Word Alignment Based Parallel Corpora Evaluation and Cleaning Using Machine Learning Techniques

Ieva Zariņa
University of Latvia

Pēteris Ņikiforovs
Tilde

Raivis Skadiņš
Tilde

{lu-ieva.zarina,peteris.nikiforovs,raivis.skadins}@tilde.lv

Abstract

This paper presents a method for cleaning and evaluating parallel corpora using word alignments and machine learning algorithms. It is based on the assumption that parallel sentences have many word alignments while non-parallel sentences have few or none. We show that it is possible to build an automatic classifier, which identifies most of non-parallel sentences in a parallel corpus. This method allows us to do (1) automatic quality evaluation of parallel corpus, and (2) automatic parallel corpus cleaning. The method allows us to get cleaner parallel corpora, smaller statistical models, and faster MT training, but this does not always guarantee higher BLEU scores.

An open-source implementation of the tool described in this paper is available from https://github.com/tilde-nlp/c-eval.

1 Introduction

In statistical machine translation, translation quality is largely dependent on the amount of parallel data available. In practice, a large chunk of data considered parallel might not be so, and it can interfere with good data and reduce translation quality.

The problem of low quality parallel corpora is getting more and more important because it is becoming popular to build parallel corpora from web data using fully automatic methods. The quality of such corpora often is very low, especially in case of multilingual corpora, which are built by people who do not know the languages they are working with. As a result, we get corpora with broken encoding, many alignment errors and even texts in different languages.

The problem can be mitigated by removing blatantly obvious non-parallel text that can be detected with handwritten rules. But that does not help in cases where there are alignment errors or two sentences are kind-of parallel but the translation is wrong or incomplete. The cleaning of such parallel text would require human involvement since devising rules for catching such errors would be nearly impossible.

The idea presented in this work is to compare word alignments in a parallel text with those found in a non-parallel text. The intuition being that truly parallel text should have many alignments on word level while unrelated non-parallel text should have few to no alignments.

Since word alignment computation is already a step in the training process of many phrase-based statistical machine translation systems, it can be used as input data for the corpus evaluation and cleaning method that we propose.

Another benefit of cleaning a corpus is a reduced size, which leads to smaller storage and computational costs of statistical machine translation systems.

2 Related Work

This paper is about evaluation and cleaning of parallel corpora, which has been researched from different aspects before. Typically corpus evaluation and cleaning are separate steps in the corpus development process, and corpus development goes through several cycles of evaluation and cleaning while corpus quality reaches acceptable level.

Corpus quality is evaluated by both calculating quantitative measurements and assessing its suitability for the purpose. One of the most important quality aspects of a parallel corpus is sentence alignment quality, which shows how accurately a corpus is broken into sentences and

whether aligned sentences are translations of each other. It is common to use the same metrics for corpus quality evaluation as for sentence alignment evaluation. The sentence alignment evaluation has been well established in ARCADE project/shared task (Langlais et al., 1998), where quality is assessed calculating precision, recall and F-measure both in segment and sub-segment levels. In the same way precision is also used for corpora evaluation. To calculate the precision we need an annotated subset of the corpus where each sentence alignment is marked as correct or not. There are different ways how to get such annotations, Smith et al. (2013), Skadiņš et al. (2014) and Seljan et al. (2010) use a human annotated random subset of corpus, while Kaalep and Veskis (2007) obtain annotations from two different but similar versions of the corpus. Another approach in corpora quality assessment has been used by Steinberger et al. (2012), they tested alignment in a production setting where translators were confronted with the automatically aligned translations and were encouraged to notify any alignment errors.

Although many parallel corpora have been declared to be suitable for different purposes, many of them have not been formally evaluated (Steinberger et al., 2012; Tiedemann, 2012; Callison-Burch, 2009, Chapter 2.2.) and many have been just partially evaluated only for suitability for MT (Koehn, 2005; Eisele & Chen, 2010; Smith et al., 2013; Skadiņš et al., 2014), i.e., authors build MT systems to illustrate that corpus is useful for MT.

