Learning to Weight Translations using Ordinal Linear Regression and Query-generated Training Data for Ad-hoc Retrieval with Long Queries

Javid Dadashkarimi **Masoud Jalili Sabet** **Azadeh Shakery**
School of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
{dadashkarimi,jalili.masoud,shakery}@ut.ac.ir

Abstract

Ordinal regression which is known with learning to rank has long been used in information retrieval (IR). Learning to rank algorithms, have been tailored in document ranking, information filtering, and building large aligned corpora successfully. In this paper, we propose to use this algorithm for query modeling in cross-language environments. To this end, first we build a query-generated training data using pseudo-relevant documents to the query and all translation candidates. The pseudo-relevant documents are obtained by top-ranked documents in response to a translation of the original query. The class of each candidate in the training data is determined based on presence/absence of the candidate in the pseudo-relevant documents. We learn an ordinal regression model to score the candidates based on their relevance to the context of the query, and after that, we construct a query-dependent translation model using a softmax function. Finally, we re-weight the query based on the obtained model. Experimental results on French, German, Spanish, and Italian CLEF collections demonstrate that the proposed method achieves better results compared to state-of-the-art cross-language information retrieval methods, particularly in long queries with large training data.

1 Introduction

The multilingual environment of the Web has long required the researchers in information retrieval (IR) to introduce powerful algorithms for bridging the gaps between the languages (Nie, 2010; Ganguly et al., 2012; Dadashkarimi et al., 2016). Generally, these algorithms can be categorized as follows: *(1)* translating the query of the user to the language of the documents (Ganguly et al., 2012), *(2)* translating all of the documents into the language of the user (Oard, 1998), *(3)* translating the query and the documents into a third language (Kishida and Kando, 2005), *(4)* bringing the query and the documents into a shared low-dimensional space (Vulic and Moens, 2015; Dadashkarimi et al., 2016), and *(5)* using semantic/concept networks (Franco-Salvador et al., 2014). Usually the query translation approach has been opted as the most efficient and effective approach in the literature (Vulic and Moens, 2015; Nie, 2010). Ma et al. (2012), have shown that cross-language information retrieval (CLIR) takes more advantage of weighting all translations than selecting the most probable ones. But, building this translation model demands a statistical analysis of translation candidates over an aligned corpus or a single target collection (Talvensaari et al., 2007; Liu et al., 2005; Ganguly et al., 2012).

Aligned corpora have been exploited in CLIR successfully (Rahimi et al., 2016; Talvensaari et al., 2007). But, these resources are either scarce in some languages or specific to a few number of domains. Therefore, recently query-dependent collections have been shown to be more effective and are available to many languages (Dadashkarimi et al., 2016; Ganguly et al., 2012). Pseudo-relevant documents are useful resources to this end. In this paper we propose to use pseudo-relevant documents to build a query-dependent translation model. To this aim, first we take top-ranked documents retrieved in response to a simple translation of the query as a pseudo-relevant collection; we expect relevant translations to appear in the collection by accepting a limited amount of noise. Thus we build a training data based

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 1725–1733, Osaka, Japan, December 11-17 2016.

on presence/absence of the translations in the collection and a number of embedded features. At the next step we aim to learn an ordinal regression model over the translation candidates and then build a translation model for the query using a softmax function. The final model is used in the second retrieval run.

Since this model requires rather large training data, it is expected to be more useful for long queries, where there is enough information about the user intention. Experimental results on French, Spanish, German, and Italian CLEF collections demonstrate that the proposed method performs better than state-of-the-art dictionary-based CLIR methods particularly in long queries.

In Section 2 we provide an overview on related works and then we propose the method and all the formulations in Section 3. Experimental results and related discussions are provided in Section 4. We conclude the paper and provide future works in Section 5.

