The ILSP/ARC submission to the WMT 2018 Parallel Corpus Filtering Shared Task

Vassilis Papavassiliou **Sokratis Sofianopoulos**

Prokopis Prokopidis **Stelios Piperidis**

Institute for Language and Speech Processing/Athena RC
Athens, Greece
{vpapa, s_sofian, prokopis, spip}@ilsp.gr

Abstract

This paper describes the submission of the Institute for Language and Speech Processing/Athena Research and Innovation Center (ILSP/ARC) for the WMT 2018 Parallel Corpus Filtering shared task. We explore several properties of sentences and sentence pairs that our system explored in the context of the task with the purpose of clustering sentence pairs according to their appropriateness in training MT systems. We also discuss alternative methods for ranking the sentence pairs of the most appropriate clusters with the aim of generating the two datasets (of 10 and 100 million words as required in the task) that were evaluated. By summarizing the results of several experiments that were carried out by the organizers during the evaluation phase, our submission achieved an average BLEU score of 26.41, even though it does not make use of any language-specific resources like bilingual lexica, monolingual corpora, or MT output, while the average score of the best participant system was 27.91.

1 Introduction

There is a growing literature on using web-acquired data for constructing various types of language resources, including monolingual and parallel corpora. As shown in, among others, Pecina et al. (2014) and Rubino et al. (2015), such resources can be exploited in training generic or domain-specific machine translation systems. Nevertheless, compared to the acquisition of monolingual data from the web, construction of parallel resources is more challenging. Apart from the identification of document pairs that are translations of each other and can be crawled from multilingual websites, the extraction of sentence pairs and, crucially, the selection of sentence pairs of good quality are far from straightforward.

Zariņa et al. (2015) exploit already available parallel corpora in order to get word alignments, which are then used to identify mistranslations. Denkowski et al. (2012) use N-gram language models built from monolingual corpora to estimate probabilities of source and target sentences, in a manner of assigning high scores to grammatical sentences and lower scores to ungrammatical sentences and non-sentences such as site maps, large lists of names, and blog comments. Aiming to select sentence pairs of good adequacy and fluency, Xu and Koehn (2017) generate probabilistic dictionaries and n-gram models from Europarl corpora. Taghipour et al. (2011) and Cui et al. (2013) extract features based on translation and language models, and word alignments from the dataset under examination (i.e. this dataset is used to train models instead of using external language resources) and then apply unsupervised techniques such as outlier detection of estimated probability density and graph-based random walk algorithm to discard sentence pairs that are of limited or no importance. In the case of web acquired data, shallow features like aligners' scores, length ratio, and patterns in URLs from which the content was originated, have been proposed (Esplà-Gomis and Forcada, 2010).

In a different manner, many researchers have approached data selection as a domain-matching issue. For instance, Duh et al. (2013) proposed the use of a neural language model trained on a domain-specific corpus to identify in-domain sentence pairs in a large corpus.

This paper describes the submission of ILSP/ARC for the WMT 2018 Parallel Corpus Filtering shared task. The task consisted in cleaning a very noisy English-German parallel corpus of 104 million sentence pairs provided by the organizers, with each EN-DE sentence pair accompanied by a score generated by the Hunalign sentence aligner.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 941–946

https://doi.org/10.18653/v1/W18-6485

The participants were to assign a quality score for each sentence pair, with higher scores indicating sentence pairs of better quality. As reported in the shared task webpage[1], "Evaluation of the quality scores will be done by subsampling 10m and 100m [EN] word corpora based on these scores, training statistical and neural machine translation systems with these corpora, and evaluating translation quality on blind test sets using the BLEU score." Given that the organizers discouraged participants from subsampling the corpus for relevance to a specific domain (e.g. the news domain), domain adaptation approaches like the ones mentioned above seem to not fit this task.

In the shared task webpage, the organizers also released a development environment with configuration files and scripts that allowed participants to subsample corpora based on quality scores and to replicate the testing procedure with a development test set.

2 System architecture

Our submission system is based on the cleaning module of the ILSP Focused Crawler (Papavassiliou et al., 2013), an open-source toolkit[2] that integrates all necessary software[3] for the creation of high-precision parallel resources from the web in a language-independent fashion.

The toolkit and its cleaning module have been used in research projects like the European Language Resource Coordination for the acquisition of high-precision parallel language resources (Papavassiliou et al., 2018).

2.1 Noise in Web acquired parallel corpora

In a pipeline for the construction of parallel corpora from the web, shortcomings of each processing step may introduce errors, usually called "noise", that affect the quality of the final output. In this shared task, the data collection pipeline of the Paracrawl[4] project was adopted for the construction of the input (i.e. the raw, very noisy parallel corpus).

