Augmenting SMT with Generated Pseudo-parallel Corpora from Monolingual News Resources

Krzysztof Wołk(✉) and Agnieszka Wołk

Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
{kwolk,awolk}@pja.edu.pl

Abstract. Several natural languages have had much processing, but the problem of limited linguistic resources remains. Manual creation of parallel corpora by humans is rather expensive and very time consuming. In addition, language data required for statistical machine translation (SMT) does not exist in adequate capacity to use its statistical information to initiate the research process. On the other hand, applying unsubstantiated approaches to build the parallel resources from multiple means like comparable corpora or quasi-comparable corpora is very complicated and provides rather noisy output. These outputs of the process would later need to be reprocessed, and in-domain adaptations would also be required. To optimize the performance of these algorithms, it is essential to use a quality parallel corpus for training of the end-to-end procedure. In the present research, we have developed a methodology to generate an accurate parallel corpus from monolingual resources through the calculation of compatibility between the results of machine translation systems. We have translations of huge, single-language resources through the application of multiple translation systems and the strict measurement of translation compatibility with rules based on the Levenshtein distance. The results produced by such an approach are very favorable. All the monolingual resources that we obtained were taken from the WMT16 conference for Czech to generate the parallel corpus, which improved translation performance.

Keywords: Parallel corpora · Corpora preparation · Generating corpora · Data mining parallel corpora

1 Introduction

Statistical machine translation (SMT) is a methodology based on statistical data analysis. Better performance of SMT systems largely depends on the quantity and quality of the parallel data it uses. If the quantity and quality of the parallel data is high, this will boost the SMT results. Even so, parallel corpus is still scarce and not easily available. Similarly, the genre and language coverage of the data should also be very limited to increase SMT performance. In particular, languages that have very few native speakers and thus offer a limited audience will lead to very little research in the field. This creates a technical gap between languages that are widely spoken in comparison to languages with

Á. Rocha et al. (eds.), *Recent Advances in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 569, DOI 10.1007/978-3-319-56535-4_31

few speakers. However, the majority of existing human languages are spoken by only a small population of native speakers.

As a result, high-quality data exists only for a few language pairs in particular domains, whereas the majority of languages lack sufficient linguistic resources such as parallel data. Building a translation system that can cover all possible language translations would require millions of translation directions and a huge amount of parallel corpora. Moreover, if we consider multiple domains in the equation, the requirements for corpus training increase dramatically. The current study explored methods to build high-quality parallel data.

Multiple studies have been performed to automatically acquire additional data to enhance SMT systems in the long term [1, 2]. All such approaches have focused on discovering the actual text for the source languages as well as the target languages. However, our study presents an alternative approach for building the parallel data. In creating virtual parallel data, as we might call it, at least one side of the parallel data is generated. For this purpose, we use monolingual text. For the other side of the parallel data, we use an automatic procedure to obtain the translation of the text. In other words, our approach generates parallel data rather than gathering it. To monitor the performance and quality of the automatically generated parallel data and to maximize its utility for SMT, we focus on compatibility between the diverse layers of an SMT system.

In classification, it is recommended that an estimate be considered reliable when multiple systems show a consensus on it. However, the output of machine translation (MT) is human language, for which it is much too complicated to seek unanimity from multiple systems to generate the same output each time we execute the translation process. In such situations, we can choose partial compatibility as an objective rather than the complete agreement of multiple systems. To evaluate the generated data, we can use the Levenshtein distance and also run through a backward translation procedure. Using this approach, only those pairs that pass an initial compatibility check, being translated back into the native language and compared to the original sentences, will be accepted. This concept is depicted in Fig. 1.

We can use this method to easily generate additional parallel data from monolingual news data provided for WMT16.[1] Retraining the newly assessed data during this procedure enhances translation system performance. Moreover, linguistic resource pairs that are rare can be improved. This methodology is not limited to languages but is also very significant for rare and important language pair resources. Most significantly, the virtual parallel corpus generated by the system is applicable to MT as well as other natural language processing (NLP) tasks.

2 State of the Art

In this study, we present an approach based on generating comprehensive multilingual resources through SMT systems. At the present time, we are working on two approaches for MT applications: self-training and translation via bridge languages. These

[1] http://www.statmt.org/wmt16/translation-task.html.