2 Previous Works

2.1 Query Translation in CLIR

Query translation is opted as an efficient way for bridging the gap between the source language of the query q^s and the language of a target collection $\mathcal{C} = \{d_1, d_2, .., d_{|\mathcal{C}|}\}$ in CLIR (Nie, 2010). In statistical language modeling, a query translation is defined as building a translation model $p(w_t|q_i^s; q^s)$ where w_t is a translation candidate and q_i^s is a query term. Monz and Dorr (2005) introduced an expectation maximization algorithm for estimating this probability: $p(w_t|q_i^s)^n = p(w_t|q_i^s)^{n-1} + \sum_{w_{t'}} a_{w_t,w_{t'}}.p(w_{t'}|q_i^s)$ where $a_{w_t,w_{t'}}$ is a mutual information of a couple of translations. This probability is computed iteratively and then is used for building query model $p(w_t|q^s)$. Dadashkarimi et al. (2014) and Cao et al. (2008), employed similar methods with bigram probabilities $p(w_t|w_{t'})$. On the other hand, Pirkola et al. (2001) introduced structured queries for CLIR in which each translation of a query term can be considered as a member of a synonym set. Structured queries use a number of operators for building this set. For example $\#sum(\#syn(w_1, .., w_k)\#syn(w'_1, .., w'_{k'}))$ treats occurrences of w_t in a document as occurrences of its set and then sums over all the sets for estimating score of a document. There are also selection-based methods that consider only a limited subset of translations in their retrieval task. Nie (2010), demonstrated that these approaches suffer from lower coverage compared to the weighting approaches.

2.2 Pseudo-relevance Feedback for Query Modeling

Top-ranked documents $F = \{d_1, d_2, .., d_{|F|}\}$ in response to the query of a user have long been considered as informative resources for query modeling (Lavrenko and Croft, 2001; Zhai and Lafferty, 2001; Lv and Zhai, 2014). Relevance models are proposed by (Lavrenko et al., 2002; Lavrenko and Croft, 2001) in both monolingual and cross-lingual environments for language modeling. To this end, Zhai and Lafferty (2001) proposed the mixture model for monolingual environments based on an expectation maximization algorithm. Lv and Zhai (2014) proposed a divergence minimization algorithm that outperforms most of the competitive baselines. There are also a further number of powerful algorithms based on machine learning methods in this area (Liu, 2009). Dadashkarimi et al. (2016), employed a divergence minimization framework for pseudo-relevance feedback using embedded features of words from a positive and a negative sample set of feedback documents. Liu et al. (2005), introduced maximum coherence model for query translation whose aim is to estimate overall coherence of translations based on their mutual information. Dadashkarimi et al. (2016), recently published another work for query translation using low-dimensional vectors of feedback terms from a couple of pseudo-relevant collections. The cross-lingual word embedding translation model (CLWETM) first learns the vectors of feedback terms separately and then aims at finding a query dependant transformation matrix $\mathbf{W}$ for projecting the source vectors to their equivalents in the target language. The projected vectors $\mathbf{W}^T\mathbf{v}_w$ are then used to build a translation model for the query. The authors have shown that CLWETM outperforms the state-of-the-art dictionary-based cross-lingual relevance models.

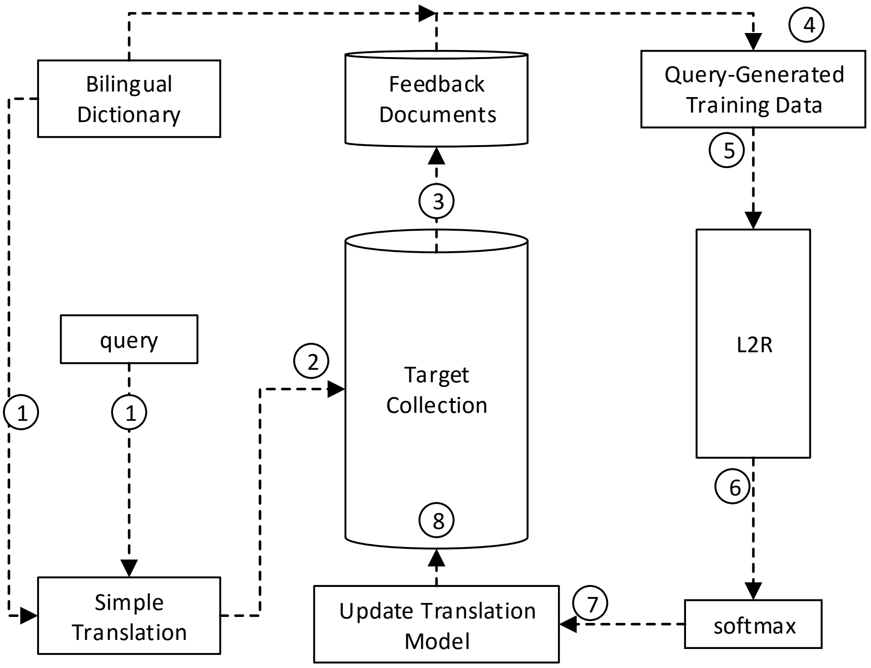

Figure 1: The whole process of building translation model using ordinal linear regression and query-generated training data.