Many types of noise occur due to misses during parsing HTML pages and extracting their textual content. Such errors are typically introduced when HTML code is considered text and/or page encoding is not successfully detected. Moreover, inaccurate identification of paragraph limits may lead to wrong sentence splitting and, eventually, in the alignment of incomplete sentences. False negatives in the detection of boilerplate text (i.e. navigation headers, disclaimers, etc.) may result in large numbers of (near-)duplicate sentence pairs, which are of only limited or no use for the production of good-quality language resources.

Other errors concern the accuracy of the language identification process. Even when the language of a web page is correctly detected at document level, it is possible that small parts of the page are written in another language. Thus, ignoring language detection at paragraph or sentence level may lead to sentence pairs with the wrong language in the source and/or the target side. Finally, misalignments at document and/or sentence level generate sentence pairs that are not translations of each other.

2.2 Filter-based clustering

Given that the existence of the types of noise discussed above is not strongly influenced by the targeted language pair, we developed a language agnostic method with the purpose of clustering sentence pairs in respect of their quality, i.e. of their correctness and usefulness for training MT engines.

The first cluster, C_0, includes obviously noisy sentence pairs. We assign to these pairs a 0 score in order to prohibit their participation in the subsamples to be used for training. Sentence pairs in C_0 match one of the following patterns:

1. sentence pairs with too short or too long EN or DE sentences (after tokenization) that would have been excluded from the training phase according to the shared task configuration. By enforcing a sentence length between 1 and 80 tokens, and a sentence length ratio less than 9 tokens (i.e. by using the default values of the Moses SMT toolkit for cleaning a corpus before training an MT system), we remove 3.42% of the sentence pairs in the input corpus. Our intuition is that most of these sentence pairs are the result of wrong HTML parsing or encoding detection.

2. sentence pairs with an EN or DE sentence that does not contain any letter in the range

[1] http://www.statmt.org/wmt18/parallel-corpus-filtering.html

[2] http://nlp.ilsp.gr/ilsp-fc/

[3] Including modules for metadata extraction, language identification, boilerplate removal, document clean-up, text classification and sentence alignment

[4] https://paracrawl.eu/releases.html

of Unicode character sets relevant to Latin scripts. This pattern discards sentence pairs (11.12% of the input corpus) that are either the result of wrong encoding detection, or contain only dates, prices, flight numbers, dimensions, products' IDs, etc.

3. sentence pairs with identical text in both languages (after removing non-Latin characters mentioned above). These sentence pairs (9.94% of the input corpus) mainly contain boilerplate elements, dates, locations, etc.[5]

4. sentence pairs for which the EN or DE parts were not in the proper language as detected by the Cybozu language detection library.[6] Sentences in these pairs (13.01% of the corpus) were often French or Spanish. As with most language detectors, the accuracy of the tool is lower during the examination of short sentences.

5. sentence pairs (1.71% of the corpus) with unusual features (e.g. words with transitions from lowercase to uppercase and vice versa, consecutive identical letters, long sequences of very short words, etc.)

6. sentence pairs consisting mostly of URLs, and emails (1.42% of the corpus)

Table 1 provides examples of sentences grouped into C_0 by some of the criteria described above.

In the next step of our language agnostic approach, we clustered the remaining sentence pairs using shallow features that are likely to be related to correctness of sentence alignment. Specifically, we compared the sequences of digits and symbols (e.g. punctuation marks, % , $, etc.) on each side of the remaining sentence pairs. Depending on the results (i.e. same/different digits and same/different symbols), the following four clusters, ordered from worst to best, were constructed:

C_1 Different digits and different symbols

C_2 Different digits and same symbols

C_3 Same digits and different symbols

C_4 Same digits and same symbols

[5] In future work we plan to reconsider the usefulness of this pattern in preparing parallel corpora for NMT engines.

[6] http://code.google.com/p/language-detection/

Table 2 contains examples of sentence pairs grouped into clusters according to this approach.

In a final step we focused on the identification of (near) duplicates. In more detail, we normalized sentence pairs by lowercasing and removing non-Latin characters, and we examined if a sentence pair was identical to or was included in another sentence pair. When a duplicate was detected, we kept the sentence pair that belonged to a better cluster. If both sentence pairs belonged to the same cluster, we kept the longer one in terms of tokens.

