Natural Language Engineering **22** (4): 627–653. © Cambridge University Press 2016
doi:10.1017/S1351324916000164

Building a multi-domain comparable corpus using a learning to rank method†

RAZIEH RAHIMI[1], AZADEH SHAKERY[1,2], JAVID DADASHKARIMI[1], MOZHDEH ARIANNEZHAD[1], MOSTAFA DEHGHANI[3] and HOSSEIN NASR ESFAHANI[1]

[1]*School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran*
[2]*School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[3]*Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands*
e-mails: razrahimi@ut.ac.ir, shakery@ut.ac.ir, dadashkarimi@ut.ac.ir, m.ariannezhad@ut.ac.ir, h_nasr@ut.ac.ir, dehghani@uva.nl

(*Received 18 December 2014; revised 1 May 2015; accepted 5 July 2015*)

Abstract

Comparable corpora are key translation resources for both languages and domains with limited linguistic resources. The existing approaches for building comparable corpora are mostly based on ranking candidate documents in the target language for each source document using a cross-lingual retrieval model. These approaches also exploit other evidence of document similarity, such as proper names and publication dates, to build more reliable alignments. However, the importance of each evidence in the scores of candidate target documents is determined heuristically. In this paper, we employ a learning to rank method for ranking candidate target documents with respect to each source document. The ranking model is constructed by defining each evidence for similarity of bilingual documents as a feature whose weight is learned automatically. Learning feature weights can significantly improve the quality of alignments, because the reliability of features depends on the characteristics of both source and target languages of a comparable corpus. We also propose a method to generate appropriate training data for the task of building comparable corpora. We employed the proposed learning-based approach to build a multi-domain English–Persian comparable corpus which covers twelve different domains obtained from Open Directory Project. Experimental results show that the created alignments have high degrees of comparability. Comparison with existing approaches for building comparable corpora shows that our learning-based approach improves both quality and coverage of alignments.

† This research was in part supported by a grant from Institute for Research in Fundamental Sciences (No. CS1393-4-43).

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

1 Introduction

Multilingual nature of the Web makes interlingual translation a crucial requirement for information management applications. Bilingual humanly constructed dictionaries are the basic available resources for translation, but they do not provide sufficient coverage for translation purpose due to out of vocabulary words, phrases and neologisms. This naturally motivates the use of data-driven translation systems. Machine translation systems need bilingual training data, which plays a central role in the accuracy of translations provided by the systems. Thus, providing appropriate training data is an important factor for machine translation systems. Parallel corpora are high quality training data that can be provided for such translation systems. However, building these corpora with suitable coverage of different domains is a costly task.

Comparable corpora are key resources for translation systems concerning languages and domains lacking linguistic resources. Aligned documents in a comparable corpus describe the same or similar topics. Loose alignments of comparable corpora compared to parallel ones can be created at lower costs. This facilitates building domain-specific comparable corpora or expanding translation resources for language pairs with limited linguistic resources. Comparable corpora can be used to directly obtain translation knowledge, or to train machine translation systems by extraction of parallel sentences/phrases (Fung and Cheung 2004; Munteanu and Marcu 2006; Smith *et al.* 2010).

The most common approach for building comparable corpora is based on Cross-Language Information Retrieval (CLIR) (Tao and Zhai 2005; Talvensaari *et al.* 2007, 2008; Huang *et al.* 2010; Paramita *et al.* 2013; Hashemi and Shakery 2014). For each document in one corpus (referred to as the source document), a driver query is generated. Documents of another corpus (referred to as target documents) are then ranked with respect to the driver query using a model for cross-lingual information retrieval. In the next step, source and target documents with a similarity score higher than a predefined threshold constitute an alignment. Finally, some heuristics, such as matched named entities, publication dates and ratio of document lengths, are used to select most reliable alignments. Using different criteria for measuring the similarity between source and target documents improves the accuracy of alignments.

Aligning documents by separately considering CLIR scores, named entities and other possibly available features such as publication dates, does not provide the flexibility and control over the impact of each feature on document similarities. This deficiency is important because the reliability of employed heuristics or features in indicating the degree of document similarity varies depending on language pairs of corpora. For example, features such as named entities, numbers and events are important indicators of comparability for languages sharing many cognates. On the other hand, when language alphabets are totally different, matching named entities and events can exhibit as much noise as matching content words. Therefore, named entity and event features for such languages are not as reliable as they are for languages with large similarity.

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

In this paper, we propose an approach to build high quality comparable corpora. The proposed approach employs a learning-based ranking algorithm to derive an alignment model for building comparable corpora. In order to train the alignment model, we need cross-lingual features and training data. We derive features based on cross-lingual information retrieval, cross-lingual document similarity and named entities. To provide appropriate training data for the learning to rank algorithm, we simulate a comparable corpus from an available parallel corpus where properties of comparable corpora, namely differences in the similarity degrees between alignments and the length of aligned documents, are considered. The proposed method has the following advantages:

(1) Our proposed approach facilitates the integration of any signal on the similarity of a specific language pair into the process of building comparable corpora.
(2) Adopting a learning-to-rank method for ranking target documents with respect to a source document enables learning the weights of various similarity features for a specific language pair.
(3) The incorporated cross-lingual features allows ranking of target documents not only based on matching translated driver queries, but also according to cross-lingual document similarities and named entities.

We exploit the proposed approach for building two comparable corpora. The first comparable corpus is built using the websites of the Open Directory Project (ODP) in different domains. A collection of documents also crawled from websites in Persian as the target collection for this comparable corpus. The built m-domain English–Persian Comparable Corpus (MUDEPCC) covers twelve different domains. MUDEPCC is 1.9 GB on disk and includes 218,140,126 terms. This corpus is publicly available to the research community. MUDEPCC is a valuable resource for training a machine translation system or obtaining translation knowledge for CLIR. Covering different domains facilitates translation using an in-domain bilingual corpus which improves the accuracy of translations. This is because the sense of extracted translations from a bilingual corpus for a word is affected by the domain of the constituting documents of the corpus.

Applying comparability measures, *binary* and *cosine comparability measures* (Saad *et al.* 2013), on MUDEPCC and *TEP* parallel corpus (Pilevar *et al.* 2011) reveals that aligned documents of MUDEPCC have high degrees of comparability. In particular, alignments of MUDEPCC achieve comparability of 49.32 percent and 40.60 percent of that of the parallel corpus based on binary and cosine comparability measures, respectively. Building MUDEPCC shows the effectiveness of our approach on large, noisy datasets of diverse documents in terms of size and writing style (formal/informal language usage).

The second comparable corpus is built using two independent news collections, BBC news[1] in English and Hamshahri news[2] in Persian. These two collections were exploited in the creation of an available comparable corpus (Hashemi and Shakery

[1] www.bbc.co.uk
[2] www.hamshahrionline.ir

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

2014). Building this comparable corpus allows comparison of our approach to the approaches based on CLIR for building comparable corpora. Experimental results show that the alignments of our comparable corpus improve binary and cosine comparability measures by 11.3 percent and 0.2 percent compared to those of the previously created comparable corpus. In addition, using translations extracted from our comparable corpus compared to those from the available comparable corpus, improves the performance of CLIR in terms of Mean Average Precision (MAP) by sixty-eight percent and seventy-five percent for two CLEF query sets (Agirre *et al.* 2009; Ferro and Peters 2009).

The remainder of the paper is organized as follows. We present related work in Section 2. The proposed learning-based approach for building comparable corpora, defined features, generated training data and the overall process of building a multi-domain comparable corpus are presented in Section 3. Section 4 describes the experimental setup, evaluation of the proposed approach for building comparable corpora and the built multi-domain comparable corpus. Finally, we conclude in Section 5.