3 Learning to Weight Translations using Query-generated Training Data and Embedded Features

In this section we propose a learning approach for weighting translations of query terms. To this end we first elaborate on building a query-generated training data in Section 3.1. In Section 3.2, we introduce the formulations of the proposed method and finally in Section 3.3 we introduce a number of embedded features used in the learning process.

3.1 Query-generated Training Data for Ordinal Regression

Let $q = \{q_1, .., q_m\}$ be the query and let $q^t = \{w_1, ..w_n\}$ be all the translation candidates of q. We expect correct translations to appear in pseudo-relevant collection F by accepting a limited amount of noise (see Section 2.2). As an example, let the query be $q = \{world, cup, 2018\}$ and assume that $q^t = \{[monde, univers], [coupe, tasse], [2018]\}$ is the set of translation candidates in French. By using a uniform distribution of weights over translation words, $q^t = \{[(1/2, monde), (1/2, univers)], [(1/2, coupe), (1/2, tasse)], [(1, 2018)]\}$ could be a simple query model in the target language. Since $\{monde, coupe, 2018\}$ are conceptually better translations, we expect them to appear in F. Thus, the presence/absence of the translations in F can be indicators of their relevance to the query. We use this information for building a query-generated training data to learn an ordinal regression model for scoring the translations. Let $y_i \in \{-1, +1\}$ indicates the presence/absence of w_i represented by feature vector $\mathbf{x}_i \in \mathbb{R}^n$, and then assume that $D = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^{|\mathbf{x}_i|} \times \{-1, +1\}\}$ is the training data. D is then be used as the training data for our regression model.

3.2 Learning to Rank for Ordinal Translation Regression

We aim to find $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^{|\mathbf{x}|}$ is the weight vector and b is a bias both specific to a query, satisfying the following constraint:

$$f(\mathbf{x}_i) > f(\mathbf{x}_j) \iff y_i > y_j \quad \forall(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in D \tag{1}$$

Table 1: Descriptions of the features in **x**.

Feature	Description
$[\mathbf{u}_{w_j}]_k$	the $k-$th dimension of w_j in its low dimensional vector $\mathbf{u}_{w_j} \in \mathbb{R}^{c\times 1}$
$p(w_j\|C)$	the maximum likelihood probability of w_j in the collection
$p(w_j\|\theta_F)$	the maximum likelihood probability of w_j in the feedback documents
$p(w_j\|q^t)$	the maximum likelihood probability of w_j in the simple translation of the query
$\sum_{w_{j'} \notin q_{w_j}} p(w_j, w_{j'})$	sum of the bi-gram probability of w_j with all translations of $q_{w_{j'}} \neq q_{w_j}$

where $f(\mathbf{x})$ should give higher rank to a pseudo relevant translation w_i compared to a non-relevant translation w_j. If we define the set of all translation words' pairs with $P = \{(i,j) : y_i > y_j\}$, finding $f(\mathbf{x})$ requires minimizing the following loss function:

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad s.t. \ \ \forall(i,j) \in P : (\mathbf{w}^T\mathbf{x}_i) \geq (\mathbf{w}^T\mathbf{x}_j) \tag{2}$$

Generally speaking, Equation 2 shows loss-function of an ordinal regression with parameter **w** (Herbrich et al., 1999; Joachims, 2006). Here, the goal is to score $w \in q^t$ based on the embedded feature vectors $\mathbf{x}_{1:n}$ and build a translation model as follows:

$$p(w_j|q) = \frac{1}{m}\frac{\delta_{w_j} e^{\mathbf{w}^T \mathrm{x}_j + b}}{\sum_{w_{j'}} \delta_{w_{j'}} e^{\mathbf{w}^T \mathrm{x}_{j'} + b}} \tag{3}$$

where δ_{w_j} is a weight function specific to each word and m is the number of query terms. We choose $\delta_{w_j} = c(w_j, F)^{\frac{1}{2}}$ equal to the count of w_j in F to the power of $\frac{1}{2}$. This power is for rewarding rare words and penalizing the common ones (Goldberg and Levy, 2014). Figure 1 shows the whole process of building training data and weighting the translations.

