Augmenting SMT with Semantically-Generated Virtual-Parallel Corpora from Monolingual Texts

Krzysztof Wołk(✉) and Agnieszka Wołk

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

Abstract. Several natural languages have undergone a great deal of processing, but the problem of limited textual linguistic resources remains. The manual creation of parallel corpora by humans is rather expensive and time consuming, while the language data required for statistical machine translation (SMT) do not exist in adequate quantities for their statistical information to be used to initiate the research process. On the other hand, applying known approaches to build parallel resources from multiple sources, such as comparable or quasi-comparable corpora, is very complicated and provides rather noisy output, which later needs to be further processed and requires in-domain adaptation. To optimize the performance of comparable corpora mining algorithms, it is essential to use a quality parallel corpus for training of a good data classifier. In this research, we have developed a methodology for generating an accurate parallel corpus (Czech-English) from monolingual resources by calculating the compatibility between the results of three machine translation systems. We have created translations of large, single-language resources by applying multiple translation systems and strictly measuring translation compatibility using rules based on the Levenshtein distance. The results produced by this approach were very favorable. The generated corpora successfully improved the quality of SMT systems and seem to be useful for many other natural language processing tasks.

Keywords: Data filtration · Corpora building · Machine learning
Data mining · Parallel corpora · Machine translation

1 Introduction

Statistical machine translation (SMT) is a methodology based on statistical data analysis. The performance quality of SMT systems largely depends on the quantity and quality of the parallel data used by these systems; that is, if the quantity and quality of the parallel data are high, this will boost the SMT results. Even so, good quality parallel corpora, without noisy data or error free, remain scarce and are not easily available [1]. Moreover, in order to increase SMT performance, the genre and language coverage of the data should be limited to a specific text domain e.g. law or medical texts. In particular, little research has been conducted on languages with few native speakers and

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 745, pp. 358–374, 2018.
https://doi.org/10.1007/978-3-319-77703-0_37

Table 1. Top languages by population: asterisks mark the 2010 estimates for the top dozen languages

Rank	Language	Native speakers in millions 2007 (2010)	Fraction of world population (2007)	Rank	Language	Native speakers in millions 2007 (2010)	Fraction of world population (2007)
1	Mandarin (entire branch)	935 (955)	14.1%	51	Igbo	24	0.36%
2	Spanish	390 (405)	5.85%	52	Azerbaijani	23	0.34%
3	English	365 (360)	5.52%	53	Awadhi	22 [4]	0.33%
4	Hindi [Note 1]	295 (310)	4.46%	54	Gan Chinese	22	0.33%
5	Arabic	280 (295)	4.23%	55	Cebuano (Visayan)	21	0.32%
6	Portuguese	205 (215)	3.08%	56	Dutch	21	0.32%
7	Bengali (Bangla)	200 (205)	3.05%	57	Kurdish	21	0.31%
8	Russian	160 (155)	2.42%	58	Serbo-Croatian	19	0.28%
9	Japanese	125 (125)	1.92%	59	Malagasy	18	0.28%
10	Punjabi	95 (100)	1.44%	60	Saraiki	17 [5]	0.26%
11	German	92 (95)	1.39%	61	Nepali	17	0.25%
12	Javanese	82	1.25%	62	Sinhalese	16	0.25%
13	Wu (inc. Shanghainese)	80	1.20%	63	Chittagonian	16	0.24%
14	Malay (inc. Malaysian and Indonesian)	77	1.16%	64	Zhuang	16	0.24%
15	Telugu	76	1.15%	65	Khmer	16	0.24%
16	Vietnamese	76	1.14%	66	Turkmen	16	0.24%
17	Korean	76	1.14%	67	Assamese	15	0.23%
18	French	75	1.12%	68	Madurese	15	0.23%
19	Marathi	73	1.10%	69	Somali	15	0.22%
20	Tamil	70	1.06%	70	Marwari	14 [4]	0.21%
21	Urdu	66	0.99%	71	Magahi	14 [4]	0.21%
22	Turkish	63	0.95%	72	Haryanvi	14 [4]	0.21%
23	Italian	59	0.90%	73	Hungarian	13	0.19%
24	Yue (incl. Cantonese)	59	0.89%	74	Chhattisgarhi	12 [4]	0.19%
25	Thai (excl. Lao)	56	0.85%	75	Greek	12	0.18%
26	Gujarati	49	0.74%	76	Chewa	12	0.17%
27	Jin	48	0.72%	77	Deccan	11	0.17%
28	Southern Min (incl. Fujianese/Hokkien)	47	0.71%	78	Akan	11	0.17%
29	Persian	45	0.68%	79	Kazakh	11	0.17%
30	Polish	40	0.61%	80	Northern Min	10.9	0.16%
31	Pashto	39	0.58%	81	Sylheti	10.7	0.16%
32	Kannada	38	0.58%	82	Zulu	10.4	0.16%
33	Xiang (Hunnanese)	38	0.58%	83	Czech	10.0	0.15%
34	Malayalam	38	0.57%	84	Kinyarwanda	9.8	0.15%
35	Sundanese	38	0.57%	85	Dhundhari	9.6 [4]	0.15%
36	Hausa	34	0.52%	86	Haitian Creole	9.6	0.15%
37	Odia (Oriya)	33	0.50%	87	Eastern Min	9.5	0.14%
38	Burmese	33	0.50%	88	Ilocano	9.1	0.14%
39	Hakka	31	0.46%	89	Quechua	8.9	0.13%