2 Related work

There are two main approaches for building comparable corpora. The most common approach is based on cross-lingual information retrieval, while in the other approach, source documents are translated using a machine translation system. The translated documents are then compared with documents in the target language to find the most similar documents. Herein, we focus on the first approach since one of the primary purposes of building comparable corpora is to provide translation resources for resource-limited language pairs, and machine translation systems may not be available for these language pairs.

Comparable corpora are usually built from news documents, Wikipedia articles, available monolingual collections, or texts mined from the Web. In the literature on comparable corpora, Sheridan and Ballerini (1996) and Braschler and Schäuble (1998) are among the first works on creating comparable corpora which exploit news collections for this purpose. In (Sheridan and Ballerini 1996), features used to create alignments are publication dates and language-independent descriptors of news articles. Braschler and Schäuble (1998) also consider proper names, numbers and shared words based on a simple dictionary.

Tao and Zhai (2005) propose a more principled method for measuring document similarities. In their method, document similarity is estimated based on the expected correlation between any pair of source and target words. They also consider heuristics like document length normalization and inverse document frequency weighting to penalize long documents and more popular words, respectively. Talvensaari *et al.* (2008) align documents of news collections based on document scores in a retrieval process and their publication dates. In their method, a query is built from each source document based on Relative Average Term Frequency criterion. The queries are translated into the target language, which are then used to retrieve relevant documents from the target collection. Target documents whose

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

relevance scores are higher than a threshold are aligned with the source document. Threshold values for aligning documents depend on proximity of documents' publication dates. This approach is further exploited for building a comparable corpus in a specific domain from the Web with the difference that the feature of publication date is not available for web pages (Talvensaari *et al.* 2007). Huang *et al.* (2010) also employ CLIR to build a comparable corpus.

In order to reduce the cost of building comparable corpora, Aker, Kanoulas and Gaizauskas (2012) suggest to obtain only the title and some meta-data information, such as publication date and time, for each document instead of its full content. Cosine similarity between the frequency vectors of titles' terms is then used to match titles, and the actual contents of paired titles are only crawled.

Strötgen, Gertz and Junghans (2011) introduce a document similarity measure based on events. To compute this measure, a document is modeled as a set of events which are temporal and geographic expressions in the document. Target documents are then ranked based on temporal and geographic hierarchies. Paramita *et al.* (2013) also propose an automatic technique to acquire a comparable corpus from the Web using news web pages, Wikipedia and Twitter. They employed extracted entities, time interval filtering, URL of web pages and document lengths as features for gathering comparable data. The scores obtained by CLIR are also used in the alignment process.

In the current study, we aim to build high quality comparable corpora from available monolingual corpora in different languages. The proposed approach is based on adopting a learning-based method to dynamically determine the weights of diverse similarity features. Finally, built comparable corpora can be used to directly obtain translation knowledge, or to train machine translation systems by extraction of parallel sentences/phrases. Different approaches are proposed to extract parallel data from noisy parallel or comparable corpora. Fung and Cheung (2004) represent each sentence as a vector of words. The cosine similarity of vectors of source and target sentences are calculated to determine parallel sentences. Munteanu and Marcu (2005) use a binary classifier to determine parallel sentences in a comparable corpus. To avoid the class imbalance problem in detecting parallel sentences, a ranking model is learned for this purpose in (Smith *et al.* 2010; Pal, Pakray and Naskar 2014) exploit a textual entailment technique and a baseline MT system. Moreover, parallel sentences can be extracted using tools such as Yalign tool[3] or ACCURAT system described in (Skadia *et al.* 2012).[4] All these approaches can be adopted on comparable corpora built using our approach in order to obtain translations.

3 A learning-based method to build comparable corpora

In this section, we describe our approach for building a comparable corpus from the Web. A comparable corpus consists of a set of alignments between documents in two languages. Two aligned documents in a comparable corpus cover the same or

[3] https://github.com/machinalis/yalign
[4] http://www.accurat-project.eu/

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

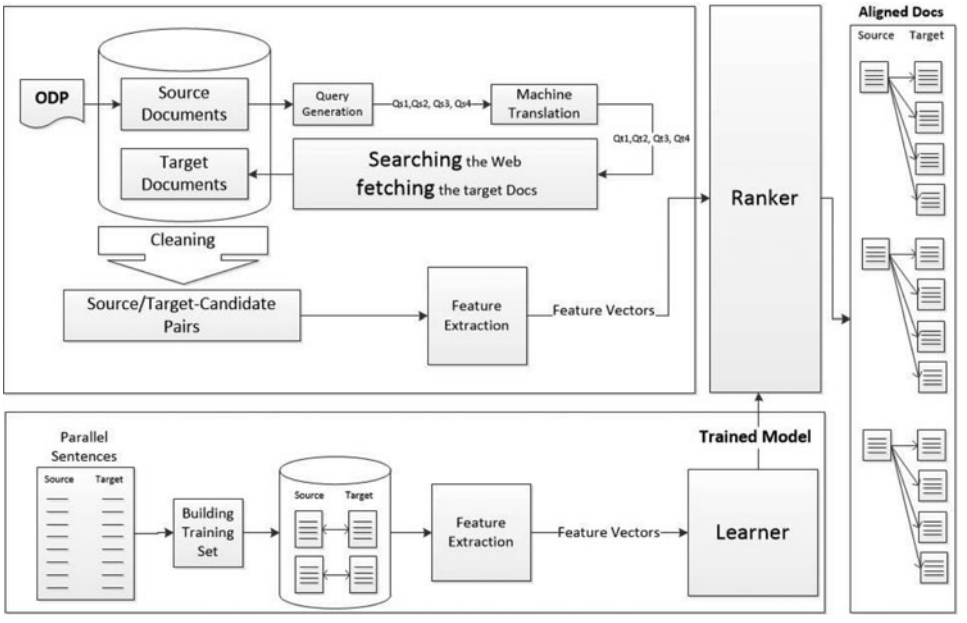


Fig. 1. The overall process of building a multi-domain comparable corpus.

similar topics, but they are not necessarily literal translations of each other. We can formally define a comparable corpus as an alignment vector (Rahimi and Shakery 2013):

$$A = \{(d_{s_1}, d_{t_1}, s_1), \ldots, (d_{s_n}, d_{t_n}, s_n)\}, \tag{1}$$

where each element is a triple consisting of *source document ID* (d_{s_i}), *target document ID* (d_{t_i}) and *similarity score* (s_i). Therefore, building a comparable corpus requires gathering documents in the source and the target languages, and then aligning the most similar documents. Figure 1 shows the overall process of building our MUDEPCC. There are two general approaches for gathering comparable collections of documents in two languages. The first approach is to build document alignments using two existing monolingual collections such as articles of two news agencies, while the other approach crawls target documents based on available source documents. The latter is more appropriate for the translation purpose since it allows building comparable corpora in a specific arbitrary domain or building large corpora with suitable coverage of different domains. Our method for building a comparable corpus exemplifies the second approach. Following, we first present the process of data gathering, and then introduce the proposed method to align documents.

3.1 Acquiring the documents

Data collection is the first step in building comparable corpora (the upper left component in Fig. 1). For the source documents, we use the ODP (data of ODP

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

2012)[5] which contains URLs of 5,169,997 websites in over 1,017,500 categories. The ODP organizes URLs in a hierarchical structure of domains where URLs of each domain are further categorized into sub-domains. Using the ODP to obtain source documents guarantees that initial seeds will cover different domains. This point is essential in building comparable corpora since increasing the size of a comparable corpus in a specific domain does not necessarily improve the quality of obtainable translations (Xu and Weischedel 2005; McNamee, Mayfield and Nicholas 2009). The contents of the websites in ODP constitute our collection of source documents.