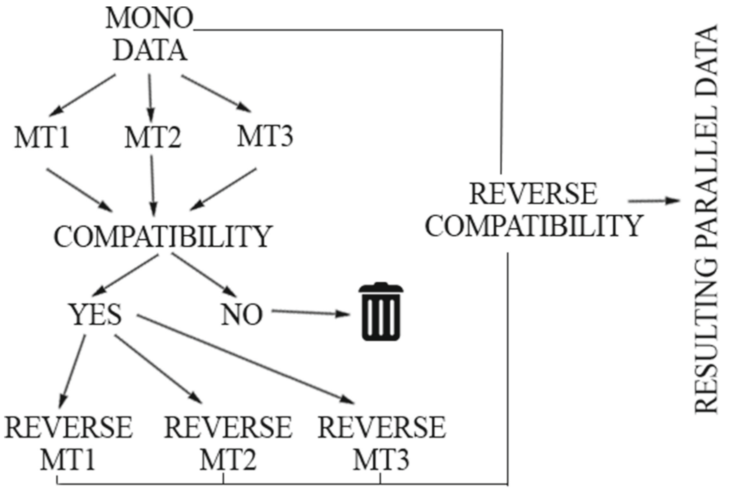

Fig. 1. Corpora generation scheme

approaches are different from what we have discussed previously. The first is focused on exploiting available bilingual data, where the linguistic resources from another language are rarely applied. The second approach focuses more on correcting the alignment of the prevailing word segment. In addition, it incorporates the phrase model concept rather than exploring the new text in context, such as translations at the word, phrase, or even the sentence level, through bridge languages. The methodology of this paper lies between the paradigm of self-training and translating via a bridge language. Our study generates data instead of gathering information for parallel data. Moreover, we have applied linguistic information and relationships between the languages to eventually perform translations between source and target languages.

Callison-Burch and Osborne in [3] presented a cooperative training method for SMT that is the consensus of several translation systems to identify the best translation resource for training. Similarly, Ueffing et al. [4] explored model adaptation methods for using monolingual data from a source language. Furthermore, as the learning progressed, application of that learned material was constrained by a multi-linguistic approach without introducing new information from a third language.

In another approach, Mann and Yarowsky [5] presented a technique to develop a translation lexicon based on transduction models of cognate pairs through a bridge language. The edit distance rate is applied to the process rather than the general MT system to limit the range of vocabulary for majority European languages. Kumar et al. [6] described the process to boost word alignment quality by using multiple bridge languages. In [7] and [8], phrase translation tables are improved through the use of phrase tables acquired in multiple ways from pivot languages. In [9], a hybrid method is combined with RBMT and SMT systems. This methodology is introduced to fill gaps in the data for pivot translation. Cohn and Lapata [10] presented another methodology to generate more reliable results of translations by generating information from small sets of data using multi-parallel data.

Contrary to the existing approaches, we returned to the black-box translation system. This means a wide generation of virtual data can be performed on translation systems that include rule-based, statistics-based, and also those based on human translations. The approach introduced in [11] pooled the results of translations of a test set made by any of the pivot MTs per unique language. However, this approach did not enhance the systems, and hence the novel training data is not used. Along with others, Bertoldi et al. [12] also conducted research on pivot languages, but they did not consider the application of universal corpus filtering, which is the measurement of compatibility to control data quality.

3 Generating Virtual Parallel Data

To generate new data, we have trained three SMT systems. The Experiment Management System [13] from the open source Moses SMT toolkit was utilized to carry out the experimentation. A 6-gram language model was trained using the SRI Language Modeling toolkit (SRILM) [14]. Word and phrase alignment was performed using the SyMGIZA ++ symmetric word alignment tool [15] instead of GIZA ++. Out-of-vocabulary (OOV) words were monitored using the Unsupervised Transliteration Model [16]. While working with the Czech (CS) and English (EN) language pair, a first SMT system was trained on TED [17], a second on the Qatar Computing Research Institute's Educational Domain Corpus (QED) [18], and a third using the News Commentary corpora provided for the WMT16[2] translation task. Official WMT16 test sets were used for system evaluation. Translation engine performance was measured by the BLEU metric [25]. The performance of the engines is shown in Table 1.

