org/resources/317-005-302-361-6/)

<sup>10</sup>[www.islrn.org/resources/378-279-085-589-0/](http://www.islrn.org/resources/378-279-085-589-0/)

	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5		ALL	
	P@20	MAP	P@20	MAP	P@20	MAP	P@20	MAP	P@20	MAP	P@20	MAP
BM25	0.338	0.234	0.318	0.233	0.403	0.230	0.359	0.220	0.375	0.202	0.359	0.224
VanillaBERT	0.413	0.253	0.409	0.262	0.461	0.260	0.440	0.265	0.452	<b>0.241</b>	0.435	0.256
SIR <sub>EN</sub>	<b>0.420</b>	<b>0.271</b>	<b>0.422</b>	<b>0.264</b>	<b>0.466</b>	<b>0.270</b>	<b>0.458</b>	<b>0.270</b>	<b>0.456</b>	0.237	<b>0.444</b>	<b>0.262</b>

Table 2: Results on each fold of TREC Robust04 for English retrieval: SIR<sub>EN</sub> outperforms VanillaBERT in both P@20 and MAP score, with larger gains in separate folds but also in ALL. Best per metric column in **bold**.

contextualized word representations that are comparable to those of ARES (see §4.2). Specifically, we use `bert-large-cased` for English and `bert-base-multilingual-cased` for other languages. Since ARES representations are conceived and computed to be in the same space as BERT representations, we do not need to train the query encoder, but rather we simply employ a 1-NN strategy. That is, for each query term encoded through BERT, we retrieve the sense (and thus the gloss) with the most similar vector. We then retain only the top  $m$  glosses to be considered for a query, and set  $m = 3$  as that is the average number of query terms in Robust04.

For the Document Ranker component, we follow MacAvaney et al. (2019, 2020b) and fine-tune a `bert-base-uncased` model for English, and `bert-base-multilingual-cased` for all the other languages of the non-English tasks. Both VanillaBERT and SIR take as input query the concatenation of the query title and its description, and the first 800 tokens of a document. We limit the maximum number of tokens for a query to 100, while for the expanded query, we additionally consider a maximum number of 100 tokens for the retrieved glosses. We choose the best model by monitoring precision@20 (P@20) score in the validation set. We include more implementation details in Appendices C and D.

**Evaluation.** To evaluate SIR and VanillaBERT models we consider the top 150 documents returned by the term-weighting unsupervised algorithms, i.e., BM25, in the English retrieval task — following MacAvaney et al. (2019), and BM25+RM3 for non-English tasks. RM3 (Abdul-Jaleel et al., 2004) shows consistent improvements over BM25, reinforcing the claims as to the effectiveness of query expansion mechanisms. Therefore, re-ranking BM25+RM3 results shows whether both VanillaBERT and SIR are able to improve the ranking even when the baseline considers extra terms for the query. We use P@20 and mean

average precision (MAP) metrics computed with the official `trec_eval`<sup>15</sup> tool to evaluate the performance of participating systems.

## 5.2 Results

**English.** In Table 2 we report the ranking results in Robust04 benchmark. Firstly, both neural re-ranking approaches, i.e., VanillaBERT and SIR<sub>EN</sub>, significantly<sup>16</sup> outperform the BM25 baseline. This result is in line with the previously reported findings in the literature. More importantly, SIR<sub>EN</sub> attains better performances than VanillaBERT, in almost all folds, both in terms of P@20 and MAP. Across all folds, we observe *relative*<sup>17</sup> improvements in MAP score from 4% to 7% (folds 1 and 3), with an overall improvement of 2.4% in ALL. As for P@20 instead, SIR<sub>EN</sub> improves VanillaBERT by 3% and 4% in folds 2 and 4, with an overall improvement of 2.2% in ALL. When considering the highest reachable performance, i.e., perfectly ranking the documents returned by BM25, SIR reduces the error rate of VanillaBERT by 3.8% in P@20 and 3.3% in MAP score overall. This shows that the sense glosses retrieved by our Query Expander (see §4.2) are of high quality and beneficial to the model, aiding to substantially reduce the error rate.