Corpus cleaning in practice has often been limited to applying a set of handwritten rules (regular expressions) to detect blatantly obvious cases where two sentences are not parallel (Rueppel et al., 2011; Ruopp, 2010; etc.). More advanced corpora cleaning includes filters that check text language (Lui & Baldwin, 2012) and spelling, and filter out machine translated content (Rarrick et al., 2011). And there are corpora cleaning methods that automatically identifies sentences that are not in conformity with the rest of the corpus; Okita (2009) removes outliers by the literalness score between a pair of sentences, Jiang et al. (2010) introduce lattice score-based data cleaning method, and Taghipour et al. (2011) use density estimators to detect the outliers. These methods allow to identify potentially non-parallel sentences and to filter out sentences with conformity level below a certain threshold; these methods filter out specified amount of data, but they do not estimate how much data should be filtered out. The method proposed in this paper deals with both issues: (1) automatic quality evaluation of parallel corpus and (2) automatic parallel corpus cleaning. Similar word alignment based corpus cleaning method is used by Stymne et al. (2013), but unlike this work they use alignment based heuristics to filter out bad sentence pairs.

3 Proposed Method

3.1 Intuition

Word alignment is a task in natural language processing of identifying translation relationships among the words in a parallel text. It is commonly used in phrase-based statistical machine translation (Koehn et al., 2003) where word alignments are used to extract phrases. One of the commonly used phrase extraction algorithms is to take sequential word alignments in a sentence and expand them as much as possible. The better the word alignments, the better the phrases.

Alignments in a parallel text can be computed with the Expectation Maximization algorithm which means that alignments in a sentence are dependent on similar alignments elsewhere in the corpus. These are called IBM Models 1-5 (Brown et al., 1993).

We can presume that if a corpus is good then there should be many word alignments in sentences. If there are mostly correct sentences in a parallel corpus then the sentences where there are few or no alignments might not be parallel.
While comparing good alignments with bad alignments for large data is a daunting task for a human, it is perfectly suited for machine learning, which we explore in this paper.

The idea is to develop a model with machine learning for classifying a pair of sentences as either parallel or not. As such, it is necessary to train such a model with positive and negative examples. Positive examples can be an approved parallel corpus while negative examples can be generated from a good corpus by shuffling translations or artificially generating bad translations.

For machine learning algorithms to do their job it is necessary to convert text into set of features (numbers), each feature representing a clue for the algorithm how to classify the input data.

3.2 Features

Fast Align word aligner (Dyer et al, 2013) which implements modified IBM Model 2 was used. It

provides us with the alignments and the statistical likelihood of each token-to-token translation. From this data we obtain the features that are used for machine learning.

We generated various probable features. For example, we calculate the Threshold score by dividing the count of alignments that are present in both alignment directions (intersection of alignment count) with the total count of alignments in the respective line (for each language direction). Further features were calculated from the alignment probability scores for each token that are provided by Fast Align in the alignment process.

From the list of probable features the most relevant ones were chosen that provide statistical significance for the machine learning.

We used WEKA (Hall et al., 2009) for 10-fold cross validation with a constant seed to evaluate all the features. Correlation-based Feature Subset Selection for Machine Learning by M. A. Hall (1999) with the best first search method was used to evaluate the significance of all features in the DGT-TM 2007 (Steinberger et al., 2012) English to Latvian corpus of 100,000 correct and 100,000 incorrect lines.

The most significant alignment feature proved to be the fourth dealing with the nth root of the multiplication of the probabilities of *n* tokens (geometric mean). The formulae of the selected features can be seen below (*n* represents the number of tokens in a line).

1) $Threshold\ score = \frac{intersection\ of\ alignment\ count}{total\ line\ alignment\ count}$
2) $\frac{Summed\ prob.\ score}{n}$
3) $lg(\frac{|Summed\ prob.\ score|}{n})$
4) $\sqrt[n]{|Multiplied\ prob.score|}$
5) $lg(\sqrt[n]{|Multiplied\ prob.score|})$

In addition to word alignments, we explored the possibility to enhance the accuracy by including features that are derived from the text itself. For example, the ratio of source sentence token count and target sentence token count, division of common number count and all unique number count in source and target sentences, etc. We calculate features from tokens, numbers, symbols, words and symbols in both source and target sentences – total 43 textual features.

The computation of textual features for a large amount of input data was about two times slower that the computation of alignment features. More importantly, the result quality including textual features together with alignment features increased the precision only by 0.2%. For these reasons, text features were discarded.