3.3 Embedded Features

In Section 3.1 we proposed a query-dependant training data. In this section, we shed light on **x**, the feature vectors in D. As shown in Table 1, we exploited two categories of features: query-dependent features and query-independent features. $p(w_j|C)$ and $[\mathbf{u}_{w_j}]_k$ are independent of the query and capture the frequency of w_j in the collection and the semantic information of w_j in the target language respectively. On the other hand, the other features are specific to the q. $p(w_j|\theta_F)$ captures frequency of w_j in the pseudo-relevant documents. For example in $q = \{world, cup, 2018\}$, although the frequency of $[tasse]$ in collection is more than $[coupe]$, but in F, $[coupe]$ is a more frequent translation compared to $[tasse]$. $p(w_j|q^t)$ is a useful feature for long queries where there are multiple instances of a topical term in the query. According to (Dadashkarimi et al., 2014; Gao et al., 2005), $\sum_{w_{j'} \notin q_{w_j}} p(w_j, w_{j'})$ captures coherence of w_j with the context of the query.

4 Experiments

4.1 Experimental Settings

Details of the used collections are provided in Table 2. As shown in the table we provided experiments on four European languages. For each collection we experiment on both short queries, derived from title of the topics, and long queries, derived from title and description of the topics. We used Lemur toolkit in all experiments[1]. All the queries and documents are stemmed using the Porter stemmer (Porter, 1997). The collections are also normalized and purified from stopwords[2]. We used Dirichlet smoothing method with prior $\mu = 1000$ in a statistical language modeling framework with KL-divergence similarity measure.

[1] http://www.lemurproject.org/

[2] `http://www.unine.ch/info/clef/`

Table 2: Collection Characteristics

ID	Lang.	Collection	Queries (title+description)	#docs	#qrels
IT	Italy	La Stampa 94, AGZ 94	CLEF 2003-2003, Q:91-140	108,577	4,327
SP	Spanish	EFE 1994	CLEF 2002, Q:91-140	215,738	1,039
DE	German	Frankfurter Rundschau 94, SDA 94, Der Spiegel 94-95	CLEF 2002-03, Q:91-140	225,371	1,938
FR	French	Le Monde 94, SDA French 94-95	CLEF 2002-03, Q:251-350	129,806	3,524

Table 3: Comparison of different query translation methods for short queries. Superscripts 1/2/3/4/5/6 indicate that the MAP improvements over the corresponding methods are statistically significant (2-tail t-test, $p \leq 0.05$). * indicates $0.05 \leq p \leq 0.1$ (compared to the proposed method L2R).

		FR (short)			DE (short)			ES (short)			IT (short)		
	ID	MAP	P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10
1	MONO	0.3262	0.412	0.374	0.2675	0.432	0.369	0.3518	0.496	0.432	0.2949	0.368	0.311
2	TOP-1	0.2211	0.312	0.273	0.2015	0.253	0.233	0.2749	0.367	0.326	0.1566	0.221	0.190
3	UNIF	0.1944	0.269	0.236	0.2148	0.282	0.237	0.236	0.294	0.249	0.1526	0.200	0.156
4	STRUCT	0.1677	0.250	0.226	0.1492	0.227	0.204	0.2472	0.335	0.328	0.0994	0.133	0.118
5	BiCTM	0.2156	0.314	0.275	0.2126	0.282	0.261	0.2652*	0.343	0.316	0.1504	0.217	0.177
6	CLWETM	**0.2312**	**0.331**	0.281	0.2158	0.282	0.255	**0.2915**	**0.384**	**0.337**	0.1630	0.221	**0.194**
7	L2R	0.2296^{2-5}	0.312	**0.288**	$\mathbf{0.2170}^{2-4}$	**0.290**	**0.265**	0.2749^{2-4}	0.380	0.320	$\mathbf{0.1638}^{2-5}$	**0.229**	0.190

The embedding features $[\mathbf{u}_{w_j}]_k$ are computed with word2vec introduced in (Mikolov et al., 2013) on each collection; size of the window, number of negative samples and size of the vectors are set to typical values of 10, 45, and 100 respectively. We also used the svm-rank toolkit for learning $\mathbf{w}$ (Joachims, 2006)[3].

