Creating Comparable Corpora through Topic Mappings

Firas Sabbah, Ahmet Aker
Department of Information Engineering, University of Duisburg-Essen
{sabbah, aker}@is.inf.uni-due.de

Abstract

Aligning multilingual documents is considered one of the most important steps in building comparable and parallel corpora. Bilingual lexicons have commonly used to detect the similarity level between the bilingual documents. However, high quality bilingual lexicons are not free and not readily available for many language pairs. In this work, we present a new approach to detect the similarity level between documents written in two different languages. The basic idea is to analyze the topical structure of texts and use it for detecting the similarity level between the documents. The results show that enhancing the lexicon-based methods by the topical structures improves the alignment process. Besides the model, this work introduces a tool for automatic comparable document search for English-Arabic languages.

Keywords: Comparable Corpora, Document Alignment, LDA, topic mapping

1 Introduction

In many cases an event is captured by many new agencies and reported in diverse languages. Being able to track all news about the same event opens many doors for different kind of analyses such as understanding how different countries observe the event, what are their agreement and disagreements in terms of argumentations, what are the reactions of respective readers[1], where are the topical focuses, etc. to name just few.

In our broader research agenda we have the vision to perform multi-lingual argument mining and perform analyses about the differences and commonalities between the arguments found in two different articles reported in two different languages. Our current focus is in English and Arabic. To perform this there are several steps: (1) determining comparable documents, (2) annotating both articles for arguments, (3) aligning arguments and finally (4) making sense of the aligned arguments. The focus of this paper, however, is at step 1 which is also the backbone of the later tasks.

Two documents written in two different languages are comparable if they talk about the same topic or event. Related work (see Section 2 for details) have investigated different ways for obtaining comparable corpora – data collection containing large sets of comparable documents.

In our work we focus on topic mappings. For this we use the Latent Dirichlet Allocation (LDA) to extract the topics of both source and target documents. Each topic is represented by a set of key words. We do this for each language separately. Then we map topics which result in a topic dictionary allowing us to query with source language topics and obtain topics in the target language. With this our approach becomes independent of translation sources which would be needed for translating source topics to target key words. However, we also extract traditional translation based features to boost the alignment performance. Both topic mappings and translation based features are combined to determine the similarity level between two documents written in two different languages. We integrated this solution into a tool enabling users to search for documents in the source language and also automatically retrieve documents in the target language which are comparable to the source documents.

In Section 2 we discuss related work. Next, in Section 4 we introduce our method of aligning the documents. We provide the evaluation results in Section 5. Next in Section 6 we present the tool for automatic comparable document search. Finally, we conclude the paper in Section 7.

2 Related work

Indeed, many approaches for creating comparable corpora were proposed. A common paradigm for obtaining a comparable corpus involves collecting monolingual data for each language and matching documents by comparing document contents (Talvensaari et al., 2007; Hashemi et al., 2010; Aker et al., 2012). These methods have one common aspect; they extract the top keywords of an English text, perform automatic translation of these to the target language and perform the pairing based on the source and translated key words.

Other studies(LU et al., 2013; Kraaij et al., 2003) use the page structure and URLs to detect the similarity level between the documents. The idea of similarity in these studies is that the HTML structure and the document path URL of the source and the target documents have to share an acceptable level of symmetry.

Since Wikipedia supports the inter-language links for its articles, we notice the intensive usage of such resource to produce such corpora (Adafre and De Rijke, 2006; Saad et al., 2013). These studies focus on how to measure the quality of similarity between the Wikipedia pages, and to select the similar articles for building comparable corpora.