(*continued*)

Table 1. (*continued*)

Rank	Language	Native speakers in millions 2007 (2010)	Fraction of world population (2007)	Rank	Language	Native speakers in millions 2007 (2010)	Fraction of world population (2007)
40	Ukrainian	30	0.46%	90	Kirundi	8.8	0.13%
41	Bhojpuri	29 [4]	0.43%	91	Swedish	8.7	0.13%
42	Tagalog/Filipino	28	0.42%	92	Hmong	8.4	0.13%
43	Yoruba	28	0.42%	93	Shona	8.3	0.13%
44	Maithili	27 [4]	0.41%	94	Uyghur	8.2	0.12%
45	Uzbek	26	0.39%	95	Hiligaynon/Ilonggo (Visayan)	8.2	0.12%
46	Sindhi	26	0.39%	96	Mossi	7.6	0.11%
47	Amharic	25	0.37%	97	Xhosa	7.6	0.11%
48	Fula	24	0.37%	98	Belarusian	7.6 [6]	0.11%
49	Romanian	24	0.37%	99	Balochi	7.6	0.11%
50	Oromo	24	0.36%	100	Konkani	7.4	0.11%
				Total		5,610	85%

thus with a limited audience, even though most existing human languages are spoken by only a small population of native speakers as showed in Table 1.

Despite the enormous number of people with technological knowledge and access, many are excluded because they cannot communicate globally due to language divides. Consistent with Anderson et al. [2], over 6,000 languages [2] are used globally; there is no universal spoken language for communication. The English language is only the third most popular (used by only 5.52% of the global population); Spanish (5.85%) and Mandarin (14.1%) are more common [3]. Moreover, fewer than 40% of citizens of the European Union (not including developing or Eastern European countries) know English [4], which makes communication a problem even within the EU [5].

This has created a technical gap between languages that are widely spoken in comparison to languages with few speakers. This also led to a big gap between quality and amount of available parallel corpora for less common language pairs, which makes natural language processing sciences slower in such countries.

As a result, high-quality data exist for just a few language pairs in particular domains (e.g. Czech-English law texts domain), whereas the majority of languages lack sufficient linguistic resources, such as parallel data for good quality research or natural language processing tasks. Building a translation system that can handle all possible language translations would require millions of translation directions and a huge volume of parallel data. Moreover, if we consider multiple domains in the equation, the requirements for corpus training in machine translation increase dramatically. Thus, the current study explored methods to build a corpus of high-quality parallel data, using Czech-English as the language pair.