As shown in Table 3 and Table 4 we have the following experimental runs: *(1)* Monolingual retrieval run (MONO). It is the primary comparison baseline for CLIR in the literature (Pirkola et al., 2001; Levow et al., 2005); *(2)* translating by top-ranked translation of a bilingual dictionary (TOP-1) (Ma et al., 2012; Esfahani et al., 2016; Dadashkarimi et al., 2014); *(3)* uniform weighting of translations in the query language modeling (UNIF); *(4)* structured query using $\#syn$ operator as described in Section 2.1 (STRUCT); *(5)* binary coherence translation model (BiCTM) introduced in (Dadashkarimi et al., 2014); cross-lingual word embedding translation model (CLWETM) recently introduced by (Dadashkarimi et al., 2016); and *(6)* the proposed learning to rank (L2R) algorithm. We used the simple STRUCT method for our initial retrieval run to build the query-generated training data as described in Equation 3.1.

4.2 Performance Comparison and Discussion

All the experimental results are provided in Table 3 and Table 4. As shown in Table 3, although L2R outperforms most of the baselines with short queries, the improvements with respect to CLWETM, the most competitive baseline, are marginal. The first reason for these outcomes could be the lower number of training data as shown in Table 6. L2R reaches 70.39%, 81.46%, 78.14%, and 55.54% of performances of the monolingual run in FR, DE, ES, and IT collections respectively.

On the other hand, the proposed L2R outperforms all the baselines with long queries in almost all the metrics. According to Table 4, L2R reaches 77.77%, 70.11%, 77.84%, 61.79% of performance of the monolingual run in FR, DE, ES, and IT collections respectively. Although CLWETM, the state-of-the-art dictionary-based translation model, takes advantage of a couple of collections in the source and target language, L2R successfully outperforms CLWETM with only one collection in the target. Nevertheless, the authors did not exploit comparable corpora for their evaluations and used a pool of multiple news agencies in the source language instead.

[3] `https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

Table 4: Comparison of different query translation methods for long queries. Superscripts 1/2/3/4/5/6 indicate that the MAP improvements over the corresponding methods are statistically significant (2-tail t-test, $p \leq 0.05$). $n - m$ indicates all methods in range $[n, .., m]$.

		FR (long)			DE (long)			ES (long)			IT (long)		
	ID	MAP	P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10
1	MONO	0.4193	0.535	0.473	0.3938	0.528	0.478	0.5281	0.672	0.596	0.3947	0.502	0.436
2	TOP-1	0.3077	0.396	0.343	0.2242	0.308	0.250	0.3762	0.480	0.432	0.2195	0.280	0.262
3	UNIF	0.2709	0.356	0.309	0.2425	0.284	0.254	0.3243	0.368	0.334	0.2095	0.231	0.200
4	STRUCT	0.1800	0.265	0.239	0.2103	0.252	0.250	0.2951	0.400	0.376	0.1942	0.244	0.224
5	BiCTM	0.3050	0.390	0.350	0.2442	0.328	0.278	0.3841	0.464	0.434	0.2172	0.262	0.242
6	CLWETM	0.3167	0.410	0.366	0.2622	0.348	0.308	0.4029	0.500	**0.462**	0.2380	0.298	**0.267**
7	L2R	**0.3261**$^{2-6}$	**0.428**	**0.368**	**0.2761**$^{2-6}$	**0.364**	**0.328**	**0.4111**$^{2-6}$	**0.504**	0.446	**0.2439**$^{2-6}$	**0.302**	0.262

Table 5: Translation model for the English topic 'Brain-Drain Impact' to French.

	UNIF		BiCTM		CLWETM		L2R	
term	candidate	$p(w\|q)$	candidate	$p(w\|q)$	candidate	$p(w\|q)$	candidate	$p(w\|q)$
impact	effet	0.125	effet	0.074646	effet	0.11913	effet	0.143442
impact	impact	0.125	impact	1.35E-03	impact	1.07E-07	impact	0.15437
impact	choc	0.125	choc	1.16E-03	choc	1.07E-07	choc	0.042613
impact	enfonc	0.125	enfonc	4.26E-04	enfonc	1.07E-07	enfonc	0.068032
impact	frapper	0.125	frapper	0.513367	frapper	0.855057	frapper	0.050397
impact	incident	0.125	incident	3.91E-01	incident	1.07E-07	incident	0.377201
impact	porte	0.125	porte	0.017560	porte	0.025813	porte	0.120816
impact	influer	0.125	influer	5.51E-05	influer	1.07E-07	influer	0.04313
brain	tete	0.340	tete	0.999197	tete	0.993176	tete	0.556568
brain	cerveau	0.340	cerveau	0.000758	cerveau	0.003412	cerveau	0.357755
brain	cervelle	0.340	cervelle	4.53E-05	cervelle	0.003412	cervelle	0.085677
drain	pert	0.143	pert	0.192359	pert	0.189706	pert	0.371849
drain	evacu	0.143	evacu	0.227306	evacu	0.216075	evacu	0.318367
drain	epuis	0.143	epuis	0.043371	epuis	0.044900	epuis	0.028666
drain	purg	0.143	purg	0.536827	purg	0.538518	purg	0.112147