**Non-English.** In Table 3 we report the performances in the CLEF 2000-2003 ad-hoc test collections. In this setting, we rerank the documents returned by BM25+RM3, as this latter achieves consistently better performances than BM25 alone. Similarly to the English retrieval task, the re-ranking systems, i.e., VanillaBERT and SIR variants, outperform both baselines in all benchmarks. When considering only the behaviour of SIR variants, we observe that using language-

<sup>15</sup>[github.com/cvangysel/pytrec\\_eval](https://github.com/cvangysel/pytrec_eval)

<sup>16</sup>Throughout §5.2, significance is computed using paired t-test with  $p\text{-value} < 0.05$ .

<sup>17</sup>We use the relative improvement with respect to VanillaBERT to comment the performances throughout this Section, calculated as  $(\text{SIR score} - \text{VanillaBERT score}) / \text{VanillaBERT score}$ .

		2000		2001		2002		2003		ALL	
		P@20	MAP	P@20	MAP	P@20	MAP	P@20	MAP	P@20	MAP
FRENCH	BM25	0.257	0.329	0.311	0.372	0.235	0.277	0.228	0.369	0.279	0.365
	BM25+RM3	0.276	0.317	0.355	0.394	0.285	0.318	0.244	0.370	0.313	0.379
	VanillaBERT	<b>0.297</b>	<b>0.383</b>	0.379	0.421	<b>0.318</b>	0.354	0.282	<b>0.487</b>	0.320	0.415
	SIR <sub>EN</sub>	0.287	0.377	<b>0.401</b>	<b>0.468</b>	0.312	<b>0.373</b>	<b>0.291</b>	0.477	<b>0.325</b>	<b>0.428</b>
	SIR <sub>TL</sub>	0.284	0.356	0.394	0.455	0.317	0.368	0.286	0.471	0.322	0.418
GERMAN	BM25	0.239	0.175	0.205	0.084	0.220	0.124	0.098	0.054	0.184	0.103
	BM25+RM3	0.276	0.210	0.227	0.091	0.251	0.154	0.104	0.048	0.207	0.118
	VanillaBERT	0.269	0.253	0.261	0.124	<b>0.291</b>	0.163	<b>0.104</b>	<b>0.057</b>	0.224	0.138
	SIR <sub>EN</sub>	<b>0.289</b>	<b>0.291</b>	0.262	<b>0.135</b>	0.282	<b>0.171</b>	0.102	<b>0.057</b>	<b>0.226</b>	<b>0.150</b>
	SIR <sub>TL</sub>	0.283	0.285	<b>0.263</b>	0.132	0.285	0.169	0.102	<b>0.057</b>	<b>0.226</b>	0.149
ITALIAN	BM25	0.097	0.124	0.247	0.250	0.276	0.253	0.136	0.198	0.195	0.213
	BM25+RM3	0.116	0.161	0.323	0.308	0.331	0.344	0.153	0.199	0.238	0.259
	VanillaBERT	<b>0.128</b>	0.238	0.363	0.328	0.341	0.367	0.154	0.208	0.254	0.288
	SIR <sub>EN</sub>	0.119	0.253	0.371	0.356	<b>0.346</b>	0.373	0.175	<b>0.248</b>	0.262	0.311
	SIR <sub>TL</sub>	0.118	<b>0.259</b>	<b>0.376</b>	<b>0.362</b>	<b>0.346</b>	<b>0.385</b>	<b>0.179</b>	0.244	<b>0.264</b>	<b>0.316</b>
SPANISH	BM25			0.448	0.419	0.427	0.350	0.362	0.359	0.410	0.375
	BM25+RM3			0.485	0.442	0.484	0.408	0.411	0.404	0.458	0.417
	VanillaBERT			0.507	<b>0.489</b>	<b>0.494</b>	<b>0.429</b>	0.430	0.430	0.475	0.448
	SIR <sub>EN</sub>			0.515	0.486	0.492	0.426	<b>0.446</b>	<b>0.453</b>	<b>0.482</b>	<b>0.455</b>
	SIR <sub>TL</sub>			<b>0.516</b>	0.486	0.492	0.419	0.435	0.451	0.479	0.452

Table 3: Results on each year of CLEF 2000-2003 for non-English retrieval: SIR<sub>EN</sub> or SIR<sub>TL</sub> outperform VanillaBERT in both P@20 and MAP score in ALL. Best per metric column in **bold**.

specific glosses (SIR<sub>TL</sub>) does not affect the performance in general. In fact, SIR<sub>TL</sub> shows mostly comparable or slightly worse results than SIR<sub>EN</sub> across all years and measures. This could be due to the fact that the model is trained on English glosses only, which come from a manually-curated English source, i.e., WordNet (see §3.2), whereas non-English glosses come from Wikipedia, which are written in a different style, and are inherently of lower quality and have limited coverage.