Multiple studies have been performed to automatically acquire additional data for enhancing SMT systems in the long term [6]. All such approaches have focused on discovering authentic text from real-world sources for both the source and target languages. However, our study presents an alternative approach for building this parallel data. In creating virtual parallel data, as we might call it, at least one side of the parallel

data is generated, for which purpose we use monolingual text (news internet crawl in Czech, in this case). For the other side of the parallel data, we use an automated procedure to obtain a translation of the text. In other words, our approach generates rather than gathers parallel data. 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.

It is recommended that an estimate be considered reliable when multiple systems show a consensus on it. However, since the output of machine translation (MT) is human language, 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 complete agreement between multiple systems. To evaluate the generated data, we can use the Levenshtein distance as well as implementing a back-translation procedure. Using this approach, only those pairs that pass an initial compatibility check, when translated back into the native language and compared to the original sentences, will be accepted. This concept is depicted in Fig. 1.

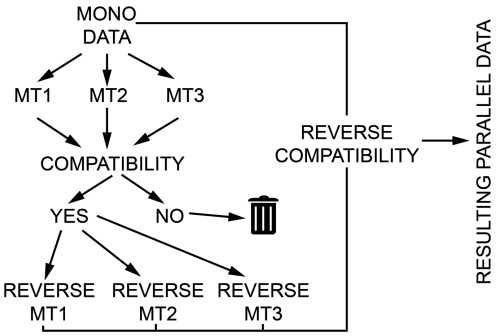

Fig. 1. Generation of artificial data

We can use this method to easily generate additional parallel data from monolingual news data provided for WMT16. 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 but important language pairs. 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. We are now working on two approaches for MT applications: self-training and translation via bridge languages (also called "pivot languages"). These approaches are different from those discussed previously:

While self-training is focused on exploiting the available bilingual data, to which the linguistic resources of a third language are rarely applied, translation via bridge languages focuses more on correcting the alignment of the prevailing word segment. This latter approach also incorporates the phrase model concept rather than exploring the new text in context, by examining translations at the word, phrase, or even sentence level, through bridge languages. The methodology of this paper lies in between the paradigm of self-training and translating via a bridge language. Our study generates data instead of gathering information for parallel data, while we also apply linguistic information and inter-language relationships to eventually produce translations between the source and target languages.

Callison-Burch and Osborne [7] presented a cooperative training method for SMT that comprises the consensus of several translation systems to identify the best translation resource for training. Similarly, Ueffing et al. [8] explored model adaptation methods to use monolingual data from a source language. Furthermore, as the learning progressed, the 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 [9] presented a technique to develop a translation lexicon based on transduction models of cognate pairs through a bridge language. In this case, the edit distance rate was applied to the process rather than the general MT system of limiting the vocabulary range for majority European languages. Kumar et al. [10] described the process of boosting word alignment quality using multiple bridge languages. In Wu and Wang [11], Habash and Hu [12], phrase translation tables were improved using phrase tables acquired in multiple ways from pivot languages. In Eisele et al. [13], a hybrid method was combined with RBMT (Rule-Based Machine Translation) and SMT systems. This methodology was introduced to fill gaps in the data for pivot translation. Cohn and Lapata [14] 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, in this study, we returned to the black-box translation system. This means that virtual data could be widely generated for translation systems, including rule-based, statistics-based, and human-based translations. The approach introduced in Leusch et al. [15] pooled the results of translations of a test set created by any of the pivot MTs per unique language. However, this approach was not found to enhance the systems, and hence the novel training data were not used. Amongst others, Bertoldi et al. [16] also conducted research on pivot languages, but did not consider applying universal corpus filtering, which is the measurement of compatibility to control data quality.

2.1 Generating Virtual Parallel Data

To generate new data, we trained three SMT systems based on TED, QED and News Commentary corpora. The Experiment Management System [17] 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) [18]. Word and phrase alignment was performed using the SyMGIZA++ symmetric word alignment tool [19] instead of GIZA++. Out-of-vocabulary (OOV) words were monitored using the Unsupervised Transliteration Model [20]. Working with the Czech (CS) and English (EN) language pair, the first SMT system was trained on TED [21], the second on the Qatar Computing Research Institute's Educational Domain Corpus (QED) [22], and the third using the News Commentary corpora provided for the WMT16 translation task. Official WMT16 test sets were used for system evaluation. Translation engine performance was measured by the BLEU metric [23]. The performance of the engines is shown in Table 2.