Table 5 shows three translation models for the topic 'Brain-Drain Impact' based on UNIF, BiCTM, CLWETM, and L2R. As shown in the table BiCTM and CLWETM are more likely to be trapped in a local optimum. BiCTM originally estimates the query model based on co-occurrences of translations through a collection and thus does not use the pseudo-relevant data. Therefore, it is possible that some translations are co-occurred with each other in the collection but not in a query-dependent collection. On the other hand, CLWETM considers semantic information of the query using low-dimensional vectors of the candidates in top-ranked documents and then combines the obtained translation model with a collection dependent model. CLWETM expects this combination to prevent the final model to be biased to each of the query-dependent/independent collection. This expectation works well in very short queries in which there is a limited information about the intention of the user (e.g., bi-gram queries). But when the original query has an informative knowledge about the intention of the user (i.e., long queries), it is better to consider statistics of the original query as a number of feature alongside the other query-dependent/independent features. For example in Table 5 [*tete*] absorbed all translation weight of 'brain' and then prevented the model to have more coverage/recall. On the other hand, appearing [*cerveau*] as a relevant observation in D, lead L2R to distribute translation probability more justly between [*tete*] and [*cerveau*]. Therefore, we believe that L2R defines a reliable hyperplane discriminating between the context words and the noisy ones more effectively.

4.3 Parameter Sensitivity

$|D|$ is the only parameter in the proposed L2R method. For each collection, we opted $|D|$ that gives the optimum MAP on L2R over a small subset of queries and then tested on remaining topics (Gao et al.,

Table 6: Expected number of query terms ($|q|$) and size of the query-generated training data ($|D|$).

FR				DE				ES				IT			
short		long		short		long		short		long		short		long	
$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$	$\|q\|$	$\|D\|$
2.76	10.62	11.58	53.44	2.8	15.62	11.54	82.7	2.8	11.6	11.56	59.9	2.82	11.14	11.73	60.76

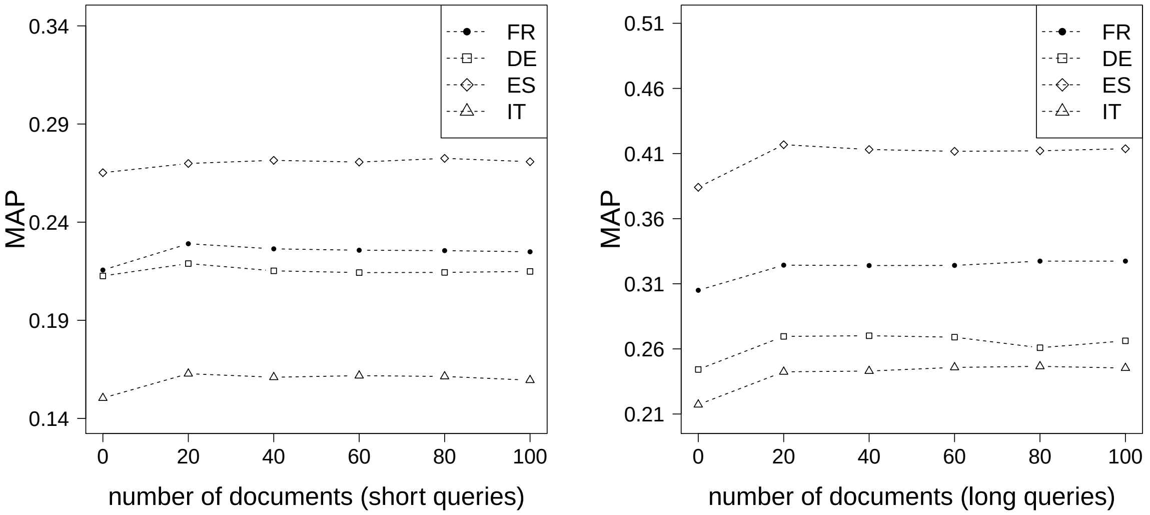

Figure 2: MAP sensitivity of L2R to the number of feedback documents in short and long queries respectively.