When compared to VanillaBERT, SIR<sub>EN</sub> attains better results across the board, showing significant improvements in MAP score on most datasets. More specifically, SIR<sub>EN</sub> improves VanillaBERT baseline by 1.6% in P@20 and 3.1% in MAP score in the ALL dataset of the French language, with significant gains in year 2001. Also, SIR<sub>EN</sub> significantly outperforms VanillaBERT with respect to the overall MAP score, and increases its performance by roughly 1% in P@20 and 8.6% in MAP score in the ALL dataset of the German language, with the largest gain in year 2000. Further-

more, both SIR<sub>EN</sub> and SIR<sub>TL</sub> significantly outperform VanillaBERT in MAP score across the row block of the Italian language, with SIR<sub>TL</sub> showing higher improvements. Indeed, on ALL, it improves the performance of the baseline by roughly 4% in P@20 and 10% in MAP score. Differently from all the other languages, although the contribution of SIR<sub>EN</sub> in Spanish is more modest across years 2001 and 2002, it brings roughly 3.5% and 1.5% improvements in both measures in the 2003 and ALL datasets, respectively.

We continue our evaluation by showing in Table 4 the results in the CLEF 2004-2008 ad-hoc News French monolingual tasks. The behaviour in these benchmarks is similar to that of CLEF 2000-2003, with SIR variants consistently improving over VanillaBERT. Differently from the trend of results in Table 3, SIR<sub>TL</sub> shows slightly higher or comparable performance than SIR<sub>EN</sub>, especially regarding P@20. In comparison to VanillaBERT, the best SIR variant improves its P@20 by 5.6% and MAP by 6.6% in the ALL dataset.

	2004		2005		2006		ALL	
	P@20	MAP	P@20	MAP	P@20	MAP	P@20	MAP
BM25	0.285	0.377	0.375	0.233	0.302	0.272	0.321	0.293
BM25+RM3	0.292	0.403	0.398	0.271	0.332	0.300	0.341	0.324
VanillaBERT	0.308	0.421	0.414	0.276	0.355	0.302	0.359	0.332
SIR <sub>EN</sub>	0.329	<b>0.461</b>	0.441	<b>0.300</b>	0.347	0.302	0.373	<b>0.354</b>
SIR <sub>TL</sub>	<b>0.336</b>	0.426	<b>0.443</b>	0.295	<b>0.356</b>	<b>0.318</b>	<b>0.379</b>	0.346

Table 4: Results on each available year of CLEF 2004-2008 for non-English French retrieval: SIR<sub>EN</sub> or SIR<sub>TL</sub> outperform VanillaBERT in both P@20 and MAP score in ALL. Best per metric column in **bold**.

SIR	VB	Query	Term definitions
<b>0.278</b>	0.092	<u>timber exports Asia:</u> What is the extent of U.S. raw timber exports to Asia	✓ <b>Asia</b> - the largest continent with 60% of the earth’s population ✓ <b>exports</b> - sell or transfer abroad ✓ <b>timber</b> - fragments of wood
0.325	<b>0.392</b>	<u>safety plastic surgery:</u> Find documents that discuss the safety of or the hazards of cosmetic plastic surgery.	✓ <b>plastic</b> - generic name for certain synthetic or semisynthetic materials that can be molded [ . . . ] ✗ <b>surgery</b> - the branch of dentistry involving surgical procedures ✗ <b>safety</b> - a safe for storing meat
0.644	<b>0.728</b>	<u>women ordained Church of England: [ . . . ] argu- ments for and against Great Britain’s approval of women being ordained as Church of England priests?</u>	✓ <b>England</b> - a division of the United Kingdom ✓ <b>ordained</b> - appoint to a clerical posts ✓ <b>Church</b> - one of the groups of Christians who have their own beliefs and forms of worship

Table 5: Excerpt of term definitions retrieved by our Query Expander: ✓ **Accurate disambiguations** improve performance by more than ✗ **inaccurate disambiguations** degrade it (upper); Even when accurately disambiguated, retrieval results decrease due to more general information in glosses (lower). VB denotes VanillaBERT.