In the next step, candidate websites in the target language are gathered for each source document. This step is performed by generating a number of driver queries for each source document and then finding documents relevant to the translations of the driver queries. To ensure obtaining proper candidate target documents for each source document (in English), we use different available information for web pages in ODP and build four driver queries for each source document. More specifically, for each source document, four driver queries are built as follows:

(1) **Keywords**: The first query is composed of keywords that are extracted from the source documents. Keywords of a document are extracted by analogy with *mixture-model feedback* method (Zhai and Lafferty 2001) in the KL-divergence retrieval framework. We estimate a language model for each document by modeling noise words in the document with the collection language model, in a way that the probability of each word given the document is a linear combination of its new language model and the collection language model as follows:

$$p(w|D) = \lambda p(w|\theta'_D) + (1 - \lambda)p(w|C), \tag{2}$$

where λ is a parameter that controls the amount of noise, $p(.|C)$ is the collection language model, and θ'_D is the new language model for document D which captures its topic. We then rank the words of document D by their probabilities in θ'_D. The five most probable words form the first query.

(2) **Title of websites**: The second query is built from terms of the websites' title after removing the stopwords.

(3) **Description of websites**: The third query is extracted from the description provided for each page in ODP. The description of each page is part of the page's metadata information and gives an objective statement of the content of the page. Queries from provided descriptions are derived based on the discrimination values of terms of description texts. Each term's discrimination value is calculated similar to IDF as $\frac{1}{df(t)}$, where $df(t)$ represents the document frequency of term t. We estimate the document frequency of a term based on its hit count in Web search results in response to a query consisting of that term using Bing search API.[6] After sorting terms of the description part based on this criterion, five top-scored terms are selected as query terms.

[5] http://www.dmoz.org/

[6] Bing Search API. http://datamarket.azure.com/dataset/bing/search

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

(4) **Domain hierarchies**: The fourth query is built using the hierarchy structure of domains in ODP where URLs are organized. Each URL is a leaf in the domain tree of ODP. For each URL, the domain labels of its parent, grandparent and great-grandparent constitute a query for the document constructed from that URL.

In the next step, we translate each derived query into the target language (Persian) using Google machine translation system. The translated queries are then searched on the Web using the Google Web Search API[7] and top eight pages are fetched for each query. After this step, we have at most thirty-two candidate documents for each source document.

3.2 Exploiting a learning-to-rank method to build alignments

Creation of alignments between source and target documents can be viewed as a task of ranking candidate target documents for each source document. Learning to rank methods are extensively used for constructing ranking models from labeled training data with promising performance. Therefore, we employ a learning to rank method for ranking target documents. In this regard, Ranking SVM (Joachims 2002) is used as the learning method in our approach since it has been shown to be effective for the ranking purpose in different tasks (Li and Hirst 2011). Ranking SVM is a pairwise ranking method that uses Support Vector Machines (Cortes and Vapnik 1995), one of the state-of-the-art classifiers in several IR and NLP applications, to learn a ranking function. The lower left component in Fig. 1 shows the learning step in the proposed approach to build comparable corpora. Following, we first introduce the employed features to build the ranking model and then describe the process of building appropriate training data for the task of creating a comparable corpus.

3.2.1 Features

In this section, we define the features exploited in the learning to rank method, which are listed in Table 1. We divide the features into four groups, frequency-based features, discrimination-based features, features based on cross-lingual document similarity, and named entity feature. Before proceeding to formal definition of features, let us define some notations. Let w_s and w_t denote source and target words, respectively. Similarly, d_s and d_t denote documents in the source and target languages, respectively.

Frequency-based features. Features in this category capture the frequencies of words of a source document in a target document. The first three features of Table 1, defined in (Azarbonyad *et al.* 2012), belong to this category. Frequency of a word belonging to one language in a document written in another language, referred to as cross-lingual term frequency, is defined as follows (Darwish and Oard

[7] https://developers.google.com/web-search/

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 1. *Features for learning a model using ranking SVM*

1.	$\sum_{w_s \in d_s} \log(\text{CLTF}(w_s, d_t))$
2.	$\sum_{w_s \in d_s} \log(1 + \frac{\text{CLTF}(w_s, d_t)}{\lvert d_t \rvert})$
3.	$\sum_{w_s \in d_s} \log(1 + \frac{\text{CLTF}(w_s, d_t)}{\lvert d_t \rvert} \times \frac{\lvert C_t \rvert}{\text{CLTF}(w_s, C_t)})$
4.	$\sum_{w_s \in d_s} \log(\text{CLIDF}(w_s))$
5.	$\sum_{w_s \in d_s} \text{CLIDF}(w_s) \times \text{BM25TF}(w_s, d_t)$
6.	$\sum_{w_s \in d_s} \log(1 + \frac{\text{CLTF}(w_s, d_t)}{\lvert d_t \rvert} \times \text{CLIDF}(w_s))$
7.	$\sum_{w_s \in d_s} \log(1 + \text{CLTF}(w_s, d_t) \times \text{CLIDF}(w_s))$
8.	$\text{NE}(d_s, d_t)$

2003):

$$\text{CLTF}(w_s, d_t) = \sum_{w_t \in V_t} p(w_t|w_s) \times \text{TF}(w_t, d_t), \tag{3}$$

where $\text{TF}(w_t, d_t)$ shows the frequency of term w_t in document d_t, $p(w_t|w_s)$ determines the probability that w_s translates into w_t and V_t is the set of the vocabulary of the target language. The probabilities $p(w_t|w_s)$ are estimated using an initial translation resource for the language pair of study.

Frequency-based features, defined using cross-lingual term frequency, reflect different roles of term-frequency heuristics in ranking documents. The first feature is based on the fact that the larger frequency of words of a source document in a target document indicates a higher similarity between two documents. However, the increase in the similarity score of two documents due to an increase in the frequency of words that are translations of each other is smaller for larger values of the frequency (see TFC2 in (Fang *et al.* 2011)). The second feature incorporates lengths of documents in order to penalize long documents. The third feature uses the inverse of collection frequency of terms to penalize matching of frequent but non-content terms in deriving document similarities.

Discrimination-based feature. Feature 4 captures the *inverse document frequency* heuristic of information retrieval in the computation of document similarities. This is important because matching more specific terms between a source and a target document is a better indicator of comparability. For this purpose, cross-lingual inverse document frequency is used which measures the popularity of a word in a collection of documents in another language of that of the word, and is defined as

$$\text{CLIDF}(w_s) = \sum_{w_t \in V_t} p(w_t|w_s) \times IDF(w_t). \tag{4}$$

Features based on cross-lingual document similarity. Features 5, 6 and 7 of Table 1 measure cross-lingual document similarity between documents in two languages using cross-lingual term frequency and cross-lingual inverse document frequency

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

defined above. In the fifth feature, $\mathrm{BM25TF}(w_s, d_t)$ is calculated as

$$\mathrm{BM25TF}(w_s, d_t) = \frac{(k+1) \times \mathrm{CLTF}(w_s, d_t)}{k((1-b) + b(\frac{|d_t|}{|d|_{avg}})) + \mathrm{CLTF}(w_s, d_t)}. \quad (5)$$

The three features of this category adopt different normalization methods on CLTF.

Named entity feature. This feature for a pair of source and target documents shows the percentage of named entities in the source document whose translations/transliterations exist in the target document. The NE feature for documents d_s and d_t is calculated as follows:

$$\mathrm{NE}(d_s, d_t) = \frac{\sum_{n \in \mathrm{NE}_s} \mathrm{trans}(n, d_t)}{|\mathrm{NE}_s|}, \quad (6)$$

where NE_s is the set of named entities extracted from document d_s, $\mathrm{trans}(n, d_t)$ equals 1 if the translation/transliteration of named entity n occurs in document d_t, otherwise it is 0, and $|\mathrm{NE}_s|$ is the number of all detected named entities in the source document d_s.