2005; Dadashkarimi et al., 2016). As shown in Figure 2, the proposed method works stably in all the collections. In long queries, amount of the improvements are clearly larger than the short ones (see the amounts of jumps from $|D| = 0$ to $|D| = 20$).

5 Conclusion and Future Works

In this paper we proposed a learning to rank method based on ordinal regression on a query-generated training data. We built the query-generated training data of translation words by using their presence/absence in pseudo-relevant documents as labels. This training data consists of embedded features representing each translation word. The result of the regression model was used in the scoring function to weight the translation words.

The method was tested on four different collections in four European languages. The experiments showed that the proposed method outperforms the state-of-the-art dictionary-based CLIR methods, especially in long queries, and it reached up to 81.46% of the performance in the monolingual task. As a future work, the authors would like to test the model on multi-lingual information filtering.

References

Guihong Cao, Stephen Robertson, and Jian-Yun Nie. 2008. Selecting query term alternations for web search by exploiting query contexts. In *ACL-08: HLT*, pages 148–155, Columbus, Ohio, June. Association for Computational Linguistics.

Javid Dadashkarimi, Azadeh Shakery, and Heshaam Faili. 2014. A Probabilistic Translation Method for Dictionary-based Cross-lingual Information Retrieval in Agglutinative Languages. In *CCL '14*, Tehran, Iran.

Javid Dadashkarimi, Mahsa S Shahshahani, Amirhossein Tebbifakhr, Heshaam Faili, and Azadeh Shakery. 2016. Dimension projection among languages based on pseudo-relevant documents for query translation. *arXiv preprint arXiv:1605.07844*.

Hossein Nasr Esfahani, Javid Dadashkarimi, and Azadeh Shakery. 2016. Profile-based translation in multilingual expertise retrieval. In *MultilingMine@ECIR '16*.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *EACL '14*, pages 414–423.

Debasis Ganguly, Johannes Leveling, and Gareth Jones. 2012. Cross-Lingual Topical Relevance Models. In *COLING '12'*, pages 927–942.

Jianfeng Gao, Haoliang Qi, Xinsong Xia, and Jian-Yun Nie. 2005. Linear discriminant model for information retrieval. In *SIGIR '05*, pages 290–297. ACM.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Support vector learning for ordinal regression. In *ICANN '99*, volume 1, pages 97–102. IET.

Thorsten Joachims. 2006. Training linear svms in linear time. In *SIGKDD '06*, pages 217–226. ACM.

Kazuaki Kishida and Noriko Kando. 2005. A hybrid approach to query and document translation using a pivot language for cross-language information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 93–101. Springer.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *SIGIR '01*, pages 120–127.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *SIGIR '02*, pages 175–182.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based Techniques for Cross-language Information Retrieval. *IP&M*, 41(3):523–547.

Yi Liu, Rong Jin, and Joyce Y. Chai. 2005. A maximum coherence model for dictionary-based cross-language information retrieval. In *SIGIR '05*, pages 536–543, Salvador, Brazil.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Yuanhua Lv and ChengXiang Zhai. 2014. Revisiting the divergence minimization feedback model. In *CIKM '14*, pages 1863–1866.

Yanjun Ma, Jian-Yun Nie, Hua Wu, and Haifeng Wang. 2012. Opening machine translation black box for cross-language information retrieval. In *CIRT '12*, pages 467–476.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS '13'*, pages 3111–3119.

Christof Monz and Bonnie J. Dorr. 2005. Iterative Translation Disambiguation for Cross-language Information Retrieval. In *SIGIR '05*, pages 520–527.

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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

Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4(3-4):209–230.

M. F. Porter. 1997. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rzieh Rahimi, Azadeh Shakery, Javid Dadashkarimi, Mozhdeh Ariannezhad, Mostafa Dehghani, and Hossein Nasr Esfahani. 2016. Building a multi-domain comparable corpus using a learning to rank method. *Natural Language Engineering*, 22(04):627–653.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, 25(1):4.

Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR '15*, pages 363–372.

ChengXiang Zhai and John Lafferty. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *CIKM '01*, pages 403–410, Atlanta, Georgia, USA.