Table 1. Corpora used for generation of SMT systems

Corpus	Direction	BLEU
TED	CS- > EN	16.17
TED	EN- > CS	10.11
QED	CS- > EN	23.64
QED	EN- > CS	21.43
News Commentary	CS- > EN	14.47
News Commentary	EN- > CS	9.87

All engines worked in accordance with Fig. 1, and the Levenshtein distance was used to measure the compatibility between translation results. The Levenshtein distance measures the diversity between two strings. Moreover, it also indicates the edit distance. It is closely linked to paired arrangement of strings [26].

Mathematically, the Levenshtein distance between two strings a, b [of length $|a|$ and $|b|$, respectively] is given by $\mathrm{lev}_{a,b}[|a|, |b|]$ where:

[2] http://www.statmt.org/wmt16/.

Table 2. Specification of generated corpora

Data set	Number of sentences		Number of Unique Czech Tokens	
	Monolingual	Generated	Monolingual	Generated
News 2007	100,766	83,440	200,830	42,954
News 2008	4,292,298	497,588	2,214,356	168,935
News 2009	4,432,383	527,865	2,172,580	232,846
News 2010	2,447,681	269,065	1,487,500	100,457
News 2011	8,746,448	895,247	2,871,190	298,476
News 2012	7,538,499	849,469	2,589,424	303,987
News 2013	8,886,151	993,576	2,768,010	354,278
News 2014	8,722,306	962,674	2,814,742	322,765
News 2015	8,234,140	830,987	2,624,473	300,456
TOTAL	53,366,020	5,944,583	8,765,548	2,125,154

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if \min(i,j) = 0 \\ min\begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ \qquad lev_{a,b}(i-1,j-1)+1_{[a_i \neq b_j]} \end{cases} & otherwise. \end{cases}$$

In this equation, $1_{[a_i \neq b_j]}$ is the display function, equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $lev_{a,b}[i,j]$ is the distance between the first i characters of a and the first j characters of b.

Using the combined methodology and monolingual data, parallel corpora were built. Statistical information on the data is provided in Table 2.

The purpose of this research was to create synthetic parallel data for training a machine translation system by translating monolingual texts with multiple machine translation systems and various filtering steps. This objective is not new; synthetic data has previously been created. The novel aspect of the present paper is to use three MT systems and apply the Levenshtein distance between their outputs as a filter and—much more importantly—use back-translation as an additional filtering step. In Table 3, we show statistical information on the corpora used without the backward translation step.

Table 3. Specification of generated corpora without backward translation

Data set	Number of sentences		Number of Unique Czech Tokens	
	Monolingual	Generated	Monolingual	Generated
News 2007	100,766	93,342	200,830	120,654
News 2008	4,292,298	1,654,233	2,214,356	1,098,432
News 2009	4,432,383	1,423,634	2,172,580	1,197,765
News 2010	2,447,681	1,176,022	1,487,500	876,654
News 2011	8,746,448	2,576,253	2,871,190	1,378,456
News 2012	7,538,499	2,365,234	2,589,424	1,297,986
News 2013	8,886,151	2,375,857	2,768,010	1,124,278
News 2014	8,722,306	1,992,876	2,814,742	1,682,673
News 2015	8,234,140	2,234,987	2,624,473	1,676,343
TOTAL	53,366,020	15,892,438	8,765,548	2,975,154

4 Experimental Setup

The machine translation experiments we conducted involve three WMT16 tasks: news translation, information technology (IT) document translation, and biomedical text translation. Our experiments were conducted on the CS-EN pair in both directions. To obtain more accurate word alignment, we have used the SyMGiza ++ tool. This tool assists in the formation of a similar word alignment model. This particular tool develops alignment models that obtain multiple many-to-one and one-to-many alignments in multiple directions between the given language pairs. SyMGiza ++ is also used to create a pool of several processors, supported by the newest threading management, which makes it a very fast process. The alignment process used in our case utilizes four unique models during the training of the system to achieve refined and enhanced alignment outcomes. The results of these approaches have been fruitful [15]. OOV words are another challenge for an SMT system. To deal with OOV words, we used the Moses toolkit and the Unsupervised Transliteration Model (UTM). The UTM is a language-independent approach that has an unsubstantiated capability for learning OOV words. We also utilized the post-decoding transliteration method from this particular toolkit. UTM is known to make use of a transliteration phrase translation table to access the probable solutions. UTM was used to score several possible transliterations and to find a translation table [16, 19].

