Creation of Comparable Corpora for English-{Urdu, Arabic, Persian}

Murad Abouammoh, Kashif Shah*, Ahmet Aker*
King Saud University, KSA, *The University of Sheffield, UK
E-mail: muabouammoh@ksu.edu.sa, kashif.shah@sheffield.ac.uk, ahmet.aker@sheffield.ac.uk

Abstract

Statistical Machine Translation (SMT) relies on the availability of rich parallel corpora. However, in the case of under-resourced languages or some specific domains, parallel corpora are not readily available. This leads to under-performing machine translation systems in those sparse data settings. To overcome the low availability of parallel resources the machine translation community has recognized the potential of using comparable resources as training data. However, most efforts have been related to European languages and less in middle-east languages. In this study, we report comparable corpora created from news articles for the pair English –{Arabic, Persian, Urdu} languages. The data has been collected over a period of a year, entails Arabic, Persian and Urdu languages. Furthermore using the English as a pivot language, comparable corpora that involve more than one language can be created, e.g. English- Arabic - Persian, English - Arabic - Urdu, English – Urdu - Persian, etc. Upon request the data can be provided for research purposes.

Keywords: Comparable Corpora for Arabic, Urdu, Persian and English

1. Paper

Statistical Machine Translation (SMT) relies heavily on the quality and quantity of bilingual parallel corpora (Brown et al., 1993, Och and Ney, 2002, Koehn, 2010). Most known parallel corpora are collected from the United Nation or the European parliament and are mostly about the legal domain (Koehn, 2005). However, often parallel resources are not readily available for under-resourced languages or other specific narrow domains. This leads to under-performing machine translation systems in those sparse data settings. To overcome the low availability of parallel resources the machine translation community has recognized the potential of using comparable resources as training data (Rapp, 1999, Munteanu and Marcu, 2002, Sharoff et al., 2006, Munteanu and Marcu, 2006, Kumano et al., 2007, Kauchak and Barzilay, 2006, Callison-Burch et al., 2006, Barzilay and McKeown, 2001, Nakov, 2008, Zhao et al., 2008, Marton et al., 2009, Aker et al., 2012)

Without a doubt the web is the largest source that can be used to gather comparable corpora (Resnik and Smith, 2003, Huang et al., 2010). Newswire have been also explored to gather comparable corpora. The ACCURAT[1] project, for instance, collected comparable corpora for 12 different European languages using the web, through Wikipedia and News articles (Skadiņa et al., 2012a, Skadiņa et al., 2012b). A follow up project, TaaS[2], expanded this idea to all EU languages. In both projects it has been shown that comparable corpora have positive impact on SMT as well as in creation bilingual terminology resources. Apart from those mentioned studies and projects there have been many workshops focusing solely on building and using comparable corpora (BUCC[3]).

The amount and diverse studies and efforts show that comparable corpora is certainly a useful resource to determine helpful material for SMT but also for related studies such as bilingual term extraction, cross-lingual information retrieval, etc. However, most efforts have been related to European languages and less in middle-east languages.

In this work we report comparable corpora created from news articles for the pair of English-{Arabic, Persian, Urdu} languages. Furthermore, when English is used as the pivot language other pairs of comparable corpora such as English-Arabic-Persian, English-Arabic-Urdu, English-Urdu-Persian, etc. can be created. We started collecting this data on 9th of November 2014 and will continue collecting till the end of May 2016. The data is saved in weekly bins. This means every week-bin contains only articles published within a week period. We believe that such corpora is publically available it will enable researchers to:

- Use the data as benchmark to perform tasks for SMT: parallel units such as sentences, phrases and terms extraction.
- Obtain not only “general” but also more domain specific data. This can be achieved by applying domain classifiers prior the SMT data extraction or by just simply following the domain information of the article URLs.
- Analyse the influences of data size in SMT quality – through gradually increasing the number of weeks to extract SMT relevant data.
- Analyse how similar events are reported in various languages.
- Analyse how topics evolve within a language over the time as well as cross-lingually.

To the best of our knowledge such data that has been collected over a period of a year, is big in volume, entails Arabic, Persian and Urdu languages as well as the settings enabling researchers to perform different analyses have not been reported by earlier studies.

[1] http://www.accurat-project.eu/
[2] http://www.taas-project.eu/
[3] https://comparable.limsi.fr/

2. Related Work

There have been number of studies in literature that reported comparable corpora for English-Arabic (Munteanu and Marcu, 2005, Abdul-Rauf and Schwenk, 2009). The most recent comparable corpora that was built in Arabic, was by (Saad et al., 2013). For the Persian the most commonly known corpora was collected by Hashemi et al. (2010) and Hashemi and Shakery (2014). In regards to the Urdu language we are not aware of any comparable corpora.