By assigning the corresponding cluster number to each sentence pair as a score (i.e. 4 to pairs of C_4, 3 to pairs of C_3, etc), the sentence pairs in the provided noisy corpus were roughly ranked. We then ran the subsampling algorithm that was provided by the organizers in order to obtain the two datasets required from each participant. We noticed that the sizes of the resulting corpora exceeded the 10M and 100M EN word thresholds. This is explained by the fact that we provided only 5 scores (as many as the clusters) and the algorithm selects all sentence pairs for a score (staring from the highest) iteratively until the size of the selected subcorpus reaches the threshold. For instance, clusters C_4, C_3 and C_2 (i.e. sentence pairs with scores 4, 3 and 2 respectively) including more than 14M English words, were sampled for the 10M corpus! To overcome this shortcoming, in our final rankings each cluster is initially assigned to an integer of different scale (e.g. C_1 to score 10, C_2 to score 1000, etc). The score of each sentence pair is then calculated by adding the Hunalign score to the initial cluster score, with the purpose of ensuring the granularity of the scores and of keeping clusters well-separated. This ranking led to corpora of 626K and 5.7M pairs for the 10M and the 100M corpora, respectively.

For a submission based on an alternative ranking, we add the character length of each pair to the initial cluster score. Compared to the Hunalign-based scoring, this variant favors long sentences and thus results in significantly smaller corpora in terms of sentence pairs (221K pairs and 5.4M pairs for the 10M and 100M corpora, respectively).

3 Evaluation Results

In the evaluation experiments conducted by the organizers, four different translation systems were trained, namely (a) a Moses statistical system

	EN	DE	Aligner score
1	Relatively extreme values are also taken into account.	Relatively extreme values are also taken into account.	2.4
2	www.gamersglobal.de about Risen 2	www.gamersglobal.de ber Risen 2	6.3
3	wie gehts denn so?	wo hast deins denn her?	1.12381
4	5103 Dec 5104 JanFebMarAprMayJun-JulAug	4574 FebMAprMaiJunJulAugSepOkt	1.46471
5	Abstr. Appl. Anal. 2014, Art. ID 363925, 7 pp. 54H25 (45G10)	Fluct. Noise Lett. 5 (2005), no. 2, L275 L282. 82C31	1.26739

Table 1: Examples of sentence pairs grouped into C_0 by filters focusing on sentence pairs with 1) identical text in both languages 2) sentences consisting mainly of URLs/emails/dates 3) the sentence in the first/second column detected as non-EN/non-DE, respectively 4) unusual patterns like mixture of upper- and lowercase ; 5) long sequences of short words.

(Koehn et al., 2007) trained on the 10M EN word parallel corpus, (b) a Moses system trained on the 100M EN word parallel corpus, (c) a Marian neural translation system (Junczys-Dowmunt et al., 2018) trained on the 10M EN word parallel corpus and (d) a Marian system trained on the 100M EN word parallel corpus. For all systems the official WMT 2017 news translation test set was used as a development set. According to the shared task's settings, the quality of the machine translation system is measured by BLEU score (Papineni et al., 2002) on the (a) official WMT 2018 news translation test set and (b) another undisclosed test set, which is the union of 5 test sets listed in Tables 3 and 4.

Table 3 summarizes the evaluation scores obtained using the ranking based on the combination of clusters and Hunalign scores on the various test sets. Our submission had an average BLEU score of 26.41 on the different test configurations (4 systems evaluated over 6 test sets), while the average score of the best participant system was 27.91.

It can be seen that for all datasets the best results are obtained by the NMT systems over their equivalent SMT ones, with the top one being the NMT trained over the 100M English token German-English filtered corpus. For both the Moses SMT and the Marian NMT systems there is a significant increase of the BLEU score when increasing the size of the training corpus from 10M to 100M English tokens. Specifically, for the Moses system the average increase is 16.6%, while for Marian the average increase is 21.5%.

Similarly, Table 4 lists the evaluation scores obtained with the alternative ranking scheme using the sentence length information. This submission had an average BLEU score of 24.98 on the different test configurations. Again, the best results are obtained with the NMT system trained over the 100M corpus. When comparing the average BLEU scores between the 10M and 100M systems, the SMT system shows an increase of 15.2%, while the NMT system shows a huge increase of 58.5%. Interestingly, the performance of the NMT system trained on the 10M corpus is lower than that of the SMT one. This can be attributed to the fact that the 10M corpus comprises 221K long sentence pairs, a relatively small number of sentences for NMT systems, which evaluate fluency over entire sentences. The equivalent SMT system is rather unaffected, presumably because SMT systems are based on n-gram models.

By comparing the results of the two alternative ranking schemes, we conclude that their performances are similar for the 100M corpora. This is explained by the fact that their intersection is extremely high: 5.2M sentence pairs are included in the 5.7M and 5.4M sentence pairs selected with the two schemes. Regarding the 10M corpora which differ significantly in number of sentence pairs (626K vs 221K), the performance of both schemes is similar for the SMT systems but differs for the NMT ones. In future work, we plan to carry out experiments that will provide evidence of how size and length of sentence pairs in a training corpus affect the performance of an NMT system.