3.2.2 Training data

Since Ranking SVM is a supervised learning method, we need training data which should be of high quality to result in high performance ranking. In our problem, the goal is to find the most similar documents to a source document in order to build alignments. Thus, we need a set of high-quality aligned documents as training data to construct a model using the Ranking SVM method. A perfect alignment includes documents that are literal translations of each other. These document pairs better reveal the importance degree of each feature in determination of comparable documents. Therefore, we employ a parallel corpus to generate training data. Most of available parallel corpora are sentence aligned, in contrast to document level alignments of comparable corpora. Thus, the fine-grained level of alignments of a parallel corpus are not suitable for our task. To resolve this issue, we construct source and target documents by concatenating a number of source sentences and their aligned target sentences in the parallel corpus, respectively.

The obvious way for building artificial documents is to select a constant number of sentences as introduced in (Ture, Elsayed and Lin 2011). However, selecting a constant number of sentences would not yield a perfect training set for our task. This is because all the built documents in the source or the target language have approximately equal lengths. This training set is not compatible with actual contents of web pages that we aim to align. Documents extracted from web pages can have different sizes. Thus, in measuring the similarity of these documents, longer documents need to be penalized more since they have more chance of being aligned with other documents. To consider this point in the learning process, we built artificial documents with different lengths from the sentence-aligned parallel corpus. In the first step, we estimate the probability mass function of the length of documents acquired for the source part of our comparable corpus from ODP. Then, a random number is repeatedly generated from a uniform distribution in the range $[0, 1]$ and mapped to the estimated probability mass function. The corresponding

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

value of document length in the probability mass function indicates how many parallel sentences should be concatenated to create a new document. For each built document in the source language, a corresponding document in the target language is also created by concatenating the sentences that are aligned with the sentences of the source document in the parallel corpus. This parallel document provides a sample of similar document for the source document and is labeled as '1'. Supervised learning to rank methods should be trained on examples of similar and dissimilar document pairs. In this regard, nine target documents are created by concatenating randomly selected target sentences. These non-parallel documents are labeled as '0'. Therefore, this training data contains only parallel and non-parallel document pairs. We refer to this training data as 'Training Data with One Similarity Degree'. However, considering only parallel documents as similar documents does not totally capture the characteristics of comparable corpora. For example, only parts of aligned documents in a comparable corpus may be on the same topic. To provide more samples of similar documents, another training data with three degrees of similarity is built.

In the second training data, there are three categories of similar documents for each built document in the source language:

(1) **A parallel document**: This document contains the literal translations of the sentences of the source document as described above.
(2) **A semi-parallel document**: This document contains the whole literal translation of the sources document along with a randomly generated part where its length is determined by a random number drawn from uniform distribution in the interval $U[0, 3 * |D_s|]$ ($|D_s|$ shows the length of the source document).
(3) **A comparable document**: The first part of this document contains the literal translation of the source document where its length is obtained randomly from a normal distribution with mean $0.5 * |D_s|$ and standard deviation $0.3 * |D_s|$. The second part of this target document is randomly generated and its length is determined by a random number drawn from uniform distribution in the interval $U[0, |D_s|]$.

All three built documents in the target language are considered as similar documents to the source document, and are labeled as '3', '2', and '1' in order of decreasing similarity degree. Therefore, in this training data, there are twelve alignments for each source document where three of them contain similar documents with different similarity degrees and the other nine alignments are random dissimilar documents. We refer to this training data as 'Training Data with Three Similarity Degrees'. Finally, both training data are used to learn a ranking model, and ranks of candidate target documents provided by the learned ranking model are used to align documents.

4 Experiments

In this section, we present our experimental results on the evaluation of the proposed approach to build comparable corpora and the quality of the two built comparable

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

corpora. We first introduce the measures for evaluating the alignment qualities of a comparable corpus. Next, the exploited datasets and linguistic resources, preprocessing steps and experimental setups are described. Finally, we report the characteristics and quality of the two built comparable corpora and compare our approach with the previous approaches.

4.1 Evaluation measures

We employ two types of evaluation metrics. The first type of metrics measure the comparability degree of aligned documents. The second type of measures evaluates the quality of alignments by assessing translation knowledge that can be extracted from comparable corpora.

4.1.1 Comparability measures

The comparability degree of alignments of a comparable corpus is measured using 'binary Comparability Measure' (binCM) and 'cosine Comparability Measure-Vector Space Model' (cosine-CM-VSM) metrics introduced in (Saad *et al.* 2013). The binCM measure is computed based on the number of words of a source document whose translations exist in its aligned document, and is defined as follows:

$$\text{binCM}(d_s, d_t) = \frac{\sum_{w_s \in d_s \cap V_s^{\text{dic}}} \text{trans}(w_s, d_t)}{|d_s \cap V_s^{\text{dic}}|}, \tag{7}$$

where V_s^{dic} is the source vocabulary of a bilingual dictionary, and d_s and d_t are the source and target documents, respectively. The binary function $\text{trans}(w_s, d_t)$ equals 1 if the translation of w_s exists in d_t. This metric has a value in $[0, 1]$ in that the higher the value, the higher the degree of comparability between aligned documents. Based on this metric, the comparability degree of an entire comparable corpus A is defined as:

$$\text{binCM}(A) = \sum_{a_i \in A} \frac{\text{binCM}(d_{s_i}, d_{t_i}) + \text{binCM}(d_{t_i}, d_{s_i})}{2|A|}, \tag{8}$$

where A is the vector representation of the comparable corpus, and d_{s_i} and d_{t_i} are the source and target documents of alignment a_i of the corpus, respectively.

The second comparability measure is cosineCM-VSM. To calculate this metric, source and target documents of an alignment are represented by vectors of size m, the number of entries in a bilingual dictionary. Let $w_s \rightarrow w_t$ denote an entry of the dictionary. The value for this entry in the vectors of the source and target documents (d_s and d_t) are $\text{tf}(w_s, d_s)$ and $\text{tf}(w_t, d_t)$, respectively, where $\text{tf}(w, d)$ shows the frequency of word w in document d. The cosineCM-VSM measure for an alignment is then estimated by computing the cosine similarity of the vectors of its source and target documents as follows:

$$\text{cosineCM} - \text{VSM}(V_{d_s}, V_{d_t}) = \frac{\sum_{i=1}^{m} V_{d_s}.V_{d_t}}{\|V_{d_s}\| \times \|V_{d_t}\|}, \tag{9}$$

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

where V_{d_s}/V_{d_t} is the vector of the source/target document of an alignment. Comparability degree of a comparable corpus using this measure is the average of cosineCM-VSM values across all alignments.

4.1.2 Evaluation based on extraction of translations

In this evaluation approach, the goal is to measure the quality of extracted translation models from comparable corpora. The higher the quality of alignments of a comparable corpus, the higher the quality of obtainable translations. There are different methods to extract translation models from comparable corpora (Gaussier *et al.* 2004; Talvensaari *et al.* 2007; AbduI-Rauf and Schwenk 2009; Garera, Callison-Burch and Yarowsky 2009; Vulić and Moens 2012). We focus here on approaches that extract translations from comparable corpora without using any other translation resources such as machine-readable dictionaries (Sheridan and Ballerini 1996; Tao and Zhai 2005; Talvensaari *et al.* 2007; Rahimi and Shakery 2013). We adopt the language modeling approach proposed in (Rahimi and Shakery 2013; Rahimi, Shakery and King 2016). In this approach, each word is modeled by a probability distribution based on the alignment representation as in Equation (1). Let w_s/w_t denote a source/target word. The parameters of generated models for w_s and w_t are $p(d_{s_i}|w_s)_{i=1}^{n}$ and $p(d_{t_i}|w_t)_{i=1}^{n}$, respectively, where n is the size of alignment vector in Equation (1). Similarity of two words is then estimated using cross entropy of their models as follows:

$$\text{sim}(w_s, w_t) \overset{\text{rank}}{=} \exp\left(\beta \sum_{i=1}^{|A|} p(d_{s_i}|w_s) \log p(d_{t_i}|w_t)\right), \tag{10}$$

where β is a parameter that controls the effect of noise words. These similarity scores are then normalized in order to compute word translation probabilities as follows:

$$p_{\text{cc}}(w_{t_i}|w_s) = \frac{\text{sim}(w_s, w_{t_i})}{\sum_{j=1}^{|V_t|} \text{sim}(w_s, w_{t_j})}, \tag{11}$$

where V_t denotes the vocabulary set of the target language. The accuracy of translation models estimated by Equation (11), can be evaluated using a machine-readable dictionary. However, due to the OOV problem, CLIR is performed to provide a more complete evaluation (Tao and Zhai 2005; Talvensaari *et al.* 2008; Rahimi and Shakery 2013).

