Multilinguality is gradually becoming ubiquitous in the sense that more and more researchers have successfully shown that using additional languages help improve the results in many Natural Language Processing tasks. Multilingual Multiway Corpora (MMC) contain the same sentence in multiple languages. Such corpora have been primarily used for Multi-Source and Pivot Language Machine Translation but are also useful for developing multilingual sequence taggers by transfer learning. While these corpora are available, they are not organized for multilingual experiments and researchers need to write boilerplate code every time they want to use said corpora. Moreover, because there is no official MMC collection it becomes difficult to compare against existing approaches. As such we present our work on creating a unified and systematically organized repository of MMC spanning a large number of languages. We also provide training, development and test splits for corpora where official splits are unavailable. We hope that this will help speed up the pace of multilingual NLP research and ensure that NLP researchers obtain results that are more trustable since they can be compared easily. We indicate corpora sources, extraction procedures if any and relevant statistics. We also make our collection public for research purposes. Our key contributions are as follows:

1. We have systematically collected, organized (or identified) and made available N lingual corpora from European, TED talks, ILCI, Bible and UN corpora.
2. We have also made available some of the scripts we used to organize our corpus collection, thereby eliminating the need to write boilerplate code.
3. Our collection spans 5 domains and can be used for studies on domain adaptation and transfer learning.
4. We have also organized some of the corpora by grouping languages according to language families to facilitate NLP for related languages and investigate how language relatedness impacts various transfer learning tasks.
3. Extraction Procedures and Languages

3.1. Corpus Extraction Mechanism

Most corpora are not directly available as N-way corpora but as N-1 bilingual corpora where one of the languages is always English. As such we simply extract the N-way corpus by retaining the N-1 sentences which have the same language as English. As such we simply extract the N-way corpora where one of the languages is always English. As such we simply extract the N-way corpora from text and XML files, publicly available. This corpus is not directly 5 lingual and thus uses English as the pivot we extracted 3, 4 and 5 lingual versions of the parallel corpus. We found that the test sets for 2011 and 2012 are not 5 lingual and thus exclude them from their collection.

We load all the corpora in a dictionary to ensure quick extraction. Although it might seem to quite memory intensive our method works well in practice since there are not too many corpora that are large enough to cause out of memory issues. Additionally we will make the scripts, for extracting the N lingual corpora from text and XML files, publicly available.

3.2. Languages

Our N-lingual corpus collection includes the following 59 languages: Afrikaans (af), Albanian (al), Arabic (ar), Bulgarian (bg), Cebuano (ce), Chinese (zh), Creole (cre), Croatian (cr), Czech (cz), Danish (da), English (en), Esperanto (es), Estonian (et), Farsi (fa), Finnish (fi), French (fr), German (de), Greek (el), Hebrew (he), Hindi (hi), Hungarian (hu), Icelandic (is), Indonesian (id), Italian (it), Japanese (ja), Kannada (kn), Korean (ko), Latin (la), Latvian (lv), Lithuanian (li), Macedonian (mk), Malayalam (ml), Malay (ms), Marathi (mr), Norwegian (no), Pashto (ps), Persian (fa), Polish (pl), Portuguese (pt), Punjabi (pu), Qırğız (qz), Romanian (ro), Russian (ru), Serbian (sr), Slovenian (sl), Somali (so), Spanish (es), Swedish (sv), Tagalog (tl), Thai (th), Turkish (tr), Ukrainian (uk), Urdu (ur), Vietnamese (vi), Xhosa (xh), Yoruba (yo), Zulu (zu).

We believe that previous IWSLT tasks showcase bilingual corpora that can be converted into N lingual versions but we leave this for future work.

Table 1: Statistics for the the N-lingual corpora extracted from the IWSLT 2016 corpus for the languages French (Fr), German (De), Arabic (Ar), Czech (Cs) and English (En)

| Type     | Languages | train | dev2010 | tst2010/tst2013 |
|----------|-----------|-------|---------|-----------------|
| 3 lingual| Fr, De, En | 191381| 880     | 1060/886        |
| 4 lingual| Fr, De, Ar, En | 84301| 880     | 1059/708        |
| 5 lingual| Fr, De, Ar, Cs, En | 45684| 461     | 1016/643        |

1In this folder we include readme files for each corpus collection for additional details which are not included in this paper due to the limited number of pages.

2https://conferences.unite.un.org/uncorpus

3We believe that previous IWSLT tasks showcase bilingual corpora that can be converted into N lingual versions but we leave this for future work.