Topical structures have been also used for building comparable corpora. (Zhang et al., 2013) propose a model to mine bilingual topics from Wikipedia in order to tackle the problem of cross-lingual linking. In this study the similarity is a score computed by the inner product of topic distributions of the documents. (Zhu et al., 2013) uses also the topics of documents to measure the similarity. The similarity value is calculated using three different measures: Kullback-Leibler (KL), Cosine Similarity and Conditional Probability. For these measures, the similarity is defined by the closeness of

[1] This assumes that each article has available reader comments.

a document to a specific topic.
In our work, however, we focus on topic mappings. The topic mappings do not rely on translation sources and are a way of bridging two articles written in two different languages. With this if a user determines topic of a source document she can easily query from the mappings how to express that topic in the target language and use the expression to look for documents expressing the mapped topic. We use this idea to align two documents written in two different languages. However, to boost the performance of the alignment we also make use of simple and light translation features and combine those together within an SVM classifier.

3 Data

For our targeted languages (English and Arabic), we extracted document collections from HuffingtonPost website [2]. However, HuffingtonPost is not the only news website that offers news in many languages. Tens of news agencies also offer multilingual news like BBC and Reuters. What makes HuffingtonPost different than other news agencies is that some HuffingtonPost news contains a specific phrase or link that leads to another version of HuffingtonPost that contains a near translation of the first article.

3.1 Collection method

For crawling the articles from HuffingtonPost we proceed the following steps: 1) We automatically track the news articles from the twitter page of the target language (Arabic-HuffingtonPost), 2) we check whether the news article has a parallel English version by searching the article text for specific phrases that indicate that news page is originally published by another source (this include phrases like "This article is translated from ..." or "this topic is originally published ..."), finally 3) we extract the texts of both documents.

3.2 The collected data

We crawled the articles over the period from July 2015 to July 2017. Over two years, Arabic-HuffingtonPost had published about 3543 Arabic Articles that have nearly parallel English versions. Table 1 presents a detailed information about the crawled data. The crawled collections cover political, sport, science, technology as well as life style domains. The data[3] is publicly available on GitHub.[4]

English articles	3543
Arabic articles	3543
Total number of English words	2320583
Total number of Arabic words	2153295
Total number of unique English words	74255
Total number of unique Arabic words	154957

Table 1: Collected data specifications

4 Methodology

In order to detect comparable documents we make use of topic mappings between source and target languages. Given a pair of documents (English-Arabic), we extract LDA topics from both documents.
Next, we measure how strongly the topics correlate and decide based on this how strongly comparable the pair of documents is. However, since the LDA topic extraction is performed independently on each document and the topic-describing words are written in two different languages it is not straightforward to compute the topic similarities. One way of doing this is through using dictionaries for translating from one language to other and compute a similarity metric over them. Another way is to generate topic mappings and use them instead of translation dictionaries. In this work we adopt the latter approach.
In the next sections we describe how we create our topics and the topic mappings. We also describe how the mapping information is transferred to features to perform the alignment process. In addition to mapping information, we also make use of traditional features which are also outlined in this section. Figure 1 presents an overview of our methodology phases.

4.1 Training LDA models

LDA (Blei et al., 2003) is a statistical unsupervised learning algorithm. It generates a distribution of how objects constitute hidden themes and how different objects constitute observable entities. LDA regards the hidden topics as a group of tightly co-occurring words.
We learn LDA models for both English and Arabic documents described in Section 3, extract topics from both document collections and align the topics. To align the topics there is an assumption that the pair of documents have an acceptable level of comparability which is the case for the HuffingtonPost data.
Before training, we pre-processed our collected dataset by removing the stop words from both languages, and applying further text processing on the Arabic dataset. To train an LDA model over a dataset, we have to know the following variables: First, we need to decide the number of topics which should be used to produce the best words divisions. Of course, the number of the topics is pertinently related to the dataset size, more documents in the dataset means more vocabulary which implies more topics. Therefore, the number of topics is determined experimentally. From the experiments, we find that the topics number around 70-75 is giving us the best results within HuffingtonPost dataset.
Secondly, we need to decide on the values of LDA parameters alpha and beta. The document-topic density is represented with alpha, the higher alpha, the more topics documents contain. The topic-word density is represented with beta. The higher beta, the more words from the corpus a topic contains. After experiments, we find that using 1.0 for alpha and 0.1 for beta is providing the best division of

[2] https://www.huffingtonpost.com

[3] Due to copyright issues we only publish the URLs to the English and Arabic articles.