For CLIR, we opt the query translation approach based on the language modeling framework (Nie 2010). In this approach, a new language model is estimated for the query, and documents are ranked using

$$\text{score}(Q, D) = \sum_{w_t \in V_t} \Big(\sum_{w_s \in V_s} p_{\text{cc}}(w_t|w_s) p(w_s|\theta_Q) \Big) \log p(w_t|\theta_D), \tag{12}$$

where V_s and V_t are source and target vocabulary sets, respectively. θ_Q and θ_D are, respectively, query and document language models. $p_{\text{cc}}(w_t|w_s)$ is the probability of translating word w_s into word w_t which is estimated using Equation (11) on the built comparable corpus.

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 2. *Statistics of training data with one similarity degree*

Average length of English documents	103
Average length of Persian documents	180
Number of English unique terms	1,034,79
Number of Persian unique terms	3,11,353

4.2 Datasets and experimental setup

Training data. The required training data is constructed using TEP parallel corpus (Pilevar *et al*. 2011). TEP parallel corpus is a sentence-aligned English–Persian corpus built from movie subtitles. The two training datasets consist of 56,846 parallel document pairs by concatenating the parallel sentences of TEP. In addition, training datasets contain 5,011,614 non-parallel document pairs where their target documents are created by concatenation of random Persian sentences. Table 2 shows the statistics of generated training data with one similarity degree.

Feature extraction. The translation model, used for feature extraction, is trained using the GIZA++ toolkit (Och and Ney 2003) on TEP parallel corpus. The obtained model is of the form $p(w|u)$ which indicates the probability of translating word u in one language to word w in another language. In order to extract the named-entity feature in our experiments, named entities in the source documents (in English) are first identified using the Stanford Named Entity Recognizer (Finkel, Grenager and Manning 2005). The identified named entities are then translated using the Google machine translation system and searched in the target documents.

Dataset. To evaluate the quality of built comparable corpora based on CLIR (Equation 12), a test collection is needed for which we use Hamshahri collection containing 166,774 documents in Persian with two sets of queries (topics in CLEF 2008 and CLEF 2009) (AleAhmad *et al*. 2009) in English and Persian languages. We use the title of Ham'08 and Ham'09 topics for the evaluation. Stopwords are removed from queries and documents of Hamshahri collection.

Tools. Experiments are done using the Lemur toolkit.[8] For training the ranking model, the SVM^{rank} tool[9] is used.

Data preprocessing. After crawling the URLs of the ODP, preprocessing steps are performed to gather only informative data from the websites. To get the main content of each web page, the jusText tool[10] (Pomikálek 2011) is used for document cleaning which eliminates noise of the web pages, such as navigation bars and advertisements. Furthermore, the target documents are refined by removing non-Persian documents. The words of remained Persian documents are then normalized by replacing all orthographic variations of letters by one form.[11]

[8] http://www.lemurproject.org/
[9] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html
[10] http://corpus.tools/wiki/Justext
[11] https://github.com/samiraabnar/PersianNormalizer

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 3. *Domains of MUDEPCC comparable corpus*

D_1	ART	D_4	GAME	D_7	NEWS	D_{10}	SCIENCE
D_2	BUSINESS	D_5	HEALTH	D_8	RECREATION	D_{11}	SOCIETY
D_3	COMPUTER	D_6	HOME	D_9	REFERENCE	D_{12}	SPORT

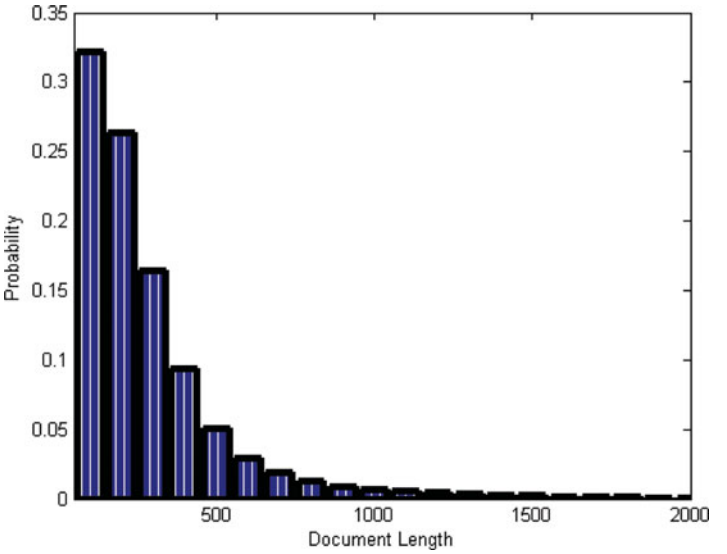


Fig. 2. (Colour online) Distribution of the length of English documents obtained by crawling the websites of ODP.

4.3 Characteristics of crawled data

In this section, we present the characteristics of the source and target collections that are used to build our multi-domain comparable corpus. The source collection is obtained by crawling the URLs provided in the ODP. The documents of this source collection cover twelve different domains that are listed in Table 3. We choose domains at the top level of the ODP hierarchy that have a large number of URLs and also few intersections with each other.

The target collection is obtained by generating driver queries and crawling documents retrieved by the Google search engine with respect to the driver queries. Statistics of each domain in the source and target collections are reported in Table 4, which include the number of documents and Average Document Lengths. In addition, Fig. 2 shows the distribution of the length of English document, which is used to generate training data.

4.4 Assessing the quality of MUDEPCC

We first construct a ranking model using 'Training Data with One-Degree Similarity'. This model is used to rank candidate Persian documents for each English document. Each English document is then aligned with its most similar document in Persian.

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 4. *Statistics of each domain (denoted by D_i) in the obtained source and target collections before the alignment of documents*

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
#EN-Docs	177,869	141,148	69,036	4,600	544	13,536	4,658	60,723	33,304	71,836	138,970	56,752
#FA-Docs	122,942	192,501	66,349	6,304	5,451	6,822	1,151	59,682	53,211	41,628	168,866	26,438
ADL-EN	386	213	409	579	646	419	500	297	336	471	455	375
ADL-FA	2,452	1,706	1,173	1,905	5,013	1,733	1,301	2,050	1,768	1,523	2,025	1,845

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 5. *MUDEPCC statistics. #UT-EN and #UT-FA show the numbers of unique terms in the English and Persian sides of the corpus, respectively*

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
#Aligns	11,883	14,635	9,060	1,259	466	1,523	250	11,035	13,774	6,907	24,686	6,418
ADL-EN	387	266	365	383	570	410	559	319	342	398	436	395
ADL-FA	2,294	1,647	1,254	1,730	5,186	1,651	1,198	1,913	1,467	1,675	1,896	1,670
#UT-EN	128,726	98,328	131,612	35,615	24,007	32,496	20,591	160,201	174,531	148,202	216,309	108,410
#UT-FA	1,039,447	939,997	484,570	122,198	168,603	158,700	42,035	784,765	708,136	528,820	1,419,853	546,245
Size(MB)	278	243	124	22.8	21.5	25.5	3.7	201	206	120	509	106

Table 6. *Comparability degrees of comparable corpora built using our approach (MUDEPCC and MUCC3), and a document-level parallel corpus*

Corpus	cosine-CM-VSM	$\text{binCM}_{(p\geq 0.1)}$	$\text{binCM}_{\text{top}}$
TEP-Parallel-Docs	0.1212	0.2733	0.2294
MUDEPCC	0.0535	0.1410	0.1087
MUCC3	0.0519	0.1290	0.0995

The obtained comparable corpus is referred to as MUDEPCC and its statistics are reported in Table 5.