3.3 Machine Learning

Once we finalized a list of possible features and selected the most relevant ones, we moved on to the next step of putting them to use with the help of machine learning algorithms.

In order to employ machine learning algorithms and to train a model, we had to provide good (correctly aligned parallel corpora) and bad (aligned corpora with shuffled lines) data. The algorithms then go through each good and bad features and produce a statistical model against which another corpus can be benchmarked.

We evaluated several machine learning algorithms and set out to find those that achieved the highest precision with acceptable performance time as well as a high rate of true positives – an important point when evaluating machine learning algorithms (Flach, 2012).

According to Hill et al. (1998) decision-tree based algorithms would be very suited for working with large data and finding the distinguishing line between data from good and bad corpus. As a result, a data model would be obtained that could be used in filtering each line of a given corpus.

Accuracy as well as training and classification run times of several machine learning algorithms were evaluated on the first 100,000 lines of the DGT-TM 2007 EN-LV corpus. The results are summarized in Table 1.

As can be seen, the algorithms perform rather similarly, though the performance time greatly varies from 15.8 seconds up to 7.5 minutes for a corpus containing 100,000 lines. The REPTree algorithm was chosen because of its high precision paired with relatively good speed.

Algorithm	Precision	Time, s
J48	98.01%	340
J48graft	98.04%	450
RandomForest	98.16%	358
RandomTree	97.43%	58
ExtraTrees	97.17%	26
REPTree	98.03%	130
NaiveBayes	95.72%	16

Table 1. Machine learning algorithm performance comparison for Fast Align features.

4 Evaluation

Firstly, we evaluated the tool by looking at the BLEU score (Papineni et al., 2002) changes, qualitative changes and the quality score of EUBokshop (OPUS edition) corpus, which is known to be cluttered with bad data. It has been automatically extracted from web data (PDF files), containing parallel corpora for 24 official European Union languages (Skadiņš et al., 2014). For testing we chose the Latvian, English and French language pairs.

We evaluated several well-known corpora with the Corpus Cleaner tool as well as whether the results were consistent with qualitative evaluation. The chosen corpora consisted of: EN-FR 10^9 parallel corpus (Callison-Burch, 2009, Chapter 2.2.), EN-DE and EN-FR versions of CommonCrawl (Smith et al., 2013), DGT-TM 2012 (Steinberger et al., 2012), EMEA (Tiedemann, 2012), Europarl (Koehn, 2005), JRC-Aquis (Steinberger et al., 2006), WIT3 (Cettolo et al., 2012).

A number of different models were built and used to test if models were language independent.

4.1 Evaluation in MT

Since the main use for this cleaning method is machine translation, we evaluated how the cleaning method affects the BLEU score.

For the MT evaluation we trained an SMT system with the original EU Bookshop corpus and noted the BLEU score.

We applied the same procedure to the cleaned version of the corpus. Table 2 summarizes the BLEU scores and the amount of good lines after cleaning for the explored language pairs can be seen.

The BLEU score for both the original and cleaned MT systems was nearly identical with the cleaned corpus having a slightly lower BLEU score than the original. However, this does not necessarily mean no improvement.

Generally, in MT systems the less data you have, the less likely you are to have correct translations, and as it has been shown by Goutte et al. (2012), phrase-based SMT is quite robust to noise. Therefore bigger corpus despite containing more corrupt lines is not that detrimental to machine translation since it gets lost in translation anyway.

Language	**BLEU score, baseline**	**BLEU score, cleaned**	***Good* lines**
LV-EN	32.54	32.50	67.19%
LV-FR	24.31	23.47	39.63%

Table 2. BLEU score for original and cleaned EU Bookshop corpora (OPUS), *good* line amount after cleaning.

While the BLEU score nearly did not change for the cleaned corpora, the corpus size, however, did. The cleaned corpora was respectively about 70% and 40% the size of the original. This means that training and memory costs were much lower than the original corpus required. Moreover, the huge difference in cleaned corpus size in comparison with the original producing the same BLEU score indicates that indeed the corrupt lines that the MT system also had deemed unfit were filtered out.