Unlike related work we report comparable corpora covering not only Arabic and Persian but also Urdu. Furthermore, unlike related studies the genre from which our data is obtained is the same in all data sets, namely news. Because of this feature our data can be split into different domains – using some classifiers or simply making use of the news article URLs information where it is clear from which domain – such as politics, sports, economy, entertainment, etc. In fact this enables the extraction of not only arbitrary parallel units but parallel units specific to certain domains. Finally, as discussed earlier our data is saved in weekly bins and spans over a period of 12 months providing a set-up for various research questions (see earlier section).

3. Method: Collecting Comparable Corpora

Our approach of collecting comparable corpora entails two steps. In step 1 we download monolingual news articles. In step 2 we pair those articles from step 1.

3.1 Step 1: Downloading news articles

For each language we have manually collected RSS feeds. The table below shows some statistics about the number of RSS feeds used and number of web pages from which we obtained RSS feeds.

Language	# RSS	# Web pages
Arabic	289	44
English	270	48
Persian	292	23
Urdu	239	16

Table 1: RSS feeds counts and number of web pages from which the RSS feeds are collected.

News web-sites provide RSS feeds for different domains. The domains our RSS feeds cover are:

- Politics
- Sport
- Economy
- Art
- Culture
- Education
- Entertainment
- Business
- Parliament
- International news
- Gulf/Middle East news
- Local news

In periodic time frames – every 15 minutes – we visit the feeds and download the updated news articles using an in-house tool. We run this process for a week and stop. The data collected in that week is saved in a separate bin, called week1-bin. After this the process for the second week is started. Data resulting from the second week is saved in week2-bin. This process was started in 9th of November 2014 and will continue till the end May 2016.

3.2 Step 2: Paring News Articles

To align articles, i.e. to create comparable corpora, we pair articles from two different bins written in two different languages. Note we ensure that the time difference between those two bins is no greater than 7 days.

To align two articles written in two languages, from now on source and target languages, we make use of core terms extracted from the source article instead of the entire article.

To define core terms we have investigated a set of 1.7K news articles along with their user generated comments -- on average we have 206 comments per news article. From each news article we have extracted terms and analysed whether they have been also used in the user generated comments. Our analysis shows that 35% of the terms extracted from the news article are also mentioned in the comments. We also found out that mostly terms from the title and first sentence (55% and 60% respectively) were mentioned in the comments. Terms extracted from other parts -- sentences from 2-6 and sentences from 7-till the end of the article) were mentioned only around 45% and 33% respectively. Around 43% of comments mentioned at least one or more terms extracted from the article.

We use this analysis to extract core terms from news articles and use them for alignment. The core terms are only extracted from the source language – that is in our case English. To extract such core terms we first extract from the source document all nouns using the OpenNLP toolkit[4]. Then each noun is ranked according whether it is mentioned in the title, in the first sentence of the source document, in the following 5 sentences after the first sentence and the remaining part of the article that follows the 6th sentence. That means each term is assigned four different scores. We treat the title and the first sentence more important than the other two parts because terms extracted from these two parts were mentioned more in the comments than the terms extracted from the other two parts. Thus we assign to the first two parts a score of 2. For the remaining two parts we assign a score of 1. This means when a noun occurs in all parts it can have a maximum score of 6. Once scores are assigned for each term they are ranked according to their scores.

For further processing we only use 35% of the top scoring nouns -- a cut-off value we obtained experimentally through our analysis with the user comments. Each of these remaining nouns is translated into the target

[4] https://opennlp.apache.org/

language using GIZA++ dictionaries[5].

GIZA++ dictionaries can contain for each source word several target translations. However, each such source-target word translation is assigned a probability score indicating how likely it can be treated as true translation. In our alignment process we make use of those probability scores and use only the most likely translation for each source term. We use the translations to construct a query -- each term translation is separated by a white space and submit it to Lucene. Prior this we split each target document by white space and remove all stop-words as well as less significant words using tf*idf and index them using Lucene[6]. For each query we retrieve top 10 target documents from the Lucene API. Finally, we use the cosine angle between the target words and the translated source core terms to determine the similarity between the source and target articles. We only consider a pair of source and target article as positive pair when their cosine similarity is at least 0.1 -- a threshold selected empirically.

4. Data

Table 2 shows statistics about the monolingual data we have collected so far. As we can see there are some variations on the number of weeks for the different languages. The reason for this is that we initially started the collection process with only English and Arabic and added later the Urdu and Persian languages.

	Weeks Count	# of Articles	Word count	Unique word count
Arabic	60	1,170,108	610,692,653	4,227,095
English	60	368,276	151,293,107	1,547,581
Persian	41	453,781	300,312,095	1,376,337
Urdu	55	132,648	55,845,160	369,735

Table 2: Statistics about the monolingual data

We used the document aligner described earlier to pair English documents with documents in the other 3 languages. We also used the English documents as pivot to create comparable corpora containing different language pair settings. The statistics about the data are shown in Table 3.

Language pairs	# Document alignments
English - Arabic	48,646
English - Persian	4,716
English - Urdu	3,434
English - Urdu-Arabic	921
English - Urdu - Persian	682
English - Arabic - Persian	1,017

Table 3: Statistics about the comparable data

From Table 3 we see that the number of document pairs vary between the languages. The smallest bilingual comparable corpora is in English-Urdu and the biggest in English-Arabic. In the triple language pairs we have English-Urdu-Persian with the least number of documents triples. The biggest is in English-Arabic-Persian.