We assess the quality of our multi-domain comparable corpus by computing the comparability measures introduced in Section 4.1.1. Calculation of these comparability measures needs a reference bilingual dictionary. Employing a machine-readable bilingual dictionary for this purpose requires a reliable morphological analyzer to derive the base form of words. However, due to the lack of such analyzer for Persian language, which yields missing some important inflectional formations (Dadashkarimi, Shakery and Heshaam 2014), we employ a dictionary which is extracted from a parallel corpus.

We use the word-to-word translation model trained on TEP parallel corpus as the translation reference. The obtained translation model is exploited in two different ways for computing the binCM measure. First, only the top translation of each source word is considered. We denote the comparability degrees computed using this dictionary as $\text{binCM}_{\text{top}}$. The cosine-CM-VSM measure is only computed based on the mentioned dictionary. The second reference dictionary is built by considering translations with probabilities greater than 0.1 per source word. To use this dictionary in computation of the binCM measure, binary function $\text{trans}(w_s, d_t)$ in Equation (7) equals 1 if at least one of the translation alternatives of w_s exists in d_t. We denote the comparability degrees computed using this probabilistic dictionary as $\text{binCM}_{(p\geq 0.1)}$. Table 6 shows the computed comparability degrees for MUDEPCC and the parallel part of our training data referred to as TEP-Parallel-Docs. Alignments of our comparable corpus, MUDEPCC, achieve comparability of 49.32 percent and 40.60 percent of that of TEP-Parallel-Docs in terms of binCM and cosine-CM-VSM, respectively.

We also built a comparable corpus by using 'Training Data with Three Similarity Degrees', which is referred to as 'MUCC3'. Table 6 also includes the comparability degrees of the alignments of this comparable corpus. The results show that the alignments of MUDEPCC have higher degrees of comparability compared to MUCC3, but the differences are not significant.

Moreover, we evaluate how well our trained ranking model performs in selection of the best target document (in Persian) for a source document (in English) by using the comparability measures. In particular, we aim to investigate whether our ranking model has selected the target document with the highest comparability degree to a source document for the alignment purpose. In this regard, we first compute the comparability measure binCM or cosine-CM-VSM between each English document

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

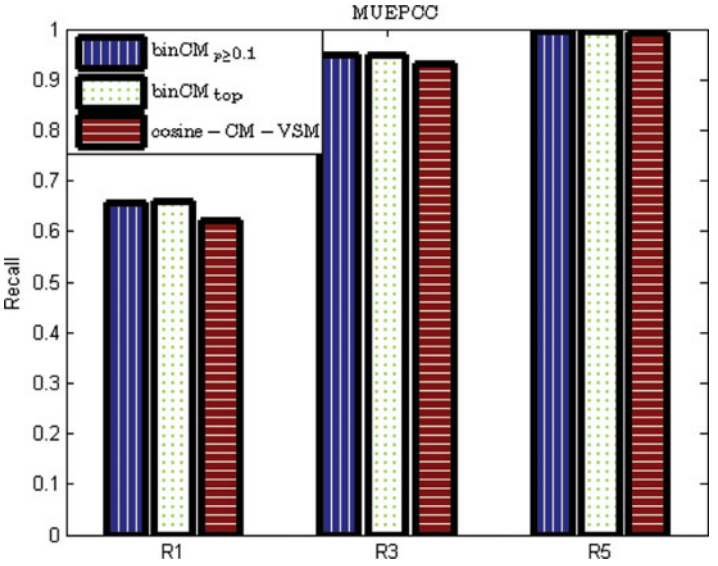


Fig. 3. (Colour online) Evaluation of MUDEPCC using binary and cosine comparability measures.

and all of its candidate documents in Persian. The candidate Persian document with the highest binCM or cosine-CM-VSM score is judged as relevant for the English document. Then, the ranked lists for each English document, provided by our learned ranking model, is evaluated against its judged Persian document according to recall at top 1, 3 and 5 documents (Saad *et al.* 2013). Figure 3 shows these recall values computed based on both comparability measures on MUDEPCC. The figure clearly indicates that our approach for ranking candidate target documents has a high performance.

In order to provide error analysis, we adapted the five-level relevance scale which was first introduced by Braschler and Schäuble (1998) and later used in other studies, such as (Talvensaari *et al.* 2007), (Hashemi, Shakery and Faili 2010) and (Rahimi and Shakery 2011). Five levels of relevance are defined as:

(1) Same story. Documents cover the same topic.
(2) Related story. Documents are about the same topic but are written from slightly different viewpoints, or one of the documents also cover other topics besides the common topic.
(3) Shared aspect. Documents are about the related events, such as events about same persons.
(4) Common terminology. Documents share a considerable amount of shared terminologies.
(5) Unrelated. Documents have no apparent relation.

In this regard, we randomly select some alignments of MUDEPCC and judge their similarities according to this five-level relevance status. Following, we mention two reasons for creation of alignments in MUDEPCC that are judged as level 4 or 5. The first reason is that there is no or very limited Persian web pages related

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

to some English web pages. For example, there is almost no web page in Persian that describes Queen Mary, University of London. Therefore, the web page of the School of History at this university[12] is aligned with the homepage of a person that has been graduated from there, which we considered to be not very related. Second, some web pages have keywords that are ambiguous such that they are specific words in one sense, while they are general words in the other sense. In this case, translation of keywords according to their general senses causes imperfect alignments. This is the case that occurred for a web page about the VW Beetle car.[13] Beetle is translated as an insect[14] in Persian. Therefore, the aligned documents are unrelated.

4.5 Comparison with previous approaches for building comparable corpora

We compare our proposed approach for building comparable corpora with the building approach exploited in (Talvensaari *et al.* 2007), (Hashemi *et al.* 2010) and (Hashemi and Shakery 2014). To provide an exact comparison, we employ our approach on two monolingual collections that are used to build UTPECC comparable corpus in (Hashemi and Shakery 2014). These source and target collections contain 53,697 English documents and 191,440 Persian documents which are gathered from BBC news[15] and Hamshahri newspapers,[16] respectively. Documents of these collections are aligned in a four-step method in (Hashemi and Shakery 2014):

(1) Extraction of keywords from the documents in the source language (English).
(2) Translation of the keywords into the target language (Persian).
(3) Ranking documents of the target collection w.r.t. the query including the translated keywords. The top r target documents are considered as candidate documents for the source document.
(4) Aligning a source and a target document based on the similarity score of the target document to the corresponding query of the source document and the proximity of their publication dates.

The details of step 4 are as follows. For each English document D, the r most similar Persian documents are searched to create alignments. First, Persian documents in r which are published in the same day as D and their retrieval scores are greater than θ_1, are aligned with D. If no candidate Persian document satisfies these conditions, then Persian documents in r which are published in $+/-$ 4 days of D are considered. However, the score threshold increases to θ_2 and Persian documents with retrieval scores greater than θ_2 are selected to be aligned. Again, in case of no alignment, the score threshold increases to θ_3 and a difference of 14 days in publication dates becomes allowed. Following (Hashemi and Shakery 2014), the score thresholds are set as $\theta_1 = 60$, $\theta_2 = 80$ and $\theta_3 = 95$. The values of

[12] http://www.history.qmul.ac.uk/
[13] http://www.maggiolinoweb.it/old_beetle.html
[14] http://en.wikipedia.org/wiki/Beetle
[15] www.bbc.co.uk
[16] www.hamshahrionline.ir

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 7. *Statistics of UTPECC and our comparable corpus (L2RCC)*

	UTPECC	L2RCC
# of alignments	7580	9037
Average length of English documents	263	257
Average length of Persian documents	396	347
# of unique English terms	33941	44564
# of unique Persian terms	47938	50669
Size (MB)	26.3	36.4

Table 8. *Evaluation of UTPECC and our comparable corpus (L2R comparable corpus) using binary and cosine comparability measures*

Corpora	cosine-CM-VSM	%Impr	$\text{binCM}_{(p\geq 0.1)}$	%Impr	$\text{binCM}_{\text{top}}$	%Impr
UTPECC	0.0975		0.2021		0.1687	
L2RCC	0.0977	+0.2%	0.2234	+10.6%	0.1878	+11.3%

score thresholds are percentiles, e.g. $\theta = 60$ means that the retrieval score is greater than sixty percent of retrieval scores of all candidate target documents w.r.t. their corresponding source documents.