In summary, the contribution of SIR is mainly evident in the MAP score across both tables, suggesting that gloss information, while not improving by a large margin in P@20, i.e., top retrieved documents, enables the system to return an overall better ranking of all the relevant documents.

### 5.3 Error Analysis

We here provide insights into the cases where the retrieved definitions do indeed help the underlying model in the retrieval tasks. To this end, we manually check the quality of the disambiguation of the query terms, and perform a comparison of VanillaBERT and SIR according to the MAP score per query. More specifically, we compute the absolute difference of MAP between systems for each query in Robust04 and pick the top ones where SIR performs better than VanillaBERT and those where it performs worse. We report an excerpt in Table 5. By inspecting the data we note that, firstly, accurate disambiguation improves performance by

a larger margin than inaccurate disambiguation degrades it. This phenomenon can be attributed to the ability of the Document Ranker to ignore noisy input while benefiting from useful extra information, and this appears to be so in the majority of cases, as demonstrated by our experimental results (see §5.2). Secondly, we notice that there are some disambiguation mistakes: we attribute this issue to the absence of any mechanism restricting the possible senses for a given word, since we base our retrieval only on representations’ L2 distance (see §4.2). For instance, the words *safety* and *surgery* in Table 5 are associated with glosses that are somehow related to but that are not specific to any sense of the target words. While this issue can be alleviated by filtering the possible senses for a word, similarly to the standard WSD task, we decide not to do so as it would require lemmatizing and POS tagging the input query and we want to keep the approach as end-to-end and scalable across languages as possible.





Alternatively, more recent WSD approaches could be useful and we leave this extensive study for future work. Another source of error, even though less frequent, concerns SIR’s failure to outperform VanillaBERT even when the disambiguation of its terms is accurate. We inspect the possible reasons behind errors of this kind by checking the top documents retrieved by each system and provide these in the second row block of Table 5. Although the retrieved glosses are factually correct for the query words, the gloss for *women* has been discarded as scoring lower than the top 3 senses in the sense retrieval step, thus the highest ranked documents generally focused on the *Church of England* rather than on *women in the Church*. This issue requires further investigation and analysis. A possible direction for future work would be to identify the most peculiar terms within the query and ensure that their definitions are included in its expanded version.

## 6 Conclusions

In this paper we presented SIR, a novel approach for ranking documents in multiple languages. Our approach is the first to take advantage of a WSD model to expand the input query with sense definitions as additional semantic information. By evaluating SIR on multiple gold Information Retrieval benchmarks across languages, we show that our approach consistently improves over its main competitors that do not have access to sense glosses, thus demonstrating that such information is beneficial for the English retrieval task, as well as in the zero-shot cross-lingual setting. In addition, through a simple qualitative analysis, we highlight the advantages and disadvantages of SIR, suggesting promising directions for better utilizing WSD to improve IR models. We release SIR at <https://github.com/SapienzaNLP/sir> to ease future research in this direction.

## Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the  ELEXIS project No. 731015 under the European Union’s Horizon 2020 research and innovation programme. 

This work was partially supported by the MIUR under the grant “Dipartimenti di eccellenza 2018-

2022” of the Department of Computer Science of Sapienza University and by the PerLIR project (Personal Linguistic resources in Information Retrieval) funded by the MIUR Progetti di ricerca di Rilevante Interesse Nazionale programme (PRIN 2017).