4https://www.ted.com/talk?page=N (where N is an integer greater than 0)

5http://lotus.kuee.kyoto-u.ac.jp/raj/mmc4nlp/iwslt2016/
Table 2: Statistics for the the N-lingual corpora extracted from the IWSLT 2017 corpus for the languages Dutch (Nl), German (De), Italian (It), Romanian (Ro) and English (En)

| Type       | Languages            | train  | dev2010 | tst2010 |
|------------|----------------------|--------|---------|---------|
| 3 lingual  | De, Ni, En           | 172735 | 715     | 1363    |
| 5 lingual  | De, Ni, It, Ro, En   | 145105 | 627     | 1153    |

and we only considered the languages which contain 30000 or more translated verses. Another problem is that some verses (each of which have a unique id in the XML file) of the bible are not available in some of the XML files leading to fewer number of N lingual entries.

In order to make it easier for researchers to work on languages belonging to the same language family we extracted N lingual corpora for the following 8 language families: Slavic, Uralic, Indo Aryan, Dravidian, Germanic, East Asian, South East Asian and Romance. For all these corpora groups we remove the last 2000 sentences (from the end of the corpus) and split them into development and test sets of 1000 sentences each. We extract the following N lingual versions of the corpus where English is always one of the languages:

- 55 lingual spanning all the languages mentioned in Section 3.2, except Punjabi, Tamil, Taiwanese and Latvian. We chose these 55 languages since these are the only ones that are completely translated (with the exception of a few accidental omissions). This 55 lingual corpus contains 26121 lines and the missing 5000 lines are a result of the randomly missing translations for a number of verses.
- 9 lingual Slavic languages corpus of 30350 lines which includes bg, cr, cz, en, po, ru, se, sl and sv.
- 8 lingual Romance languages corpus of 30133 lines which includes en, fr, it, la, pt, ro and es.
- 5 lingual Indo Aryan languages corpus of 30049 lines which includes en, hi, mr, my and ne.
- 8 lingual Germanic languages corpus of 28854 lines which includes af, da, nl, en, de, ic, no and sw.
- 4 lingual Dravidian languages corpus of 30651 lines which includes en, kn, ml and te.
- 4 lingual East Asian languages corpus of 31063 lines which includes en, zh, ja and ko.
- 7 lingual South-east Asian languages corpus of 29621 lines which includes en, ce, id, ma, ta, th and vi.
- 4 lingual Uralic languages corpus of 30885 lines which includes en, et, fi and hu.

4.2.2. IWSLT 2017 corpus

The IWSLT 2017 task focused on 5 languages (de, nl, it, ro and en) but the objective was on a single multilingual system. One development set (dev2010) and one test set (tst2010) were provided. Like the IWSLT 2016 corpus this corpus is not directly 5-lingual either. We extracted 3 lingual (de, nl and en) and 5 lingual (de, nl, it, ro and en) versions of the corpus. We do not list the 4 lingual version since it is of roughly the same as the 5 lingual corpus. Refer to table 1 for details.

4.2.3. Generic Ted Talks corpus

We found an unofficial 13 lingual (ar, de, es, fr, he, it, ja, ko, nl, pt (Brazilian Portuguese), ru, zh (Mainland Chinese) and tw (Taiwanese Chinese)) TED talks corpus of 349049 lines which was crawled from the TED talks site. This repository also contains many pairs of bilingual corpus but one unusual aspect of this corpus is that it does not contain English as either a source or a target language. A 4 lingual version of this corpus which spans only ja, ko zh and tw contains an additional 40K lines for a total of 389764 lines. Since, there is no specific development or test set in the case of this 13 lingual corpus we created our own splits. For both these 4 and 13 lingual corpora we remove the last 4000 sentences (from the end of the corpus) and split them into development and test sets of 2000 sentences each. We believe that this corpus should be useful for future IWSLT tasks which focus on multilinguality.

4.3. Bible corpus

The Bible corpus is probably the only corpus which is translated into over 100 languages. However the corpus downloadable online is present in XML format and needs to be pre-processed. We developed a simple XML parsing script that can produce a N-lingual version given the XML files for each language. While inspecting the corpus we discovered that the bible is not fully translated into about 40 of the languages and thus we exclude them from our collections. The English version of the bible contains about 31102 verses.

6http://lotus.kuee.kyoto-u.ac.jp/raj/mmcr4nlp/bible
7https://github.com/ajinkyakulkarni14/TED-Multilingual-Parallel-Corpus
8http://lotus.kuee.kyoto-u.ac.jp/raj/mmcr4nlp/ted

4.4. ILCI corpus

The ILCI corpus has been used frequently in the Indian Languages Machine Translation shared tasks in ICON 2014 and 2015. The ILCI corpus is a 6 lingual multilingual corpus spanning the languages Hindi, English, Tamil, Telugu, Marathi and Bengali was provided as a part of the task. The training, development and test set contain 45600, 1000 and 2400 6-lingual sentences respectively. Half of the corpus (train/dev/test) belongs to the tourism domain and the other half to the health domain. There is work...
on a 12 lingual equivalent of the ILCI corpus but it is not publicly available. Most of the sentences in the corpus are short and simple in terms of grammatical complexity. The average sentence length varies from 12 for Tamil (morphologically rich) to 17 for English (morphologically poor) indicating that most of the sentences are intended to be used in survival situations. This is quite different from the case of the UN corpus where the average sentence length for English is around 25.