We also use this approach to align the documents, but the following parts are implemented differently:

(1) **Keyword extraction**: Keywords are extracted based on the language modeling framework, Equation (2), in our approach, while 'relative average term frequency' is used in (Hashemi and Shakery 2014).
(2) **Selection of candidate documents**: The KL-divergence retrieval model is used with pseudo relevance feedback to select r candidate documents per source document in (Hashemi and Shakery 2014), whereas we use the same model without pseudo relevance feedback.
(3) **Number of candidate documents** : Value of r in our approach is 15, while r is set to 50 in (Hashemi and Shakery 2014).
(4) **Ranking of candidate documents**: The candidate documents are ranked based on their retrieval scores with respect to generated queries from source documents in (Hashemi and Shakery 2014), whereas we use a learning to rank method, as described in Section 3.2, for ranking candidate documents.

Statistics of the exploited news collections, UTPECC comparable corpus which is built in (Hashemi and Shakery 2014), and the built comparable corpus using the proposed method (referred to as L2R Comparable Corpus) on the news collections, are reported in Table 7. The statistics show that more documents of the news collections are aligned using the proposed learning-based method compared to the previous method. Following, we evaluate the quality of alignments in two comparable corpora.

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

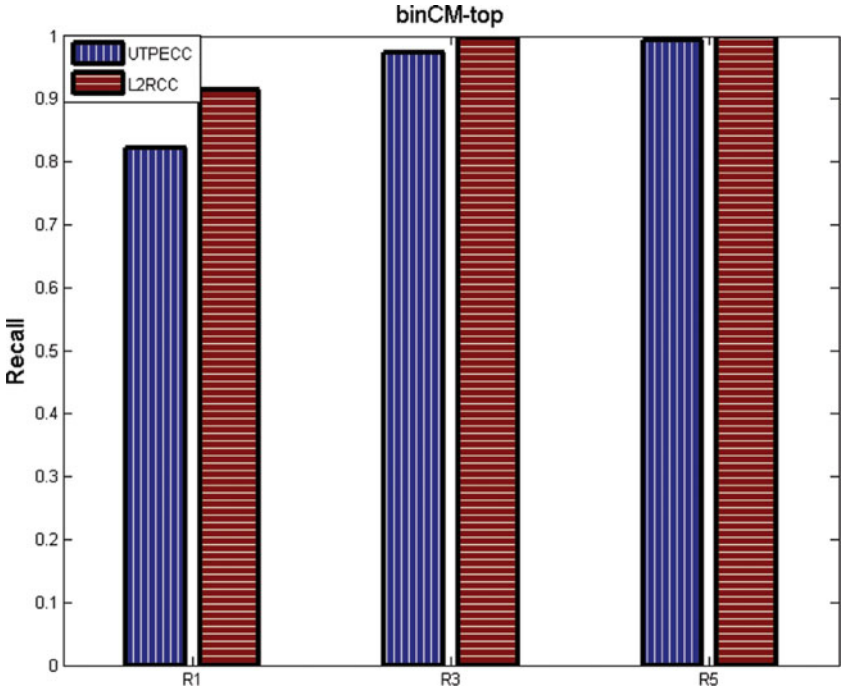


Fig. 4. (Colour online) Comparing the quality of UTPECC and our L2R comparable corpus using the binary comparability measure.

Comparability measures. Table 8 shows the results of comparing UTPECC and L2R comparable corpus in terms of the comparability measures. The results show that our L2R comparable corpus has alignments with higher quality and improves binCM by 11.3 percent compared to UTPECC.

In addition, we compare the two comparable corpora based on the recall measure as described before, where relevant documents are only judged by the cosine comparability measure. The results are depicted in Fig. 4, which show that our approach outperforms the previous approach.

Performance of CLIR. We also compare the L2R comparable corpus (built using our approach) with UTPECC (built in (Hashemi *et al.* 2010)) in terms of the quality of translations extracted from both comparable corpora. First, two translation models are obtained from UTPECC and L2RCC comparable corpora. For this purpose, stopwords are removed from both sides of comparable corpora. English words are stemmed using Porter stemmer. The translation models are then evaluated based on the performance of CLIR. For each experiment, MAP and Precision at top ten documents (P@10) are reported. We also report the performance of monolingual retrieval using Ham'08 and Ham'09 query sets as a baseline for CLIR evaluation in Table 9. These results are obtained using the KL-divergence retrieval model.

To perform CLIR using the language modeling framework, we smooth document language models using Dirichlet Prior smoothing and do not tune the smoothing parameter μ which is set to the default value of 1,000. Performance of CLIR is reported in Table 9, which shows that using translations extracted from our

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Table 9. *CLIR performance using translations extracted from UTPECC and L2R corpus. '*' sign indicates statistical significant differences at the 0.01 level using a two-tailed paired t-test*

	Ham'08				Ham'09			
	MAP	Percent Mono	P@10	Percent Mono	MAP	Percent Mono	P@10	Percent Mono
Mono	0.4231	–	0.6460	–	0.3710	–	0.6020	–
UTPECC	0.1404	33.18	0.2340	36.2	0.0905	24.39	0.1653	27.45
L2RCC	0.2372*	56.06	0.3960*	61.3	0.1588*	42.8	0.2920*	48.5

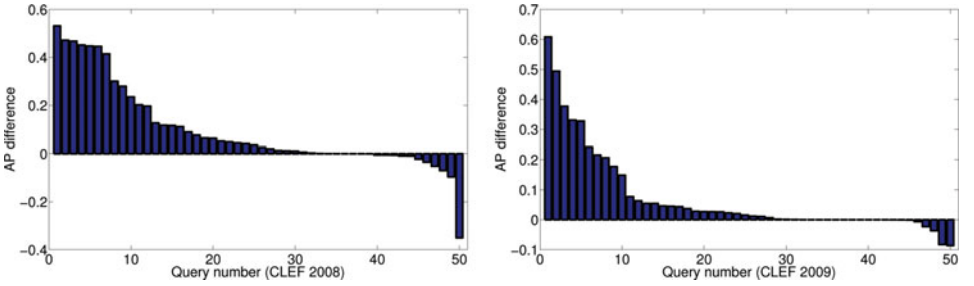


Fig. 5. (Colour online) Per query AP difference for topics in Ham'08 and Ham'09.

comparable corpus compared to those from UTPECC, MAP performance of CLIR is improved by sixty-eight percent and seventy-five percent for Ham'08 and Ham'09, respectively. This shows the higher quality of alignments built using our learning-based approach.

To provide a more complete analysis, we present the Average Precision (AP) difference between results obtained for each query using our L2R comparable corpus and UTPECC comparable corpus as the translation resource in Fig. 5. As the figure illustrates, AP of thirty-five out of fifty queries in Ham'08 is improved. In addition, seven out of nine queries that have zero AP using UTPECC (Hashemi *et al.* 2010) corpus, get non-zero precision values using our L2R corpus. Similarly, we reached better AP for forty of fifty queries in Ham'09 .