4.2 Qualitative Evaluation

To qualitatively evaluate the cleaning method, we randomly took 200 lines from the original as well as the cleaned corpora for Latvian-English and Latvian-French language pairs. We manually evaluated them for incorrect or erroneous alignment. The results are shown in Table 3. The manual evaluation was done by one evaluator.

	LV-EN	LV-FR
Sentences from the original corpus that were classified as *good* by the human evaluator	78%	72%
Sentences that were classified as *good* by the human evaluator from sentences that were classified as *good* by the corpus cleaner.	90%	95%
Sentences that were classified as *good* by the human evaluator from sentences that were classified as *bad* by the corpus cleaner.	11%	10%

Table 3. The amount of good lines in EU Bookshop corpora

The qualitative results clearly show the improvement in corpus quality. Taking into account that the size of corpora was approximately 30% smaller after cleaning and performance rate of about 90%, it can be concluded that a significant part of bad data was removed.

4.3 Corpora Evaluation with Different Models

As a part of the corpora cleaning process, we implemented a corpus evaluation solution. The percentage score of a corpus shows the amount of good lines in the text.

As models for cleaning could be constructed from any corpora that is recognized of good quality, we set to determine if the models are language independent. That is, if different models (made from approximately equal quality corpora) would produce the same results for a given parallel corpus.

The models were trained on the DGT-TM 2007 corpus consisting of EN-LV, EN-FR, EN-LT, and FR-LV language pairs. The graph lines represent the score of each corpus using the corresponding model (along the X axis). Models themselves were evaluated using WEKA tool. The results are shown below in Figure 1.

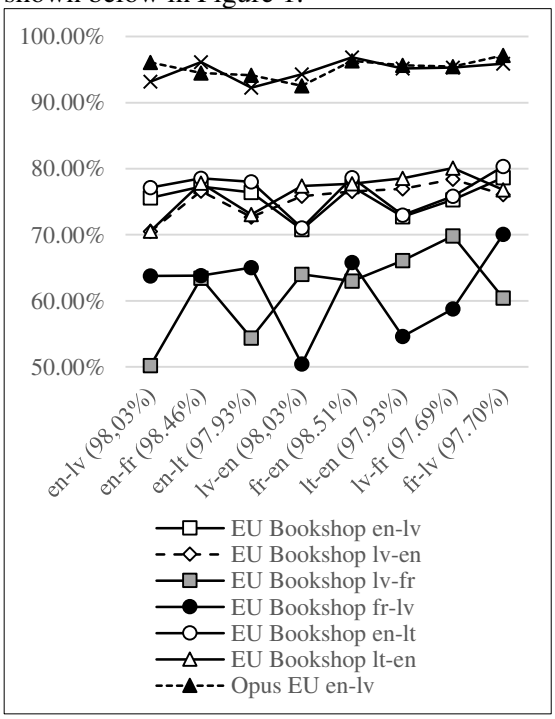

Figure 1. Corpora evaluation with different models.

The results show, overall, that the lower the quality of corpus, the more varied the cleaning results from different models will be.

It can be concluded that while there is a difference in the performance of the models (worst case up to 20%), it evens out with the increase of the quality of the corpora (approx. 5% variation). To sum up, for precise corpus evaluation, it would be best to use a model that has been built for the particular language pair.

To see how the method fares with already good data, we evaluated the DGT-TM English-Lithuanian corpus with the DGT-TM English-Latvian model as well as the DGT-TM French-English corpus with the DGT-TM Latvian-English model. It removed approximately 3% of good sentences, which we think is acceptable. Similarly OPUS EU Constitution corpus, which is considered fairly accurate, saw about 5% cut and showed considerably more stable results across all models than EU Bookshop corpora signaling reliable performance in case of high quality corpora.

4.4 Evaluated Corpora Comparison

Initially we started our evaluations using well known good quality corpora. As can be seen in Table 4, all of the evaluated corpora are of high quality (around 98%) corresponding with previous evaluations and qualitative evaluations of 100 sentences randomly taken from the English to Latvian language pair. The quality of the above corpora was measured with corresponding models built from the first 100,000 lines of the DGT-TM-2007 corpus.