[4] https://github.com/fsabbah/lda-comparable-corpora/blob/master/en_ar_URLs.csv

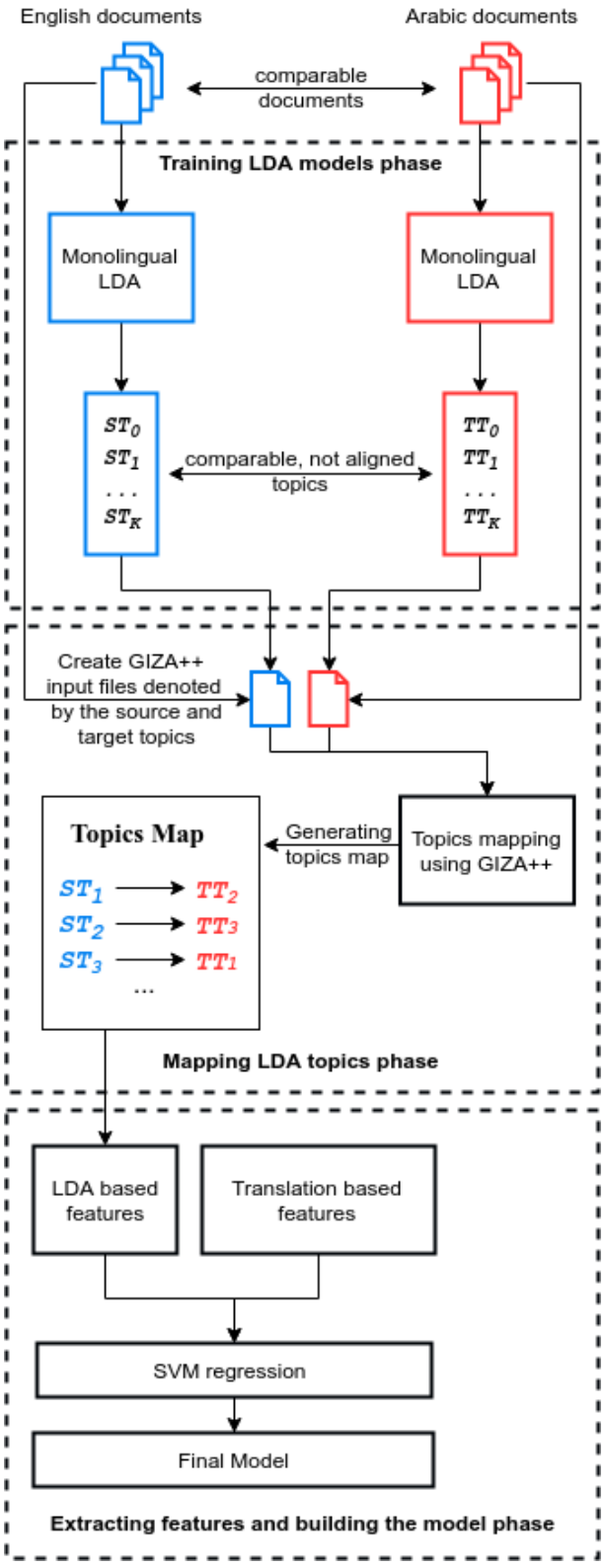

Figure 1: Methodology phases

topics. To perform topic modeling, we used the LDA implementation within the mallet project[5].