4.5. Europarl corpus

The Europarl corpus covers the following 21 European languages: bg, cz, da, de, gr, en, es, et, fi, fr, hu, it, lt, nl, po, pt, sl, sv, and sw. Since the corpus available online is not directly N lingual but is available as XX-English bilingual pairs we used English to extract a 21 lingual corpus. One issue with extracting a large 21 lingual corpus is that for 11 out of the 21 languages only 30% of the full corpus is available. Countries such as Bulgaria and Romania joined the European Union in 2007 and hence the Europarl corpus has about 400K sentences for Bulgarian and Romanian as compared to about 2M sentences for English and Spanish, languages which have been present since the beginning. Just like we did for the Bible corpus, we also extract N lingual corpora for the following language families: Germanic, Slavic, Uralic, Baltic and Romance. Since English is the pivot language we used for extraction we include it in all the collections. As in the case of the TED corpus we remove the last 4000 sentences (from the end of the corpus for each group) and split them into development and test sets of 2000 sentences each. The details of the corpora are as follows:

- 21 lingual spanning all the languages. This corpus is of 189310 lines.
- 10 lingual spanning en, da, de, es, fi, fr, it, nl, pt and sv. These 10 languages are the largest in the collection. This corpus is of 1.071M lines.
- 6 lingual Slavic languages corpus of 342845 lines which includes bg, cz, en, po, sl and sv. We also extract a 4 lingual Slavic languages corpus of 569962 lines which includes sl, sv, cz and en. This increase in the number of lines is due to the exclusion of Polish and Bulgarian which contain significantly fewer number of entries in their respective monolingual corpora.
- 6 lingual Romance languages corpus of 294192 lines which includes en, fr, it, pt, ro and es. We also extract a 5 lingual version of 1.454M lines by excluding Romanian.
- 5 lingual Germanic languages corpus of 1.408M lines which includes da, nl, en, de, and sw.
- 4 lingual Uralic languages corpus of 563761 lines which includes en, et, fi and hu.
- 3 lingual Baltic languages corpus of 588273 lines which includes lt, li and en.

5. Conclusion

In this paper we have described our collection of Multilingual Multiway corpora. Our collection, which we believe to be the first of its kind, spans 59 languages and 5 domains. We have also extracted N lingual development and test sets from existing bilingual development and test sets for the IWSLT corpora. For the Bible, Europarl and TED corpora where official N lingual development and test sets are unavailable we defined our own training, development and test splits and encourage other researchers to use these splits for ease of comparison. Our collection can be used for NLP research in low (Bible and ILCI), medium (IWSLT and TED) as well as high resource (Europarl and UN) scenarios. In the cases of the Bible and Europarl corpora we have extracted N lingual corpora for various language families to facilitate research on how language relatedness affects the final results of NLP tasks. Due to page limit restrictions we do not give various statistics such as word count and average sentence length for each instance of the N lingual corpora we extracted but plan to include it later. We make our collection available to the public and also plan on expanding it in the future to improve coverage in terms of domains, number of lines and number of languages. Some additional and promising sources of N lingual corpora are: BTEC corpus (Paul et al., 2013) and the QED corpus. We also plan on looking into extracting bilingual parallel corpora from Wikipedia and then extract multilingual multiway versions of those corpora.

6. References

Agić, Z., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 268–272. Association for Computational Linguistics.

Agic, Z., Johannsen, A., Plank, B., Alonso, H. M., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. TACL, 4:301–312.

Asgari, E. and Moftakir, M. R. (2016). Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (weld) as a quantitative measure of language distance.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wi: Web inventory of transcribed and translated talk. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), pages 261–268, Trento, Italy, May.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. Language Resources and Evaluation, 49(2):375–395.

13 http://sanskrit.jnu.ac.in/ilci/index.jsp
14 http://www.statmt.org/europarl/
15 http://lotus.kuee.kyoto-u.ac.jp/raj/mmcr4nlp/europarl
16 http://lotus.kuee.kyoto-u.ac.jp/~raj/mmcr4nlp/
17 Although this corpus is large in terms of number of languages and lines it is not easily available
18 http://alt.qcri.org/resources/qedcorpus/
Jha, G. N. (2010). The tdil program and the indian language corpora initiative (ilci). In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta, may. European Language Resources Association (ELRA).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Conference Proceedings: the tenth Machine Translation Summit, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Och, F. J. and Ney, H. (2001). Statistical multi-source translation. In Proceedings of MT Summit, volume 8, pages 253–258.

Paul, M., Finch, A. M., and Sumita, E. (2013). How to choose the best pivot language for automatic translation of low-resource languages. ACM Trans. Asian Lang. Inf. Process., 12:14:1–14:17.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may. European Language Resources Association (ELRA).

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 30–34, San Diego, California, June. Association for Computational Linguistics.