5. Conclusion

In this study, we collected monolingual corpora for the English, Arabic, Urdu and Persian languages from news using RSS feeds. This data has been collected over a period of a year. We use nouns extracted from the English documents, translate them into the target language using GIZA++ dictionaries and determine based on the translated nouns target documents that are comparable to the source (English) document. Using this method we created comparable corpora for the pair English –{Arabic, Persian, Urdu} languages. We also used English as the pivot language and obtained comparable corpora that involves more than 2 languages. Upon request the data can be provided for research purposes.

6. Acknowledgements

The authors would like to thank the Deanship of Scientific Research, King Saud University for its academic and financial support.

7. Bibliographical References

Abdul-Rauf, S., & Schwenk, H. (2009). *Exploiting comparable corpora with TER and TERp*. Paper presented at the Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora.

Aker, A., Kanoulas, E., & Gaizauskas, R. J. (2012). *A light way to collect comparable corpora from the Web*. Paper presented at the LREC.

Barzilay, R., & McKeown, K. R. (2001). *Extracting paraphrases from a parallel corpus*. Paper presented at the Proceedings of the 39th Annual Meeting on Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). *The mathematics of statistical machine translation: parameter estimation*. Comput. Linguist., 19(2), 263-311.

Callison-Burch, C., Koehn, P., & Osborne, M. (2006). *Improved statistical machine translation using paraphrases*. Paper presented at the Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the

[5] We automatically extracted the dictionaries for all language pairs using moses training scripts. We first downloaded the aligned.grow-diag-final-and files from OPUS \footnote {http://opus.lingfil.uu.se/} along with parallel corpus used to train these models. These files contain word alignments based on GIZA++ tool. Given these alignments, we extract lex files that contain maximum likelihood estimate. The parallel data to train these models consist of 955K sentences for persian-english, 10M for arabic-english and 726K for urdu-englsih.

[6] http:// lucene.apache.org

Association of Computational Linguistics.

DictMetric: ACCURAT D2.6 (2011). Toolkit for multi-level alignment and information extraction from comparable corpora, version 3.0. 29th June, 2012 164 pages. (http://www.accurat-project.eu/),

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

Hashemi, H. B., Shakery, A., & Faili, H. (2010). Creating a Persian-English comparable corpus *Multilingual and Multimodal Information Access Evaluation* (pp. 27-39): Springer.

Huang, D., Zhao, L., Li, L., & Yu, H. (2010). *Mining large-scale comparable corpora from Chinese-English news collections*. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters.

Kauchak, D., & Barzilay, R. (2006). *Paraphrasing for automatic evaluation*. Paper presented at the Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.

Koehn, P. (2005). *Europarl: A parallel corpus for statistical machine translation*. Paper presented at the MT summit.

Koehn, P. (2010). *Statistical Machine Translation*: Cambridge University Press.

Kumano, T., Tanaka, H., & Tokunaga, T. (2007). Extracting phrasal alignments from comparable corpora by using joint probability smt model. *Proceedings of TMI.*

Marton, Y., Callison-Burch, C., & Resnik, P. (2009). *Improved statistical machine translation using monolingually-derived paraphrases*. Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.

Munteanu, D. S., & Marcu, D. (2002). *Processing comparable corpora with Bilingual Suffix Trees*. Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10.

Munteanu, D. S., & Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Comput. Linguist.*, 31(4), 477-504. doi: 10.1162/089120105775299168

Munteanu, D. S., & Marcu, D. (2006). *Extracting parallel sub-sentential fragments from non-parallel corpora*. Paper presented at the Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.

Nakov, P. (2008). *Paraphrasing verbs for noun compound interpretation*. Paper presented at the Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008).

Och, F. J., & Ney, H. (2002). *Discriminative training and maximum entropy models for statistical machine translation*. Paper presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Rapp, R. (1999). *Automatic identification of word translations from unrelated English and German corpora*. Paper presented at the Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland.

Resnik, P., & Smith, N. A. (2003). The Web as a parallel corpus. *Comput. Linguist.*, 29(3), 349-380. doi: 10.1162/089120103322711578

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

Sharoff, S., Babych, B., & Hartley, A. (2006). *Using comparable corpora to solve problems difficult for human translators*. Paper presented at the Proceedings of the COLING/ACL on Main conference poster sessions.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., . Gaizauskas, R. (2012). *Collecting and using comparable corpora for statistical machine translation*. Paper presented at the Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., & Gornostay, T. (2012). *Analysis and evaluation of comparable corpora for under resourced areas of machine translation*. Paper presented at the The 5th Workshop on Building and Using Comparable Corpora.

Zhao, S., Niu, C., Zhou, M., Liu, T., & Li, S. (2008). *Combining Multiple Resources to Improve SMT-based Paraphrasing Model*. Paper presented at the ACL.