## References

- Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. [Umass at trec 2004: Novelty and hard](#). *Computer Science Department Faculty Publication Series*, page 189.
- Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. [Cross-domain modeling of sentence-level evidence for document retrieval](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China. Association for Computational Linguistics.
- Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. [ESC: Redesigning WSD with extractive sense comprehension](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.
- Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. [ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Michele Bevilacqua and Roberto Navigli. 2020. [Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. [Recent trends in word sense disambiguation: A survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Terra Blevins and Luke Zettlemoyer. 2020. [Moving down the long tail of word sense disambiguation with gloss informed bi-encoders](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

- Martin Braschler. 2003. [CLEF 2003—Overview of results](#). In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 44–63. Springer.
- Martin Braschler, Carol Peters, and Peter Schäuble. 2000. [Cross-Language Information Retrieval \(CLIR\) Track Overview](#). In *Text REtrieval Conference (TREC) 2000*.
- Guihong Cao, Stephen Robertson, and Jian-Yun Nie. 2008. [Selecting query term alternations for web search by exploiting query contexts](#). In *Proceedings of ACL-08: HLT*, pages 148–155, Columbus, Ohio. Association for Computational Linguistics.
- Simone Conia and Roberto Navigli. 2021. [Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Zhuyun Dai and Jamie Callan. 2019. [Deeper text understanding for IR with contextual neural language modeling](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 985–988. ACM.
- 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.
- Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. [Query expansion with locally-trained word embeddings](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 367–377, Berlin, Germany. Association for Computational Linguistics.
- Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. [A deep relevance matching model for ad-hoc retrieval](#). In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 55–64. ACM.
- Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. [A deep look into neural ranking models for information retrieval](#). *Information Processing and Management*, 57(6):102067.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. [REALM: retrieval-augmented language model pre-training](#). *ArXiv preprint*, abs/2002.08909.
- Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. [Learning deep structured semantic models for web search using clickthrough data](#). In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 2333–2338. ACM.
- Samuel Huston and W. Bruce Croft. 2014. [Parameters learned in the comparison of retrieval models using term dependencies](#). *Online preprint*.
- Jeff Johnson, M. Douze, and H. Jégou. 2021. [Billion-scale similarity search with gpus](#). *IEEE Transactions on Big Data*, 7:535–547.
- Robert Krovetz and W. Bruce Croft. 1992. [Lexical ambiguity and information retrieval](#). *ACM Transactions on Information Systems*, 10(2):115–141.
- Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. [CSI: A coarse sense inventory for 85% word sense disambiguation](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8123–8130. AAAI Press.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Jimmy Lin. 2019. [The neural hype and comparisons against weak baselines](#). *ACM SIGIR Forum*, 52(2):40–51.
- Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonello, Nazli Goharian, and Ophir Frieder. 2020a. [Expansion via prediction of importance with contextualization](#). In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1573–1576. ACM.

- Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020b. [Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning](#). In *Proceedings of the 42nd European Conference on Information Retrieval Research*, pages 246–254.
- 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 2019, Paris, France, July 21-25, 2019*, pages 1101–1104. ACM.
- Donald Metzler and W. Bruce Croft. 2007. [Latent concept expansion using markov random fields](#). In *Proceedings of the 30th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR) 2007*, page 311–318, New York, USA. Association for Computing Machinery.
- George A. Miller. 1992. [WordNet: A lexical database for English](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Bhaskar Mitra and Nick Craswell. 2018. [An introduction to neural information retrieval](#). *Foundations and Trends® in Information Retrieval*, 13(1):1–126.
- Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. 2016. [A dual embedding space model for document ranking](#). *ArXiv preprint*, abs/1602.01137.
- Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Ceconi. 2021. [Ten Years of BabelNet: A Survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567. ijcai.org.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. [BabelNet: Building a very large multilingual semantic network](#). In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage Re-ranking with BERT](#). *ArXiv preprint*, abs/1901.04085.
- Rodrigo Nogueira and Jimmy Lin. 2019. [From doc2query to docTTTTTquery](#). *Online preprint*.
- Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. [Document expansion by query prediction](#). *ArXiv preprint*, abs/1904.08375.
- Douglas W Oard and Fredric C Gey. 2002. The TREC 2002 arabic/english clir track. In *Text REtrieval Conference (TREC) 2002*.
- Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 2021. [MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3915–3921. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. 1996. [Okapi at TREC-4](#). *Nist Special Publication Sp*, pages 73–96.
- Joseph John Rocchio. 1971. [The smart retrieval system: Experiments in automatic document processing](#). *Relevance feedback in information retrieval*, pages 313–323.
- Mark Sanderson. 1994. [Word sense disambiguation and information retrieval](#). In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR) 1994*, pages 142–151, London. Springer London.
- Mark Sanderson. 2000. [Retrieving with good sense](#). *Information Retrieval*, 2(1):49–69.
- Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. [With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.
- Peng Shi, He Bai, and Jimmy Lin. 2020. [Cross-lingual training of neural models for document ranking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, Online. Association for Computational Linguistics.
- Ellen Voorhees. 2004. [Overview of the trec 2004 robust retrieval track](#). In *Text Retrieval Conference (TREC) 2004*.
- Ellen M. Voorhees. 1993. [Using wordnet to disambiguate word senses for text retrieval](#). In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93*, page 171–180, New York, USA. Association for Computing Machinery.
- Peilin Yang, Hui Fang, and Jimmy Lin. 2018. [Anserini: Reproducible ranking baselines using lucene](#). *J. Data and Information Quality*, 10(4).