The KenLM tool was applied to language model training. This library helps to resolve typical problems of language models, reducing execution time and memory usage. To reorder the phrase probability, the lexical values of the sentences were used. We also used KenLM for lexical reordering. Three directional types are based on both targets: swap (S), monotone (M), and discontinuous (D). All three models were used in a hierarchical model. The bidirectional restructuring model examines the phrase arrangement probabilities [20–22].

The quality of domain adaptation largely depends on training data, which helps in incorporating the linguistic and translation models. The acquisition of domain-centric data helps much in this regard [23]. A parallel, generalized domain corpus and a monolingual corpus were used in this process, as identified by Wang et al. [24]. First, sentence pairs of the parallel data were weighted based on their significance to the targeted domain. Second, reorganization was conducted to get the best sentence pairs. After obtaining the required sentence pairs, these models were trained for the target domain [24].

For similarity measurement, we used three approaches: word overlap analysis, the cosine term frequency-inverse document frequency (tf-idf) criterion, and perplexity measurement. However, the third approach, which incorporates the best of the first two, is the strictest one. Wang et al. [24] observed that a combination of these approaches provides the best possible solution for domain adaptation for Chinese-English corpora [24]. Inspired by Wang et al.'s approach [24], we utilized a combination of these models. Similarly, the three measurements were combined for domain adaptation. Wang et al. found that the performance of this process yields around 20 percent of the domain analogous data [24].

5 MT Results and Conclusions

Numerous human languages are used around the world. Millions of translation systems have been introduced for the possible language pairs. These translation systems struggle largely due to the limited availability of language resources such as parallel data.

We have attempted to supplement these limited resources. Additional parallel corpora can be utilized to improve the quality and performance of linguistic resources, as well as individual NLP systems. In the MT application (Table 3), our data generation approach has increased translation performance. Although the results appear very promising, there is still a lot of room for improvement. Performance improvements can be attained by the application of more sophisticated algorithms to quantify the comparison among different MT engines. In Table 4, we present the baseline (BASE) outcomes for the MT systems we obtained for three diverse domains (news, IT, and biomedical—using official WMT16 test sets). Second, we generated a virtual corpus and adapted it to the domain (FINAL). The generated corpora demonstrate improvements in SMT quality and utility as NLP resources. From Table 2, it can be concluded that a generated virtual corpus is morphologically rich, which makes it acceptable as a linguistic resource. In addition, by retraining with a virtual corpus SMT system and repeating all the steps, it is possible to obtain more virtual data of higher quality.

Table 4. Evaluation of generated corpora

Domain	Direction	System	BLEU
News	CS- > EN	BASE	15.26
	CS- > EN	FINAL	18.11
	EN- > CS	BASE	11.64
	EN- > CS	FINAL	13.43
IT	CS- > EN	BASE	12.86
	CS- > EN	FINAL	14.12
	EN- > CS	BASE	10.19
	EN- > CS	FINAL	11.87
Bio-Medical	CS- > EN	BASE	16.75
	CS- > EN	FINAL	18.33
	EN- > CS	BASE	14.25
	EN- > CS	FINAL	15.93

Lastly, in Table 5 we replicate the same quality experiment but using generated data without the backward translation step. Even as shown in Table 3, more data can be obtained in such a manner. However, the SMT results are not as good as the ones obtained with the backward translation step. This means that the generated data must be noisy and most likely contain incomplete sentences that are removed after backward translation.

Table 5. Evaluation of corpora generated without backward translation step

Domain	Direction	System	BLEU
News	CS- > EN	BASE	15.26
	CS- > EN	FINAL	17.32
	EN- > CS	BASE	11.64
	EN- > CS	FINAL	12.73
IT	CS- > EN	BASE	12.86
	CS- > EN	FINAL	13.52
	EN- > CS	BASE	10.19
	EN- > CS	FINAL	10.74
Bio-Medical	CS- > EN	BASE	16.75
	CS- > EN	FINAL	11.32
	EN- > CS	BASE	14.25
	EN- > CS	FINAL	15.03

Acknowledgements. Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and was backed by the PJATK legal resources.