4 Conclusions

In this paper we described the ILSP/ARC submission to the WMT 2018 Parallel Corpus Filtering

Cluster	EN	DE	Aligner score
C_2	We offer 2 comfortable bedrooms, sleeping up to 4 guests, a cot	Zwei komfortable Schlafzimmer für bis zu 4 Personen, Kinderbett	0.41805
C_2	The table now has 2 columns for the 2 euro commemorative coins, because some countries will issue two different 2 euro special coins. A description can be viewed by holding the mouse over the i-symbol for a while.	Es gibt in der Tabelle 2 Spalten für 2 Euro Gedenkmnzen, da seit 2007 einige Länder mehrere 2 Euro Sondermünzen ausgeben. Über das i-Symbol kann die entsprechende Bezeichnung der Münzen angezeigt werden.	0.49576
C_3	Our club for runners who have finished in Düsseldorf 10 times. We would like to honour this accomplishment.	Unser Club für alle Läufer, die bereits 10 Mal in Düsseldorf gefinished haben. Diese besondere Leistung, möchten wir auch besonders würdigen.	1.5466
C_4	Austrian declaration of principles at the Conference on Security and Cooperation in Europe (Helsinki, December 1972)	Grundsatzerklärung Österreichs auf der Konferenz über Sicherheit und Zusammenarbeit in Europa (Helsinki, Dezember 1972)	3.9431
C_4	A current application: The turbine sheets of the new Airbus A 380 were manufactured by a milling machine equipped by a self carrying product of WeBe Electronic GmbH.	Eine aktuelle Applikation: Die Turbinenblätter des neuen Airbusses A 380 von einer mit einem selbsttragenden WeBe-Produkt ausgerüsteten Fräsmaschine gefertigt.	2.6620

Table 2: Examples of sentence pairs grouped to different clusters based on the shallow features detailed in Section 2.2.

	SMT 10M	SMT 100M	NMT 10M	NMT 100M
news2017	20.49	25.28	26.09	31.46
news2018	26.30	30.58	31.32	38.99
iwslt2017	18.83	22.82	21.20	26.57
Acquis	18.71	22.27	22.94	27.63
EMEA	26.50	30.88	30.17	35.96
GlobalVoices	20.20	23.43	23.39	28.20
KDE	23.78	26.74	25.73	30.63
average	22.39	26.12	25.79	31.33

Table 3: BLEU evaluation scores (ranking was based on the combination of clusters and Hunalign scores)

	SMT 10M	SMT 100M	NMT 10M	NMT 100M
news2017	20.82	25.50	16.32	31.38
news2018	26.91	30.80	20.33	39.01
iwslt2017	18.91	22.70	11.40	26.60
Acquis	19.34	22.35	21.13	27.82
EMEA	27.24	30.86	27.43	35.89
GlobalVoices	20.38	23.49	14.67	28.32
KDE	23.32	26.59	23.68	30.37
average	22.68	26.13	19.77	31.34

Table 4: BLEU evaluation scores (ranking was based on the combination of clusters' scores and sentences' length)

Shared Task. We explored shallow features of sentences and sentence pairs and grouped the task data in 5 clusters according to their presumed usefulness for training MT systems. Our language-pair independent submissions were not based on MT output or bilingual lexica, i.e. on resources which are often scarce or simply not available for many language pairs. Nevertheless, the results obtained from the systems trained on our submissions indicate that this language-pair independent approach yields datasets on which competitive MT systems can be built.

Acknowledgments

This work has been supported by SMART 2015/1091 LOT 3 & LOT 2, a service contract that

implements the acquisition of language resources for the EC's Connecting Europe Facility (CEF) eTranslation platform.

References

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for SMT using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–345, Sofia, Bulgaria. Association for Computational Linguistics.

Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The CMU-Avenue French-English Translation System. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 261–266, Montréal, Canada. Association for Computational Linguistics.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683. Association for Computational Linguistics.

Miquel Esplà-Gomis and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathemathical Linguistics*, 93:77–86.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vassilis Papavassiliou, Prokopis Prokopidis, and Stelios Piperidis. 2018. Discovering parallel language resources for training MT engines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vassilis Papavassiliou, Prokopis Prokopidis, and Gregor Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Aleš Tamchyna, Andy Way, and Josef Genabith. 2014. Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, 49(1):147–193.

Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. 2015. Abu-MaTran at WMT 2015 Translation Task: Morphological Segmentation and Web Crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 184–191, Lisbon, Portugal. Association for Computational Linguistics.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 414–421.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *EMNLP*, pages 2945–2950. Association for Computational Linguistics.

Ieva Zariņa, Pēteris Ņikiforovs, and Raivis Skadiņš. 2015. Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192, Antalya, Turkey.