Language	F <sub>1</sub> score
English	77.3
French	81.2
German	79.6
Italian	77.0
Spanish	75.3

Table 6: ARES performance in SemEval-2013 benchmark of all-words multilingual WSD.

Hamed Zamani and W Bruce Croft. 2016. [Embedding-based query language models](#). In *Proceedings of the 2016 ACM international conference on the theory of information retrieval*, pages 147–156.

Zhi Zhong and Hwee Tou Ng. 2012. [Word sense disambiguation improves information retrieval](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–282, Jeju Island, Korea. Association for Computational Linguistics.

## A Sense Retrieval

As per the standard practice, we tokenize  $Q$  by applying wordpiece tokenization, adding the [CLS] prefix and the [SEP] suffix. Following ARES, we represent the query term vector  $\mathbf{V}_{q_i}$  as the sum of the BERT representations of the last four hidden layers, and average the wordpiece vectors belonging to the same query term. Moreover, since ARES vectors are composed of two stacked BERT representations, we concatenate  $\mathbf{V}_{q_i}$  with itself. We search the most related senses for each term  $q_i$  within the query, first normalizing  $q_i$  vectors and those of ARES and then employing L2 distance search index provided by the FAISS (Johnson et al., 2021) library.

## B ARES WSD Performance

In Table 6 we show the results obtained by ARES in the SemEval-2013 benchmark of all-words WSD task in different languages as reported by Scarlini et al. (2020). We choose to report SemEval-2013 only as it comprises all the languages of interest. We direct the reader to Scarlini et al. (2020) for the complete evaluation of ARES in WSD.

## C Document Ranker Details

For the Document Ranker component, we follow MacAvaney et al. (2019, 2020b) and finetune a bert-base-uncased model for English, and bert-base-multilingual-cased for all

the non-English tasks. Both VanillaBERT and SIR take as input query the concatenation of the query title and its description, and the first 800 tokens of a document. We limit the maximum number of tokens for a query to 100, while for the expanded query, we additionally consider a maximum number of 100 tokens for the retrieved glosses. Since BERT supports 512 tokens, we split longer documents into segments, separately encoding each with the query. Then we average the multiple [CLS] tokens to compute the final query-document pair representation used for classification.

## D Training Hyperparameters

We employ the hyperparameters of VanillaBERT (MacAvaney et al., 2019) on top of which we show the improvements of our contribution. The models are trained with Adam optimizer with learning rate 0.001 for the classifier and  $2 \times 10^{-5}$  for BERT layers. The training process is carried out on a single GPU (Nvidia GeForce GTX 1080Ti), for 100 epochs each of which is trained on 32 batches comprising 16 query-document pairs. We validate by monitoring P@20 and employ early stopping with patience 20 epochs. Training takes 5-10 hours for both VanillaBERT and SIR, depending on whether the early stopping is triggered. VanillaBERT and SIR have 110M and 179M trainable parameters when trained with bert-base-uncased and bert-base-multilingual-cased BERT models<sup>18</sup>, respectively.

<sup>18</sup>We use the pretrained models by pytorch-pretrained-bert library.