References

1. Munteanu, D.S., Fraser, A.M., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In: HLT-NAACL, pp. 265–272 (2004)
2. Smith, J.R., Quirk, C., Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 403–411. Association for Computational Linguistics (2010)
3. Callison-Burch, C., Osborne, M.: Co-training for statistical machine translation. Ph.D. Thesis. Master's thesis, School of Informatics, University of Edinburgh (2002)
4. Ueffing, N., Haffari, G., Sarkar, A.: Semisupervised learning for machine translation. In: Learning Machine Translation, Pittsburgh, Pennsylvania, pp. 237–256. The MIT Press, February 2009
5. Mann, G.S., Yarowsky, D.: Multipath translation lexicon induction via bridge languages. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, pp. 1–8. Association for Computational Linguistics (2001)
6. Kumar, S., Och, F.J., Macherey, W.: Improving word alignment with bridge languages. In: EMNLP-CoNLL, pp. 42–50 (2007)
7. Wu, H., Wang, H.: Pivot language approach for phrase-based statistical machine translation. Mach. Transl. **21**(3), 165–181 (2007)
8. Habash, N., Hu, J.: Improving arabic-chinese statistical machine translation using english as pivot language. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 173–181. Association for Computational Linguistics (2009)
9. Eisele, A., et al.: Hybrid machine translation architectures within and beyond the EuroMatrix project. In: Proceedings of the 12th Annual Conference of the European Association for Machine Translation (EAMT), pp. 27–34 (2008)

10. Cohn, T., Lapata, M.: Machine translation by triangulation: making effective use of multi-parallel corpora. In: Annual Meeting, p. 728. Association for Computational Linguistics (2007)
11. Leusch, G., et al.: Multi-pivot translation by system combination. In: IWSLT, pp. 299–306 (2010)
12. Bertoldi, N., et al.: Phrase-based statistical machine translation with pivot languages. In: IWSLT, pp. 143–149 (2008)
13. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
14. Stolcke, A., et al.: SRILM-an extensible language modeling toolkit. In: INTERSPEECH, p. 2002 (2002)
15. Junczys-Dowmunt, M., Szał, A.: SyMGiza ++: symmetrized word alignment models for statistical machine translation. In: Bouvry, P., Kłopotek, Mieczysław A., Leprévost, F., Marciniak, M., Mykowiecka, A., Rybiński, H. (eds.) SIIS 2011. LNCS, vol. 7053, pp. 379–390. Springer, Heidelberg (2012). doi:10.1007/978-3-642-25261-7_30
16. Durrani, N., et al.: Integrating an unsupervised transliteration model into statistical machine translation. In: EACL, pp. 148–153 (2014)
17. Cettolo, M., Girardi, C., Federico, M.: 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 (2012)
18. Abdelali, A., et al.: The AMARA corpus: building parallel language resources for the educational domain. In: LREC, pp. 1044–1054 (2014)
19. Moses statistical machine translation, ‘‘OOVs’’ Last revised 13 Feb 2015. http://www.statmt.org/moses/?n=Advanced.OOVs#ntoc2. Accessed 27 Sep 2015
20. Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187–197. Association for Computational Linguistics (2011)
21. Ruiz costa-jussà, M., Rodríguez fonollosa, J.A.: Using linear interpolation and weighted reordering hypotheses in the moses system. In: Seventh Conference on International Language Resources and Evaluation, pp. 1712–1718 (2011)
22. Moses statistical machine translation, ‘‘Build reordering model.” Last revised 28 Jul 2013. http://www.statmt.org/moses/?n=FactoredTraining.BuildReorderingModel. Accessed 10 Oct 2015
23. Amittai, A., He, X., Gao, J.: Domain adaptation via pseudo in-domain data selection In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, p. 362. Association for Computational Linguistics (2011)
24. Wang, L., et al.: A systematic comparison of data selection criteria for SMT domain adaptation. Sci. World J. (2014). https://www.hindawi.com/journals/tswj/2014/745485/
25. Papineni, K., et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
26. Yujian, L., Bo, L.: A normalized levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1091–1095 (2007)