5 Conclusion and future work

In this paper, we proposed a new approach for building comparable corpora from Web data. Our approach is based on a learning to rank method to derive a ranking of candidate target documents for each source document. In this regard, we exploited features based on CLIR, cross-language document similarity and named entity. We also proposed to generate training data from a parallel corpus so that training data simulates the characteristics of comparable corpora. Employing a learning-based approach for building alignments of a comparable corpus allows the addition of any language pair- or domain-specific feature. We used our approach to build a large-scale multi-domain comparable corpus from the Web and demonstrated its

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

quality by different evaluation measures. In addition, we compared the proposed approach for building comparable corpora with the existing methods and indicated that our approach builds higher quality alignments through several experiments.

In the current work, we employed *de-facto* standard methods for removing noise from translation models and matching named entities. However, using more sophisticated methods, such as methods of Aker, Paramita and Gaizauskas (2013) and (Aker *et al.* 2014), for these steps may help to improve the quality of alignments. Another direction is to expand training data with aligned Wikipedia articles using multilingualism characteristic of Wikipedia, and study how much this training data affects the quality of learned ranking model and subsequently comparable corpora.

Looking forward, we are interested in defining more features to measure document similarity. We also aim to explore other learning to rank methods for training a ranking model suitable for the task of creating comparable corpora. In addition, extracting parallel segments from built comparable corpora to train a SMT is an important future work.

6 Acknowledgments

This research was in part supported by a grant from Institute for Research in Fundamental Sciences (No. CS1393-4-43).

References

AbduI-Rauf, S., and Schwenk, H. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'09, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 16–23.

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. 2009. Clef 2008: ad hoc track overview. In *Proceedings of the 9th Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*, CLEF'08, Berlin, Heidelberg, Springer-Verlag, pp. 15–37.

Aker, A., Kanoulas, E., and Gaizauskas, R. 2012. A light way to collect comparable corpora from the web. In N. C. C. Chair, K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Aker, A., Paramita, M., and Gaizauskas, R. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Soa, Bulgaria: Association for Computational Linguistics, pp. 402–411.

Aker, A., Paramita, M. L., Pinnis, M., and Gaizauskas, R. 2014. Bilingual dictionaries for all EU languages. In *The Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., and Oroumchian, F. 2009. Hamshahri: a standard Persian text collection. *Knowledge Based Systems* **22**(5): 382–387.

Azarbonyad, H., Shakery, A., and Faili, H. 2012. Using learning to rank approach for parallel corpora based cross-language information retrieval. In *Proceedings of 20th European Conference on Artificial Intelligence (ECAI)*, Montpellier, France, pp. 79–84.

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Braschler, M., and Schäuble, P. 1998. Multilingual information retrieval based on document alignment techniques. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, ECDL'98*, London, UK: Springer-Verlag, pp. 183–197.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* **20**(3): 273–297.

Dadashkarimi, J., Shakery, A., and Heshaam, F. 2014. A probabilistic translation method for dictionary-based cross-lingual information retrieval in agglutinative languages. In *Proceedings of the 3th Conference on Computational Linguistic*, CLConference'14, Tehran, Iran.

Darwish, K., and Oard, D. W. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR'03, New York, NY, USA. ACM, pp. 338–344.

Fang, H., Tao, T., and Zhai, C. (2011). Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems* **29**(2): 7:1–7:42.

Ferro, N., and Peters, C. 2009. Clef 2009 ad hoc track overview: TEL and Persian tasks. In *Proceedings of the 10th Cross-Language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*, CLEF'09, Berlin, Heidelberg, Springer-Verlag, pp. 13–35.

Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL'05, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 363–370.

Fung, P., and Cheung, P. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING'04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Garera, N., Callison-Burch, C., and Yarowsky, D. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL'09, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 129–137.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL'04*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 527–534.

Hashemi, H. B., and Shakery, A. 2014. Mining a Persian-English comparable corpus for cross-language information retrieval. *Information Processing and Management* **50**(2): 384–398.

Hashemi, H. B., Shakery, A., and Faili, H. 2010. Creating a Persian-English comparable corpus. In *Proceedings of the 2010 International Conference on Multilingual and Multimodal Information Access Evaluation: Cross-Language Evaluation Forum*, CLEF'10, Berlin, Heidelberg, Springer-Verlag, pp. 27–39.

Huang, D., Zhao, L., Li, L., and Yu, H. 2010. Mining large-scale comparable corpora from Chinese-English news collections. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China.* COLING'10, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 472–480.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'02, New York, NY, USA. ACM, pp. 133–142.

Li, H., and Hirst, G. 2011. *Learning to Rank for Information Retrieval and Natural Language Processing*. G - Reference, Information and Interdisciplinary Subjects Series. California, USA: Morgan & Claypool Publishers.

McNamee, P., Mayfield, J., and Nicholas, C. 2009. Translation corpus source and size in bilingual retrieval. In *Proceedings of Human Language Technologies: The 2009 Annual*

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short'09 Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 25–28.

Munteanu, D. S., and Marcu, D. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* **31**(4): 477–504.

Munteanu, D. S., and Marcu, D. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 81–88.

Nie, J.-Y. 2010. *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. California, USA: Morgan & Claypool Publishers.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1): 19–51.

Pal, S., Pakray, P., and Naskar, K. S. 2014. Automatic Building and Using Parallel Resources for SMT from Comparable Corpora. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra) at EACL*. Association for Computational Linguistics, pp. 48–57.

Paramita, M. L., Guthrie, D., Kanoulas, E., Gaizauskas, R., Clough, P., and Sanderson, M. 2013. Methods for collection and evaluation of comparable documents. In S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung (eds), *Building and Using Comparable Corpora*, pp. 93–112. Berlin Heidelberg: Springer.

Pilevar, M. T., Faili, H., and Pilevar, A. H. 2011. TEP: Tehran English-Persian parallel corpus. In *Proceedings of the 12th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, Berlin, Heidelberg, Springer-Verlag, pp. 68–79.

Pomikálek, J. 2011. *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Rahimi, R., and Shakery, A. 2013. A language modeling approach for extracting translation knowledge from comparable corpora. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, Berlin, Heidelberg. Springer-Verlag, pp. 606–617.

Rahimi, R., Shakery, A., and King, I. 2016. Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework. *Information Processing & Management* **52**(2): 299–318.

Rahimi, Z., and Shakery, A. 2011. Topic based creation of a Persian-English comparable corpus. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology*, AIRS'11, Berlin, Heidelberg, Springer-Verlag, pp. 458–469.

Saad, M., Langlois, D., and Smaïli, K. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia - Social and Behavioral Sciences* **95**: 40–47.

Sheridan, P., and Ballerini, J. P. 1996. Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'96, New York, NY, USA. ACM, pp. 58–65.

Skadia, I., Aker, A., Mastropavlos, N., Su, F., Tufi, D., Verlic, M., Vasijevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In N. C. C. Chair, K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.), In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association 824 (ELRA).

Smith, J. R., Quirk, C., and Toutanova, K. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010*

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press

Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT'10, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 403–411.

Strötgen, J., Gertz, M., and Junghans, C. 2011. An event-centric model for multilingual document similarity. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'11, New York, NY, USA, ACM, pp. 953–962.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems* **25**(4), 1.

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. 2008. Focused web crawling in the acquisition of comparable corpora. *Information Retrieval* **11**(5): 427–445.

Tao, T. and Zhai, C. 2005. Mining comparable bilingual text corpora for cross-language information integration. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD'05*, New York, NY, USA: ACM, pp. 691–696.

Ture, F., Elsayed, T., and Lin, J. 2011. No free lunch: brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'11, New York, NY, USA, ACM, pp. 943–952.

Vulić, I., and Moens, M.-F. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'12, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 449–459.

Xu, J., and Weischedel, R. 2005. Empirical studies on the impact of lexical resources on CLIR performance. *Information Processing & Management* **41**(3): 475–487.

Zhai, C., and Lafferty, J. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, CIKM'01, New York, NY, USA, ACM, pp. 403–410.

https://doi.org/10.1017/S1351324916000164 Published online by Cambridge University Press