	DGT-TM 2012	EMEA	Europarl	JRC Acquis	WIT[3]
EN-DE	98.91%	95.54%	99.01%	99.30%	97.65%
EN-ES	98.24%	96.74%	99.36%	99.18%	98.46%
EN-FR	98.84%	96.39%	99.58%	98.89%	99.30%
EN-IT	98.01%	95.65%	98.94%	99.02%	97.74%
EN-LV	97.75%	94.26%	99.67%	98.36%	98.34%
EN-LV QE	99%	91%	99%	98%	97%

Table 4. Corpora quality evaluation by Corpus Cleaner and qualitative evaluation (QE)

We also evaluated less credible corpora (See Table 5). Significant differences can be seen between EUBookshop Tilde and OPUS editions with approximately 20% increase in quality. This result is understandable as Tilde has considerably improved the quality of EUBookshop by filtering and manually editing it (Skadiņš et al., 2014).

In order to compare the results of the CommonCrawl EN-DE corpus quality with the work done by Stymne et al. (2013), it was additionally cleaned by removing sentence pairs with larger than three ratio, sentences with more than 60 tokens as well as the corpus was lowercased. This reduced the corpus by 4.28%. Consequently filtering the original CommonCrawl reduced the amount by 16%, while 13% was removed from the cleaned version of the CommonCrawl corpus.

Corpus	Language pair	Corpus Cleaner Quality	QE
EN-FR 10[9]	EN-FR	84.20%	89%
CommonCrawl	EN-FR	80.02%	70%
CommonCrawl (original)	EN-DE	83.94%	55%
CommonCrawl (filtered)	EN-DE	87.25%	59%
EUBookshop (TILDE)	EN-LV	96.19%	93%
	EN-FR		77%
EUBookshop (OPUS)	EN-LV	76.45%	67%
	FR-LV	71.52%	73%

Table 5. Corpora quality evaluation by Corpus Cleaner and qualitative evaluation (QE)

Stymne's et al. research shows a considerably larger corpus reduction (27%) based on alignment evaluation, 5.3% reduction by cleaning the text and in addition 8.8% by removing sentences with wrong detected language. The approach taken by Stymne et al. looks at a manually annotated gold corpus of 100 lines, and extrapolates from that good calculated values from alignment intersection against sentence length, similarly as Threshold score described previously. This manual method generates more strict results and consequently marks more lines as bad. However, the qualitative evaluation of CommonCrawl both original and cleaned versions correspond to that in Stymne's et al. work signaling that the used methods should be looked into more thoroughly.

Language detection as employed by Stymne et al. produced high quality results. While, wrong language use shows up in the alignment quality up to a certain level producing a small intersection set, it could, nevertheless, be considered as an additional feature in the corpus cleaner tool.

English-French10[9] and CommonCrawl EN-FR corpora show a moderate level of accuracy as well as the qualitative evaluation confirms this result deviating by 5% and 10% respectively.

5 Conclusion and Future Work

We have shown that by using word alignment features we can build an automatic classifier, which identifies most non-parallel sentences in a parallel corpus. This method allows us to do (1) automatic quality evaluation of a parallel corpus, and (2) automatic parallel corpus cleaning. The method allows us to get cleaner parallel corpora, smaller statistical models, and faster MT training, but unfortunately this does not always guarantee higher BLEU scores.

In this paper, we are reporting our first results. It is still necessary, however, to test the method for a much wider range of languages and corpora to verify that the method is applicable for other language pairs and to see whether the automatic corpora quality evaluation correlates with human judgment.

We used Fast Align, which is based on IBM Model 2; but IBM Model 1, which requires less computation power, may prove just as effective. Similarly, it would be useful to evaluate higher IBM Models to see how much the results are improved at the cost of longer running time.

We discarded text features for use as the input data for the classifier, but that does not mean that they are not useful. They might as well be used with handwritten rules as an additional step in the cleaning pipeline, either before this method is applied or afterwards. We are planning to revise textual features. In this research, we focused on identifying alignment errors, but textual features can be useful to identify broken encoding, texts in wrong language and other corpora quality issues.

More consistent results across language models could be achieved improving bad training data generation. It is possible that during the shuffling process some lines are aligned in a way that produces a somewhat valid translation, therefore yielding inconsistent data for the machine-learning algorithm.