Table 2. 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 and is closely linked to the paired arrangement of strings [24].

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:

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

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 $\mathrm{lev}_{a,b}[i, j]$ 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 3.

Table 3. 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	19,743,105	2,125,154

The purpose of this research was to create synthetic parallel data to train a machine translation system by translating monolingual texts with multiple machine translation systems and various filtering steps. This objective is not new; synthetic data have been created in the past. However, the novel aspect of the present paper is its use of three MT systems, application of the Levenshtein distance between their outputs as a filter, and—much more importantly—its use of back-translation as an additional filtering step. In Table 4, we show statistical information on the corpora used without the back-translation step.

Table 4. Specification of generated corpora without back-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	19,743,105	10,453,241

2.2 Semantically-Enhanced Generated Corpora

The artificially generated corpora presented in Table 3 were obtained using statistical translation models, which are based purely on how frequently "things" happen, and not on what they really mean. This means that they do not really understand what was translated. In this research, these data were additionally extended with semantic information so as to improve the quality and scope of the data domain. The word relationships were integrated into generated data using the WordNet database.

The way in which WordNet was used to obtain a probability estimator was shown in Cao et al. [25]. In particular, we wanted to obtain $P(w_i|w)$, where w_i and w are assumed to have a relationship in WordNet. The formula is as follows:

$$P(w_i|w) = \frac{c(w_i, w|W,L)}{\sum_{w_j} c(w_j, w|W,L)}$$

where W is a window size and $c(w_i, w|W, L)$ is the count of wi and w appearing together within W-window. This can be obtained simply by counting each within a certain corpus. In order to smooth the model, we applied interpolated Kneser-Ney [26] smoothing strategies.

The following relationships were considered: synonym, hypernym, hyponym, and hierarchical distance between words.

In Table 5, we show statistical information on the semantically enhanced corpora produced previously and shown in Table 3.

Table 5. Specification of semantically generated corpora without back-translation

Data set	Number of sentences		Number of unique czech tokens	
	Monolingual	Generated	Monolingual	Generated
News 2007	100,766	122,234	200,830	98,275
News 2008	4,292,298	1,467,243	2,214,356	803,852
News 2009	4,432,383	1,110,234	2,172,580	959,847
News 2010	2,447,681	982,747	1,487,500	585,852
News 2011	8,746,448	1,397,975	2,871,190	1,119,281
News 2012	7,538,499	1,759,285	2,589,424	968,975
News 2013	8,886,151	1,693,267	2,768,010	982,948
News 2014	8,722,306	1,462,827	2,814,742	1,243,286
News 2015	8,234,140	1,839,297	2,624,473	1,273,578
Total	53,366,020	11,835,109	19,743,105	8,035,470

Another common approach to semantic analysis that is also used within this research is latent semantic analysis (LSA). LSA has already been shown to be very helpful in automatic speech recognition (ASR) [27] and many other applications, which was the reason for incorporating it within the scope of this research. The high-level idea

of LSA is to convert words into concept representations and to assume that if the occurrence of word patterns in documents is similar, then the words are also similar. The mathematical model can be defined as follows:

In order to build the LSA model, a co-occurrence matrix W will first be built, where w_{ij} is a weighted count of word w_j and document d_j.

$$w_{ij} = G_i L_{ij} C_{ij}$$

where C_{ij} is the count of w_i in document d_j; L_{ij} is local weight; and G_i is global weight. Usually, L_{ij} and G_i can use TF/IDF.

Then, singular value decomposition (SVD) analysis will be applied to W, as

$$W = U\ S\ V^T$$

where W is a M * N matrix (M is vocabulary size, N is document size); U is M * R, S is R * R, and V is a R * N matrix. R is usually a predefined dimension number between 100 and 500.

After that, each word w_i can be denoted as a new vector $U_i = u_i * S$. Based on this new vector, the distance between two words is defined as:

$$K(U_i,\ U_j) = \left\{u_i * S^2 * u_m^T\right\}\{|u_i * S| * |u_m * S|\}$$

Therefore, clustering can be performed to organize words into K clusters, $C_1, C_2, \ldots, C_K$.

If H_{q-1} is the history for word W_q, then it is possible to obtain the probability of W_q given H_{q-1} using the following formula:

$$\begin{aligned} P(W_q|H_{q-1}) &= P(W_q|W_{q-1}, W_{q-2}, \ldots W_{q-n+1},\ d_{q_1}) \\ &= P(W_q|W_{q-1}, W_{q-2}, \ldots W_{q-n+1}) * P(W_q|d_{q_1}|) \end{aligned}$$

where $P(W_q|W_{q-1}, W_{q-2}, \ldots W_{q-n+1},\ d_{q_1})$ is the N-gram model; $P(d_{q_1}|W_q)$ is the LSA model.

Additionally,

$$P(W_q|d_{q_1}) = P(U_q|V_q) = K(U_q, V_{q_1})/Z(U,V)K(U_q, V_{q_1}) = \frac{U_q * S * V_{q-1}^T}{|U_q * S^{1/2}| * |V_{q-1} * S^{1/2}|},$$

where Z(U, V) is the normalized factor.

It is possible to also apply word smoothing to the model-based K-Clustering as follows:

$$P(W_q|d_{q_1}) = \sum_{k=1}^{K} P(W_q|C_k)P(C_k|d_{q_1})$$

where $P(W_q|C_k)$, $P(C_k|d_{q_1})$ can be computed using the distance measurement given above by a normalized factor.

In this way, the N-gram and LSA model are combined into a single language model and can be used for word comparison and text generation. The Python code for such LSA analysis was implemented in Thomo's [28] research.

In Table 6, we show statistical information on the semantically enhanced corpora produced previously and shown in Table 3.

Table 6. Specification of semantically generated corpora using LSA

Data set	Number of sentences		Number of unique czech tokens	
	Monolingual	Generated	Monolingual	Generated
News 2007	100,766	98,726	200,830	72,975
News 2008	4,292,298	868,862	2,214,356	592,862
News 2009	4,432,383	895,127	2,172,580	729,972
News 2010	2,447,681	725,751	1,487,500	472,976
News 2011	8,746,448	1,197,762	2,871,190	829,927
News 2012	7,538,499	1,298,765	2,589,424	750,865
News 2013	8,886,151	1,314,276	2,768,010	694,290
News 2014	8,722,306	1,267,862	2,814,742	992,893
News 2015	8,234,140	1,471,287	2,624,473	892,291
Total	53,366,020	9,138,418	19,743,105	6,029,051

2.3 Experimental Setup

The machine translation experiments we conducted involved 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 used the SyMGiza++ tool, which assisted 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 shown to be fruitful in previous research [19]. OOV words are another challenge for an SMT system and to deal with such 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 probable solutions. UTM was used to score several possible transliterations and to find a translation table [20, 29].

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 each target–swap (S), monotone (M), and discontinuous (D)–all three of which were used in a hierarchical model. The bidirectional restructuring model was used to examine the phrase arrangement probabilities [30–32].

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 greatly in this regard [33]. A parallel, generalized domain corpus and monolingual corpus were used in this process, as identified by Wang et al. [34]. First, sentence pairs of the parallel data were weighted based on their significance to the targeted domain. Second, reorganization was conducted to obtain the best sentence pairs. After obtaining the required sentence pairs, these models were trained for the target domain [34].

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. Moreover, Wang et al. observed that a combination of these approaches provides the best possible solution for domain adaptation for Chinese-English corpora [34]. Thus, inspired by Wang et al.'s approach, 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 approximately 20% of the domain analogous data.

2.4 Evaluation

To make progress in machine translation (MT), the quality of its results must be evaluated. It has been recognized for quite some time that using humans to evaluate MT approaches is very expensive and time-consuming [35]. As a result, human evaluation cannot keep up with the growing and continual need for MT evaluation, leading to the recognition that the development of automated MT evaluation techniques is critical. Evaluation is particularly crucial for translation between languages from different families (i.e., Germanic and Slavic), such as Polish and English [35, 36].

Vanni and Reeder [36] compiled an initial list of SMT evaluation metrics. Further research has led to the development of newer metrics. Prominent metrics include Bilingual Evaluation Understudy (BLEU), the National Institute of Standards and Technology (NIST) metric, Translation Error Rate (TER), and the Metric for Evaluation of Translation with Explicit Ordering (METEOR). These metrics were used in this research for evaluation.

In this research, we used the most popular metric BLEU, which was developed based on a premise similar to that used for speech recognition, described by Papineni et al. [23] as "The closer a machine translation is to a professional human translation, the better it is." Thus, the BLEU metric is designed to measure how close SMT output is to the output of human reference translations. It is important to note that translations,

be they SMT or human, may differ significantly in terms of word usage, word order, and phrase length [23].

2.4.1 Statistical Significance Tests

In cases where the differences in metrics described above do not deviate greatly from each other, a statistical significance test can be performed. The Wilcoxon test [37] (also known as the signed-rank or matched-pairs test) is one of the most popular alternatives to the Student's t-test for dependent samples. It belongs to the group of non-parametric tests and is used to compare two (and only two) dependent groups that involve two measurement variables.

The Wilcoxon test is used when the assumptions for the Student's t-test for dependent samples are not valid; for this reason, it is considered an alternative to this test. The Wilcoxon test is also used when variables are measured on an ordinal scale (in the Student's t-test, the variables must be measured on a quantitative scale). The requirement for application of the Wilcoxon test is the potential to rank differences between the first and second variable (the measurement). On an ordinal scale, it is possible to calculate the difference in levels between two variables; therefore, the test can be used for variables calculated on such a scale. In the case of quantitative scales, this test is used if the distributions of these variables are not close to the normal distribution.

3 Results and Discussion

Numerous human languages are used around the world and millions of translation systems have been introduced for the possible language pairs. However, these translation systems struggle with high quality performance, largely due to the limited availability of language resources such as parallel data.

In this study, 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 4), our data generation approach has increased translation performance. Although the results appear very promising, there remains a great deal of room for improvement. Performance improvements can be attained by applying more sophisticated algorithms to quantify the comparison among different MT engines. In Table 6, 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 3, 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. Statistically significant results in accordance with the Wilcoxon test are marked with * and those that are very significant with ** (Table 7).

Table 7. 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*

Next, in Table 8, we replicate the same quality experiment but using generated data without the back-translation step. As shown in Table 4, more data can be obtained in such a manner. However, the SMT results are not as good as those obtained using back-translation. This means that the generated data must be noisy and most likely contain incomplete sentences that are removed after back-translation.

Table 8. Evaluation of corpora generated without the back-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	16.83*
	EN → CS	BASE	14.25
	EN → CS	FINAL	15.03**

Next, in Table 9, we replicate the same quality experiment but using generated data from Table 5. As shown in Table 9, augmenting virtual corpora with semantic information makes a positive impact on not only the data volume but also data quality. Semantic relations improve the MT quality even more.

Table 9. Evaluation of semantically generated corpora without the back-translation step

Domain	Direction	System	BLEU
News	CS → EN	BASE	15.26
	CS → EN	FINAL	19.31**
	EN → CS	BASE	11.64
	EN → CS	FINAL	14.87**
IT	CS → EN	BASE	12.86
	CS → EN	FINAL	15.42**
	EN → CS	BASE	10.19
	EN → CS	FINAL	12.17**
Bio-medical	CS → EN	BASE	16.75
	CS → EN	FINAL	19.47**
	EN → CS	BASE	14.25
	EN → CS	FINAL	16.13**

Finally, in Table 10, we replicate the same quality experiment but using generated data from Table 6 (LSA). As shown in Table 10, augmenting virtual corpora with semantic information by facilitating LSA makes an even more positive impact on data quality. LSA-based semantic relations improve the MT quality even more. It is worth mentioning that LSA provided us with less data but we believe that it was more accurate and more domain-specific than the data generated using Wordnet.

Table 10. Evaluation of semantically generated corpora using LSA

Domain	Direction	System	BLEU
News23	CS → EN	BASE	15.26
	CS → EN	FINAL	19.87**
	EN → CS	BASE	11.64
	EN → CS	FINAL	15.61**
IT	CS → EN	BASE	12.86
	CS → EN	FINAL	16.18**
	EN → CS	BASE	10.19
	EN → CS	FINAL	13.04**
Bio-medical	CS → EN	BASE	16.75
	CS → EN	FINAL	20.37**
	EN → CS	BASE	14.25
	EN → CS	FINAL	17.28**

4 Conclusions

Summing up, in this study, we successfully built parallel corpora of satisfying quality from monolingual resources. This method is very time and cost effective and can be applied to any bilingual pair. In addition, it might prove very useful for rare and

under-resourced languages. However, there is still room for improvement, for example, by using better alignment models, neural machine translation, or adding more machine translation engines to our methodology. Moreover, using Framenet, which provides semantic roles for a word and shows restrictions in word usage, in that only several kinds of word can be followed by a certain word, might be of interest for future research [38].

References

1. Wołk, K., Marasek, K., Wołk, A.: Exploration for Polish-* bi-lingual translation equivalents from comparable and quasi-comparable corpora. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, pp. 517–525 (2016)
2. Anderson, S.R., Harrison, D., Horn, L., Zanuttini, R., Lightfoot, D.: How many languages are there in the world?: linguistic society of America (2010). http://www.linguisticsociety.org/sites/default/files/how-many-languages.pdf. Accessed 16 Feb 2017
3. List of languages by number of native speakers (2016). Wikipedia, https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers. Accessed 16 Feb 2016
4. Paolillo, J., Anupam, D.: Evaluating language statistics: the ethnologue and beyond (2006). http://www.uis.unesco.org/Library/Documents/evaluating-language-statistics-ethnologue-beyond-culture-2006-en.pdf. Accessed 8 Oct 2015
5. English language in Europe 2016 Wikipedia. https://en.wikipedia.org/wiki/English_language_in_Europe. Accessed 16 Feb 2017
6. Munteanu, D., Fraser, A., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In: Human Language Technologies-The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Marina del Rey, pp. 265–272 (2004)
7. Callison-Burch, C., Osborne, M.: Co-training for statistical machine translation. Dissertation, School of Informatics, University of Edinburgh (2002)
8. Ueffing, N., Haffari, G., Sarkar, A.: Semisupervised learning for machine translation. In: Goutte, C., Cancedda, N., Dymetman, M., Foster, G. (eds.) Learning Machine Translation, pp. 237–256. MIT Press, Pittsburgh (2009)
9. Mann, G., 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, Pittsburgh, pp. 1–8 (2001)
10. Kumar, S., Och, F., Macherey, W.: Improving word alignment with bridge languages. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pp. 42–50 (2007)
11. Wu, H., Wang, H.: Pivot language approach for phrase-based statistical machine translation. Mach. Transl. **21**(3), 165–181 (2007)
12. 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. Association of Computational Linguistics, Athens, pp. 173–181 (2009)

13. Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., Chen, Y.: Hybrid machine translation architectures within and beyond the EuroMatrix project. In: Hutchins, J., Hahn, W.V. (eds.) Hybrid MT Methods in Practice: Their Use in Multilingual Extraction, Cross-Language Information Retrieval, Multilingual Summarization, and Applications in Hand-Held Devices. Proceedings of the European Machine Translation Conference, Proceedings of the 12th Annual Conference of the European Association for Machine Translation. HITEC e.V., European Association for Machine Translation, Hamburg, Germany, pp. 27–34 (2008)
14. Cohn, T., Lapata, M.: Machine translation by triangulation: making effective use of multi-parallel corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, pp. 728–735 (2007)
15. Leusch, G., Max, A., Crego, J.M., Ney, H.: Multi-pivot translation by system combination. In: Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT), Paris, pp. 299–306 (2010)
16. Bertoldi, N., Barbaiani, M., Federico, M., Cattoni, R.: Phrase-based statistical machine translation with pivot languages. In: Proceedings of IWSLT, Hawaii, pp. 143–149 (2008)
17. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association of Computational Linguistics, Prague, pp. 177–180 (2007)
18. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Proceedings of International Conference Spoken Language Processing, Denver, pp. 901–904 (2002)
19. Junczys-Dowmunt, M., Szal, A.: SyMGiza ++: symmetrized word alignment models for statistical machine translation. In: Bouvry, P., Kłopotek, M.A., Leprévost, F., Marciniak, M., Mykowiecka, A., Rybiński, H. (eds.) Security and Intelligent Information Systems: International Joint Conferences, 2011, Warsaw, pp. 379–390. Springer, Heidelberg (2012)
20. Durrani, N., Sajjad, H., Hoang, H., Koehn, P.: Integrating an unsupervised transliteration model into statistical machine translation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, pp. 148–153 (2014)
21. Cettolo, M., Girardi, C., Fedirico, M.: WIT3: web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation, Trento, pp. 261–268 (2012)
22. Abdelali, A., Guzman, F., Sajjad, H., Vogel, S.: The AMARA corpus: building parallel language resources for the educational domain. In: Ninth International Conference on Language Resources and Evaluation (LREC14), Reykjavik, pp. 1044–1054 (2014)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of Association for Computational Linguistics, Philadelphia, pp. 311–318 (2002)
24. Yujian, L., Bo, L.: A normalized Levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1091–1095 (2007)
25. Cao, G., Nie, J., Bai, J.: Integrating term relationships into language models. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, pp. 298–305 (2005)
26. Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Comput. Speech Lang. **13**(4), 359–394 (1999)
27. Bellegarda, J.: Data-driven semantic language modeling, Institute for Mathematics and Its Applications Workshop (2000). http://cmusphinx.sourceforge.net/wiki/semanticlanguagemodel. Accessed 16 Feb 2017

28. Thomo, A.: Latent semantic analysis (LSA) tutorial (2009). http://webhome.cs.uvic.ca/~thomo/svd.pdf. Accessed 16 Feb 2007
29. Moses statistical machine translation, OOVs (2015). http://www.statmt.org/moses/?n=Advanced.OOVs#ntoc2. Accessed 27 Sept 2015
30. Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association of Computational Linguistics, Edinburgh, pp. 187–197 (2011)
31. Costa-jussa, M.R., Fonollosa, J.R.: Using linear interpolation and weighted reordering hypotheses in the Moses system. In: Seventh Conference on International Language Resources and Evaluation, Valletta, pp. 1712–1718 (2011)
32. Moses statistical machine translation, Build reordering model (2013) http://www.statmt.org/moses/?n=FactoredTraining.Build. Reordering Model. Accessed 10 Oct 2015
33. Axelrod, 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. Association of Computational Linguistics, Edinburgh, pp. 355–362 (2011)
34. Wang, L., Wong, D.F., Chao, L.S., Lu, Y., Xing, J.: A systematic comparison of data selection criteria for SMT domain adaptation. Sci. World J. **2014**, 745485 (2014)
35. Hovy, E.: Toward finely differentiated evaluation metrics for machine translation. In: Proceedings of the EAGLES Workshop on Standards and Evaluation, Pisa, pp. 127–133 (1999)
36. Vanni, M., Reeder, F.: How are you doing? A look at MT evaluation. In: White, J.S. (eds.), Envisioning Machine Translation in the Information Future, AMTA 2000. LNCS, vol. 1934. Springer, Heidelberg (2000)
37. Oyeka, I.C.A., Ebuh, G.U.: Modified Wilcoxon signed-rank test. Open J. Stat. **2**, 172–176 (2012)
38. Lin, S., Verspoor, K.: A semantics-enhanced language model for unsupervised word sense disambiguation. In: Ninth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2008). Lecture Notes in Computer Science (LNCS), Haifa, pp. 287–298 (2008)