[5] http://mallet.cs.umass.edu/

4.2 Mapping LDA topics

The training phase of LDA produces two sets of topics, one for each language. The topic mapping process aims to match the LDA topics of the source and the target languages. For matching topics, we use GIZA++ tool. Since each pair of English and Arabic documents are translation of each other or strongly comparable, we can assume that they share exactly the same or similar topics, just expressed in different languages. Our aim is to find the topic mappings and use this knowledge for finding comparable corpora. For this purpose we create analog to parallel sentence files parallel topic files where the English file contains the ids of the English topics and the Arabic file contains the ids of the Arabic topics. The files are line-aligned where a pair of English-Arabic lines represent a pair of English-Arabic documents. In our approach, we use topics which have at least 5% probability according to LDA. To also express the frequency of a topic or its coverage within a document in each line, we repeat the topic id according to its probability in the original document. For example, if we have probability of 80% of a topic within a document, then we repeat for the document line the topic id eight times in case LDA topics number is selected as K=10.

In its original setting Giza++ produces words alignment. In our case the words are topics. Using this GIZA++ output, we are able to build a mapping matrix between the source and the target topics. Table 2 presents an example map between topics.

	ST_0	ST_1	...	ST_k
TT_0	0.12	0.29	...	0.02
TT_1	0.81	0.05	...	0.01
...	...	...	...	...
TT_k	0.49	0.28	...	0.03

Table 2: Alignment of source and target topics. TT stands for target topic and ST for source topic.

As we see in Table 2 each source/target topic is aligned with every target/source topic. Each alignment is associated with a probability score which is computed by GIZA++. With this matrix it is possible to obtain for a given source topic all target topics which are above a specific probability, determine target documents entailing those topics and based on results make statements about the similarity between the source document and the determined target documents.

Table 3 presents examples of the aligned pairs of topics. These topics contain only the top 20 terms per topic.

4.3 Extracting features and building the model

We use Support Vector Machines (SVMs) with a linear kernel and the trade-off between training error and margin parameter C = 1 for the alignment purposes. Within the classifier, the used features are extracted from the trained LDA models and their topics mapping for the 3366 near parallel articles. Furthermore, we also make use of features extracted using a home-trained GIZA++ dictionary.

English topic	Best aligned Arabic topic	Translation of the Arabic topic
sugar, diet, fat, weight, foods, eat, eating, healthy, health, food, calories, high, body, drinks, risk, energy, low, per, blood, protein	تناول، سكر، غذائية، وزن، طعام، اطعمة، غذائي، جسم، صحية، دهون، تحتوي، نظام، حرارية، كمية، نسبة، سعرات، قلب، تغذية، اصابة، صحي	eating, sugar, food, weight, food, foods, body, healthy, fat, contain, system, calories, quantity, ratio, calories, heart, nutrition, injury, healthy
trump, president, trumps, donald, house, white, obama, washington, campaign, former, election, elect, administration, york, national, presidential, bush, presidency, office, america	ترامب، رئيس، اوباما، اميركية، اميركي، ولايات، دونالد، ابيض، اميركا، بيت، اميركيين، واشنطن، ادارة، جمهوري، لترامب، منتخب، بوش، رئاسة، انتخابية، حملة	trump, president, obama, american, american, states, donald, white, america, house, americans, washington, administration, republican, trump, team, bush, presidency, electoral, campaign
iraq, isis, iraqi, mosul, city, forces, islamic, baghdad, state, sunni, shia, war, saddam, battle, fighting, falluja, government, kurdish, people, iraqs	تنظيم، داعش، عراق، دولة، قوات، اسلامية، مدينة، قاعدة، موصل، معركة، عراقية، ابو، عراقي، ميليشيات، بغداد، عمليات، حسين، حرب، عبد، سوريا	organization, isis, iraq, state, forces, islamic, city, base, mosul, battle, iraqi, abu, iraqi, militias, baghdad, operations, hussein, war, abdul, syria

Table 3: English and Arabic topics represented by top 20 LDA words.

4.3.1 LDA-based features

The procedure of extracting the LDA based features proceeds the following steps: 1) for each document in the training dataset, we fetch the top LDA topics from the trained LDA model, 2) we connect each document in the source dataset to two documents in the target dataset (correct and incorrect target documents), 3) from each connection, we extract four features related to each top LDA topic.

To fetch the top LDA topics of a document we infer the probabilities of topics from a document. We sort the topics according to their probabilities. After that, we define two relationships between the source document and the target documents. The first one represents a positive case; it represents a connection between the source document and its correct aligned target document. The second represents the negative case; it is a link between the source document and a random document from the target dataset. We make sure the random document must not be the aligned target document of that source document. That means we create one correct connection and another wrong connection. For each connection, we extract the following features:

1. The probability of the top LDA source topic.
2. The probability of the top aligned target LDA topic (we find the top aligned target topic using the GIZA++ topics mapping).
3. The probability of the top LDA target topic.
4. The probability of the top aligned source LDA topics to that target topic.

Figure 2 shows the process of extracting the features of the top LDA source and target topics.

However, using only the features of the top topic is not enough to capture all topics within a document. To solve this, we used the top ten LDA topics. As described in the procedure above, we extract the same four features for each of these top 10 topics leading to 40 features in total.

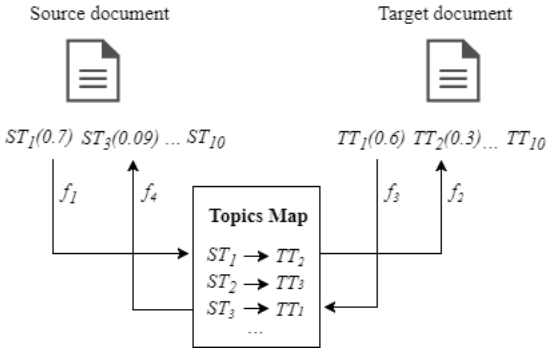

Figure 2: LDA topics features

4.3.2 Translation-based features

In order to improve the accuracy results, we added more features to the LDA-based features described above. This time we extracted the features from the texts. Since we need to determine the similarity between two different texts written in two different languages, we need to convert these texts or parts of them into one language. A translation system is the perfect tool used in these cases.

However, translation systems are not readily available. To overcome this problem, we used a home-trained GIZA++ dictionary. The parallel resources used for training and building this dictionary are brought from the OPUS project[6]. The main idea of using translation is to find how many similar words are shared between different parts of the source and the target texts. Such parts include titles, first and second sentences of both documents. In addition, we extracted also the most important 20 words of each document by calculating tf*idf values of the documents words. As we need numerical values for the classifier, we use cosine similarity to define a numerical value of the similarity between the original texts and the translated texts. As a result, we created the following features:

[6] http://opus.lingfil.uu.se/

1. The cosine similarity between the source document's title and the translated title of the target document,
2. The cosine similarity between the target document's title and the translated title of the source document,
3. Repeating these also for the first and second sentences in the source and target documents,
4. The cosine similarity between the top 20 tf*idf words of the source text and the translated top 20 words of the target text and
5. As in feature 4 with changing the direction of translation, i.e., from target text to source text.

In total, we collected 48 features, 40 from LDA topics and eight based on GIZA++ dictionaries. We set the similarity value 1.0 for each correctly aligned pair of documents, 0 for the connections that are not correctly paired.

5 Evaluation

For evaluation purposes we again use the huffingtonPost data described in Section 3. We split this data into a training (3366 articles) and a testing (177 articles) set. The training data is used to extract topic models and later to create the topic mappings (see previous section).
To evaluate our approach, we perform an automatic evaluation on the testing data. We compare LDA based features against the translation-based ones. In our evaluation we pair each English document with every Arabic document resulting in 177 pairs for each English document. Note among these 177 pairs there is only one pair that is correct. For all pairs features are extracted and SVM used to rank them. The document pair that is ranked top is evaluated whether it is the correct pair. If yes then we have a positive hit otherwise negative. Once we have repeated this for every English document we compute the accuracy scores which is the ratio of positive hits to all hits. Results are shown in Table 4. Note that the table shows only accuracy figures of the translation features. From the table we can see that best results are obtained when all translation-based features are combined. The combined translation-based features lead to close 69% accuracy.

Experiment	SVM classifier
Title	29.94%
Title + First sentence	40.67%
Title + First sentence + Second sentence	44.63%
20-top ranked tf*idf words	50.84%
Title + 20-top ranked tf*idf words	62.71%
Title + 20-top ranked tf*idf words + First sentence + Second sentence	68.92%

Table 4: Accuracy of the translation-based features

Figure 3 presents the results of the LDA-based features. Unfortunately LDA based features are not able to outperform the translation based features and achieve maximum

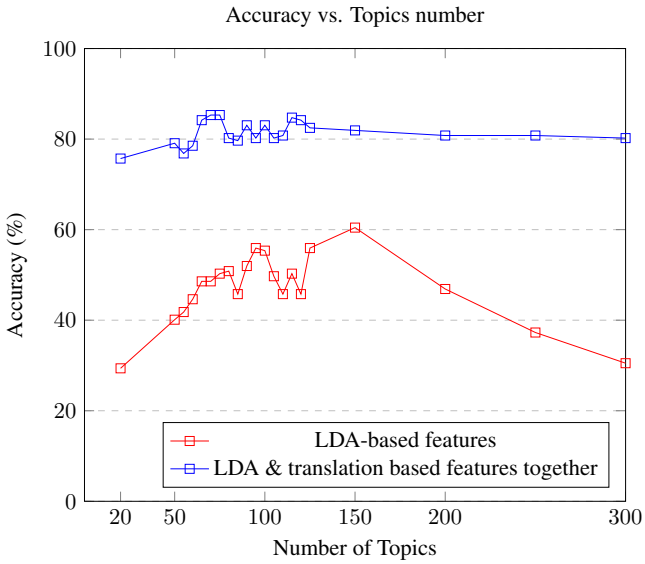

Figure 3: Accuracy of LDA-based features (alone and combined with translaiton based features).

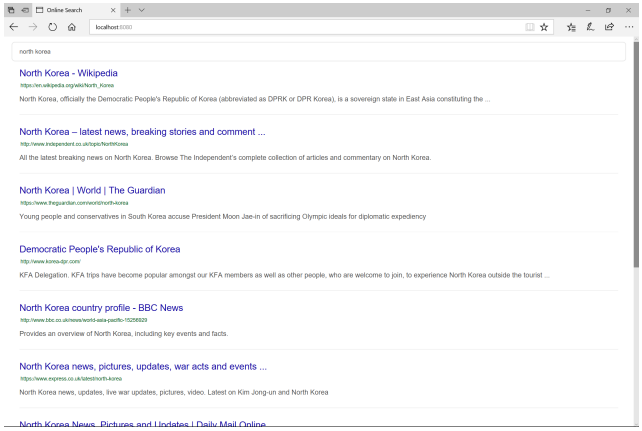

Figure 4: Search results for the English query.

60% accuracy (with K=150). However, we see that the LDA based features boost the results when they are combined with the translation ones. This is again shown in Figure 3 but this time with K=70-75. The combined approach leads to an accuracy of around 85%. This means in 85% the case our alignment is able to capture the correct target document of each source one. In our tool we use this combined approach to align documents.

6 Tool for comparable document search

Our current tool supports the gathering of Arabic documents comparable to an English document. The system allows users to enter English queries to search for English documents. The tool uses the Bing search API to search the web. The retrieved English documents are shown in a list similar to a search engine result list (see Figure 4).
Within the tool the user can select any English document and preview it before asking for comparable documents. To find the comparable Arabic documents, the tool first translates the title of the selected English document using an in

house created GIZA++ dictionary and uses the translated title to query for Arabic documents. Once the Arabic documents are retrieved it applies the alignment method described earlier to rank them. The user can then select the Arabic documents to display – this time the English and the Arabic document are displayed side by side (see Figure 5).

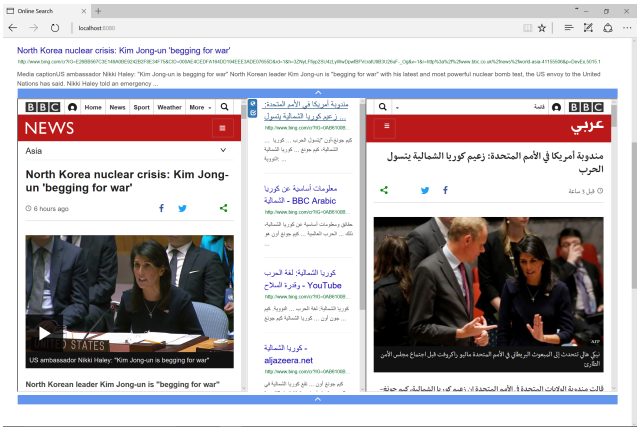

Figure 5: Documents are displayed side by side

7 Conclusion and Outlook

In this work we described a new approach for aligning English and Arabic documents for the purpose of comparable corpora construction. The proposed approach make use of LDA topics to analyze the topical structures of the documents. Based on the LDA topics we created topic mapping dictionary to automatically transfer a set of key-words describing the topics within the source document to the target language and use the transferred knowledge to judge whether two documents written in English and Arabic are comparable. Besides the topical mappings, we also use the traditional translation-based features to boost the alignment performance. Our results show that topic mappings as well as traditional features alone have performance around 60% to 70% accuracy. However, when both are combined the performance increases to the 80% level. We also integrated our alignment approach within a search tool that enables users to search for English documents, select an English document and retrieve Arabic documents comparable to the selected English document. The Arabic documents are ranked according to how comparable they are to the selected English document. In both cases the tools lets the user to read the articles.

In future we plan to further work on our vision to have a complete tool that supports multi-lingual argument mining. We will enhance our current tool with state-of-the-art argument mining approaches to determine arguments in the English and Arabic documents. However, due to the lack of argumentative training data for the Arabic language we will use for now only English argument mining solutions, tag English arguments and investigate mappings of those English arguments to the Arabic document. In terms of argument mapping we will follow the strategy discussed in (Aker and Zhang, 2017). However, in close future we aim to construct Arabic argument mining solutions using the data collection idea described in (Sliwa et al., 2018).

8 Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

9 Bibliographical References

Adafre, S. F. and De Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Aker, A. and Zhang, H. (2017). Projection of argumentative corpora from source to target languages. In *Proceedings of the 4th Workshop on Argument Mining*, pages 67–72.

Aker, A., Kanoulas, E., and Gaizauskas, R. J. (2012). A light way to collect comparable corpora from the web. In *LREC*, pages 15–20. Citeseer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Hashemi, H. B., Shakery, A., and Faili, H. (2010). Creating a persian-english comparable corpus. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 27–39. Springer.

Kraaij, W., Nie, J.-Y., and Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29(3):381–419.

LU, Y., ZHANG, X., and ZHENG, D. (2013). Automatic english-chinese parallel corpus acquisition and sentences extraction. *Journal of Computational Information Systems*, 9(6):2365–2372.

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.

Sliwa, A., Ma, Y., Liu, R., Borad, N., Fatemeh Ziyaei, S., Ghobadi, M., Sabbah, F., and Aker, A. (2018). Multilingual argumentative corpora in english, turkish, greek, albanian, croatian, serbian, macedonian, bulgarian, romanian and arabic. In *Proceedings of LREC 2018*.

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

Zhang, T., Liu, K., Zhao, J., et al. (2013). Cross lingual entity linking with bilingual topic model. In *IJCAI*.

Zhu, Z., Li, M., Chen, L., and Yang, Z. (2013). Building comparable corpora based on bilingual lda model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 278–282.