Acknowledgements

The research leading to these results has received funding from the research project "Optimization methods of large scale statistical models for innovative machine translation technologies", project financed by The State Education Development Agency (Latvia) and European Regional Development Fund, contract nr. 2013/0038/2DP/2.1.1.1.0/13/APIA/VIAA/029.
We would like to thank Valdis Girgždis and Maija Kāle for their contribution to this research.

References

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

Callison-Burch, C., Koehn, P., Monz, C., & Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation (pp. 1–28). Athens, Greece: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W09/W09-0401

Cettolo, M., Girardi, C., & Federico, M. (2012, May). WIT3: Web inventory of transcribed and translated talks. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT) (pp. 261-268).

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In HLT-NAACL (pp. 644-648).

Eisele, A., & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, B. Maegaard, … N. C. (Conference Chair) (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (pp. 2868–2872). European Language Resources Association (ELRA).

Flach, P. (2012). The Art and Science of Algorithms that Make Sense of Data (pp. 55). New York, USA: Cambridge University Press.

Goutte, C., Carpuat, M., & Foster, G. (2012). The impact of sentence alignment errors on phrase-based machine translation performance. In Conference of the Association for Machine Translation in the Americas (AMTA). San Diego, CA.

Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl., 11(1), 10–18. doi:10.1145/1656274.1656278

Hill, L. O., Chawla, N., Bowyer, K. W. (1998) Decision Tree Learning on Very Large Data Sets. Department of Computer Science and Engineering, University of South Florida. Retrieved from https://www3.nd.edu/~dial/papers/SMC98.pdf

Jiang, J., Way, A., & Carson-Berndsen, J. (2010). Lattice Score Based Data Cleaning For Phrase-Based Statistical Machine Translation.

Kaalep, H. J., & Veskis, K. (2007). Comparing parallel corpora and evaluating their quality. Proceedings of MT Summit XI, 275-279.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical phrase based translation. Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL).

Koehn, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. MT Summit, 11, 79–86. Retrieved from http://mt-archive.info/MTS-2005-Koehn.pdf

Langlais, P., Simard, M., Veronis, J., Armstrong, S., Bonhomme, P., Debili, F., ... & Theron, P. (1998). Arcade: A cooperative research project on parallel text alignment evaluation.

Lui, M., & Baldwin, T. (2012). Langid.Py: An Off-the-shelf Language Identification Tool. In Proceedings of the ACL 2012 System Demonstrations (pp. 25–30). Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Okita, T. (2009). Data Cleaning for Word Alignment. In Proceedings of the ACL-IJCNLP 2009 Student Research Workshop (pp. 72–80). Stroudsburg, PA, USA: Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics.: ACL

Rarrick, S., Quirk, C., & Lewis, W. (2011). MT Detection in Web-Scraped Parallel Corpora. In Proceedings of MT Summit XIII. Asia-Pacific Association for Machine Translation.

Rueppel, J., Jiang, L., Yu, G., and Flournoy, R. (2011). AIR-based light clients for supporting Moses engine training. In Proceedings of the 13th Machine Translation Summit (pp. 503–506). Xiamen.

Ruopp, A. (2010). How to implement open source machine translation solutions (TAUS report): TAUS BV.

Seljan, S., Tadić, M., Agić, Ž., Šnajder, J., Bašić, B. D., & Osmann, V. (2010). Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora. In N. C. (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA).

Skadiņš, R., Tiedemann, J., Rozis, R., & Deksne, D. (2014). Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. In N. C. (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 1850–1855). Reykjavik, Iceland: European Language Resources Association (ELRA).

Smith, R. J., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., & Lopez, A. (2013). Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1374–1383). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/P13-1135

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), (pp. 24-26). Genoa, Italy

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely available Translation Memory in 22 languages. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA).

Stymne, S., Hardmeier, C., Tiedemann, J., & Nivre, J. (2013). Tunable distortion limits and corpus cleaning for SMT. In WMT 2013; 8-9 August; Sofia, Bulgaria (pp. 225-231). Association for Computational Linguistics.

Taghipour, K., Khadivi, S., & Xu, J. (2011). Parallel Corpus Refinement as an Outlier Detection Algorithm. MT Summit XIII. Machine Translation Summit (MT-Summit-11), 13. September 19-23, Xiamen, China. NA, Xiamen.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA).