Generating Virtual Parallel Corpus: A Compatibility Centric Method

Jia Xu^{+} and Weiwei Sun^{+*}
$^{+}$German Research Center for Artificial Intelligence (DFKI)
*Department of Computational Linguistics, Saarland University
D-66123, Saarbrücken, Germany
Jia.Xu@dfki.de,wsun@coli.uni-saarland.de

Abstract

The processing of many natural languages suffers from scarce linguistic resources. We introduce the idea of compatibility to extend training data for machine translation: If translation hypotheses by multiple systems are measured as compatible, they are considered as reliable predictions. By this way, we generate virtual parallel data per bridge language, and re-compiling on this corpus improves our machine translation quality by more than 30% relatively.

1 Introduction

Statistical machine translation (SMT) is a data-driven approach. The quantity and quality of parallel data is crucial to build high performance SMT systems. Nonetheless, nowadays parallel corpora are still limited in quantity, genre and language coverage. In particular, the languages with less native speakers are less investigated. This results in a technical gap between the translation on widely spoken languages and on other languages. On the other hand, the overwhelming majority of human languages are spoken by minority population of native speakers. Due to limited human efforts, rich labeled data are only available for few language pairs in certain domains, while most of languages are lacking sufficient linguistic resources such as parallel data. In order to build translation systems covering all language directions, which can be millions, a great amount of parallel corpora are required. Additionally, if we take various domains into account, human annotation alone can hardly meet the increasing demand on the huge amount of training corpus.

In this paper, we are concerned with building *high-quality*, *virtual* parallel data for machine translation. Many investigations have been performed to obtain extra data automatically in order to improve machine translation systems, not the least among them being (Munteanu et al., 2004), (Smith et al., 2010) etc. These approaches have been focused on exploring *real* text of the source and/or target languages. Different from their work, we present an alternative idea to build *virtual* (i.e. pseudo) parallel data. By virtual, we mean at least one side of the parallel data is artificial. In this work, for one side of the new data, we use real monolingual texts; for the other side, we preform an automatic or semi-automatic procedure to obtain the translated results of these texts. In other words, we focus on *generating*, rather than *gathering*, parallel data.

To control the quality of the automatically generated data and to ensure its usefulness for MT, we introduce the idea of compatibility. Our method is inspired by research on agreement-based semi-supervised learning methods, such as co-training, and leverages multiple MT systems. Generally speaking, the compatible predictions provided by multiple systems is more reliable. For simple classification problems, it is reasonable to regard a prediction as good when the multiple systems agree on it. However, the output of MT is in human languages, which is too complex. It is too strict and unreasonable to ask multiple systems to provide exactly the same translated sentence for an input. In this case, we consider the compatibility, rather than agreement, of multiple predictions. Assume without loss of generality that we have two MT systems, A and B, both of which can translate language L^s to language L^t. To obtain a pseudo sentence pair $(s, t) \in L^s \times L^t$, we first pick up a "real" sentence s from monolingual data. Both A and B can provide translated results t_A and t_B. If t_A and t_B are compatible, either t_A or t_B is allowed to be collected as reliable virtual data. The result of compatibility

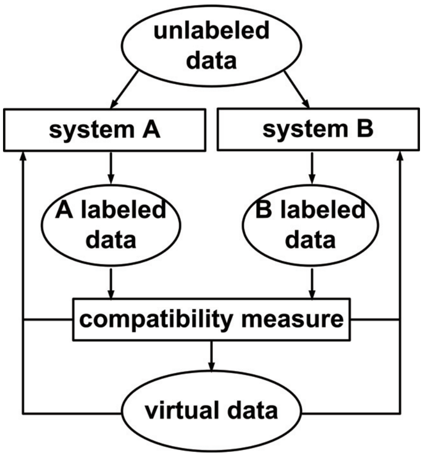

Figure 1: Compatibility and virtual Data.

measure and virtual data can be applied back to train better systems. This procedure is shown in Figure 1. Note that our approach is independent of MT system without assumption on any translation algorithm.

There are two key problems in our compatibility-centric method: (1) How to build multiple MT systems? (2) How to measure compatibility of the outputs of multiple systems? Note that our purpose is to generate virtual data. For the first problem, this purpose implies that we can choose the inputs to make sure that at least one system works well on these inputs. For the second problem, since it is not necessary to put both t_A and t_B into the virtual data collection, there is no need to explicitly measure the compatibility of two sentences. Our method can still work when we only evaluate the compatibility of a translated sentence t with another MT system.

To implicate our idea, we take the Romanian-German translation as an example in this work. In particular, our special solution leverages multiple parallel corpora in different language pairs and MT systems correlated to these languages. For the first aforementioned problem, we build a complementary system per a bridge language, English. The system of Romanian-German, which is suboptimal due to sparse training data, can be improved through translation systems with richer resources, i.e. Romanian-English and English-German system. Let Romanian-German MT be system A and Romanian-English and English-German MT be system B. Since the system A has scarce resource, we need to improve system A with the help of system B. For the second problem, we use the sentence level confidence measure calculated based on word alignment models to measure the compatibility of the system A and the outputs given by B.

In our experiment, we translate the English text of Romanian-English parallel corpus into German using our existing MT system, in order to generate the parallel corpus of Romanian-German; we can also generate it from the other direction by translating English text of German-English parallel corpus into Romanian automatically. Using this method, we obtain around 20% additional parallel data to the existing parallel data of German and Romanian. Re-training the MT systems using our obtained data improves the translation performance by more than 30% relatively, in both German-Romanian and Romanian-German translation systems. Linguistic resources in the scarce resourced language pairs are greatly advanced. This method is not limited to languages but is applicable for all scarce resourced language pairs. The generated virtual parallel corpus can not only be applied into MT but also other NLP tasks.

2 Generating Virtual Parallel Data

2.1 Background and Motivation

There are only a few parallel corpora publicly available for some languages we work on. The JRC-Acquis(JRC) is a huge collection of European Union legislative documents translated into more than twenty official European languages (Steinberger et al., 2006). The European Parliament Proceedings Parallel Corpus (Europarl corpus) was extracted from the proceedings of the European Parliament (1996-today) (Koehn, 2005). News Commentary(NC) (SMT, 2011) and SETimes (SETIMES, 2011) are corpora collected from the news domains.

In this paper, we are concerned with generating *high-quality*, *virtual* parallel data for machine translation. To do this, we exploit multiple parallel corpora in different language pairs. In particular, we generate parallel corpora for scarce resourced languages, taking Romanian to German as a case study for simplicity. We can also take German to Romanian or other language directions.

In order to find out the gap between the translation quality on better studied language pairs and that on less studied language pairs, we consider the widely spoken language English as a bridge and perform baseline translation experiments on all directions of Romanian, German and English. These MT experiments are setup in the same way using the method to be described in Section 3.1.

Our test corpus is a collection of multilingual corpus in ten European languages. The domain of our test corpus is a mixture of general information about European Union, popular scientific and educational, official and legal documents, news and magazine ar-

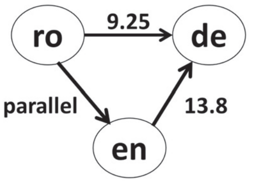

Figure 2: Learning the ro-de virtual parallel corpus per ro-en parallel corpus and en→de MT. The translation performance in the BLEU[%] score of the baseline systems is annotated on the edge.

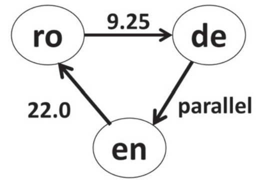

Figure 3: Learning the ro-de virtual parallel corpus per de-en parallel corpus and en→ro MT. The translation performance in the BLEU[%] score of the baseline systems is annotated on the edge.

src	tgt	BLEU[%]	corpus	#[M]
de	en	28.3	Europarl, NC	1.86
en	de	13.8	Europarl, NC	1.86
en	ro	22.0	SETimes	0.81
ro	en	27.7	SETimes	0.81
de	ro	8.87	DGT, JRC	0.56
ro	de	9.25	DGT, JRC	0.56

Table 1: Baseline translation performance in BLEU[%] and domains and the number of sentences (#) in millions of available training corpora for different translation directions.

ticles, information technology articles, letters and fictions. We did not select test sentences from the JRC corpus, because the JRC corpus contains many redundant sentences, and the evaluation will be overestimated. The training corpora are listed in Table 1. We applied Europarl and NC corpus for English-German, SETimes for English-Romanian and JRC plus DGT for Romanian-German in training.

Table 1 shows the baseline translation performance between German-English, Romanian-English and German-Romanian, as well as the domains and number of sentences of the available parallel corpus. Although the numerical performance measured in the BLEU score (Papineni et al., 2002) on the multilingual test set is not a fair comparison criterion across different language pairs, it still indicates the degree of the translation quality. Translations between Romanian and German has a low quality due to out-of-domain training corpus. Therefore, we generate the virtual Romanian-German corpus on the in-domain news corpus, NC and SETimes through English.

2.2 Bi-directional Generation

Since the translation performance of Romanian-German (9.25% in BLEU) is lower and the performance of English-German is higher, we can learn Romanian-German translations across English, which means the Romanian text is translated into English and the English output is translated into German. Aligning the Romanian text and the translated German text leads to automatically generated Romanian-German translations. Learning translations here can either be manually, using labeled data, or automatically. In practice, translating twice automatically, for example from Romanian to English and from English to German, can result in multiplied errors. Therefore, we used parallel data of Romanian-English and only translate English into German using our machine translation system. The process is illustrated in Figure 2. As the second approach, this process can also go from the other direction as shown in Figure 3, namely, German to English and English to Romanian, where we use the German-English parallel data and the English to Romanian translation system. As the third approach, we combine the obtained parallel data from the first and the second approach, which gives us the virtual parallel data both in the NC and SETimes. Europarl data is not applied here due to its legalism domain.

The parallel corpus generated in this way can be directly used as part of the training data. One advantage of this method is that we can generate translation of unknown words using the existing system to reduce OOVs. These additional word or phrase translations come from the training data per bridge language. However, the corpus obtained can be noisy. Therefore, we applied compatibility centric approach to generate high quality data for efficient training and for better translation performance.

2.3 The Compatibility Idea

The quality of the automatically generated data is very important for its application in MT. Our solution controls the quality of the virtual data by two ways. As mentioned above, the first control is the use of "real" Romain-English and German-English parallel data rather than automatic Romain-English and German-English systems, which may cause numerous errors. The second control is based on the idea of compatibility. Generally speaking, the quality of compatible predictions provided by multiple systems is more reliable. For simple classification problems, it is reasonable to take a prediction as good which the multiple systems agree on. This idea is widely used in ensemble learning and semi-supervised learning. Take Bootstrap aggregating, a meta-algorithm for ensemble learning as an example, multiple models are separately trained on randomly generated sub-samples, and then vote to achieve final predictions. Another example closely related to our method is co-training such as in (Callison-Burch, 2002). One way to select automatic predictions for re-training in co-training is to choose the agreed ones.

Different from simple classification problems, even complex structured prediction problems such as parsing, the output of MT is in human languages, which may be the most complicated way to represent the meaning of another human language. It is too strict to ask multiple systems to provide exactly the same translated sentence for an input. We extend the agreement idea to the compatibility idea. Informally, two sentences are called *compatible* if they express the same meaning to some extent. We collect compatible translations which are more reliable into the virtual data set.

2.4 Compatibility Measure

The realizations of compatibility measure varies for different applications and from case to case. Here we present a method for machine translation. As shown in Figure 1, the output by system A and the output by system B can be applied to find the compatibility measure rules and the virtual data. We do not generate the single best using each system but use the word alignment models of system A to calculate the confidence of the single best output of system B. The advantage of this approach over directly comparing the single best output by each system is that the underlying translation models can be considered. Better score suggested by system A indicates a higher compatibility between system A and system B. Note, that the compatibility measure is different than the confidence measure, where the former one can take the latter as a realization but is not limited to.

We evaluate the quality of each sentence pair and choose a certain percentage of the best scored sentences for training. In order to include information from various resources, the quality of a sentence pair is measured using a log-linear model combining different sub-models. Let (f_1^J, e_1^I) be a bilingual sentence, the evaluation is performed using the following Equation:

$$H(f_1^J, e_1^I) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I)$$

$h_m(f_1^J, e_1^I)$ is a score evaluated on this sentence pair using sub-model m. Each model m is assigned with a feature weight λ_m. For simplicity, we only include the negative logarithm of IBM model 1 (Brown et al., 1993) in normal and inverse direction as sub-models. We combine the IBM model 1 in both directions in the log-linear model with an equal weight for each direction. We use the training software GIZA++ (Och and Ney, 2003) to obtain the lexicon probability.

3 Experimental Results

3.1 MT Setup

We apply Moses (Koehn et al., 2007) as our baseline translation system and train standard alignment models in both directions with GIZA++ (Och and Ney, 2003) using models of IBM-1 (Brown et al., 1993), HMM (Vogel et al., 1996) and IBM-4 (Brown et al., 1993) which brings us the optimal translation performance and efficiency based on empirical evaluations. Features in the log-linear model include translation models in two directions, a language model, a distortion model and a sentence length penalty. The language model is a statistical 5-gram model with modified Kneser-Ney smoothing estimated using SRI-LM toolkit (Stolcke, 2002). Each language model is trained with the target side of the parallel data. We do not apply any zmert tuning in EMS because it does not improve our translation results on the evaluation set. Importantly, our proposed method is independent on the SMT systems, i.e. the generation and evaluation of virtual data can be applied on any SMT system with various algorithms and configurations. The training and the test corpora are described in Section 2.1.

We perform machine translation experiments on Romanian to German to evaluate the quality of our generated corpus. The following translation systems are built and tested:

1. baseline: The baseline system is trained on the JRC and DGT corpora.
2. ro-en-de-80%: We translate a portion of English side of the Romanian-English parallel corpus in SETimes to obtain the virtual data. We rank the sentence pairs using the method described in Section 2.4, then include 80% best sentence pairs in the training, together with the baseline corpus, to re-train the translation system.
3. de-en-ro-all: We translate the English side of the German-English parallel corpus in NC to obtain the virtual data, then include all the virtual data, together with the baseline corpus, to

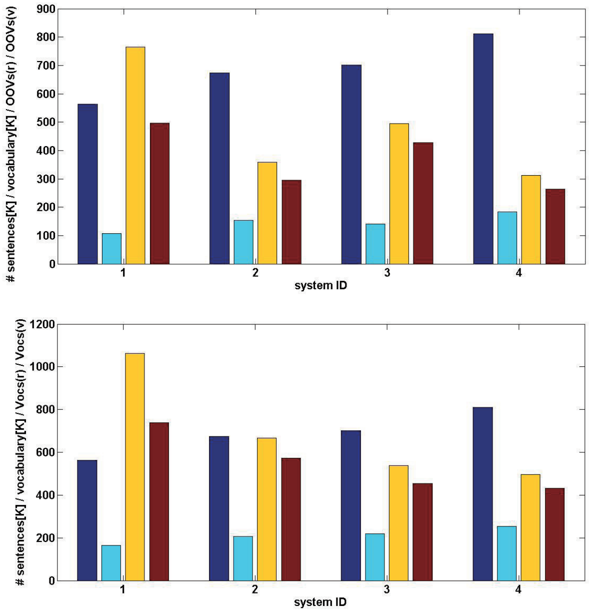

Figure 4: Corpus statistics on the training corpus in the baseline (1), ro-en-de-80% (2), de-en-ro (3) and combined (4) system for Romanian (upper bar) and for German (lower bar). Number of training sentences are increased, OOVs in running words and OOVs in vocabulary are significantly reduced by including virtual corpus.

re-train the translation system, as shown in Figure 3.

4. combined: The virtual data obtained using system ro-en-de-80% and de-en-ro-all are combined and included into the training. We only perform corpus selection for the ro-en-de system for convenience.

5. ro-en-de-all: We add all the virtual data obtained in ro-en-de-80% without filtering. This process is shown in Figure 2.

6. direct translation: As a comparison, we directly translate the Romanian text into English using our standard ro-en system and then translate this English output into German using our standard en-de system.

3.2 Corpus Statistics

Figure 4 shows the corpus statistics of different translation systems with their IDs described in Section 3.1. We compare the number of sentences[K], vocabulary size[K], OOVs of running words and OOVs of vocabulary. The upper bar shows the statistics of the source language, Romanian. The lower bar shows the statistics of the target language, German. The test corpus contains 512 sentences with 13.1K and 12.3K running words in Romanian(ro) and German(de), respectively. The vocabulary size of the test corpus is around 3.8K. In the baseline translation system, we use JRC and DGT as training corpus containing 563K sentence pairs. The OOVs are 764 and 1063 in running words (tokens) and 496 and 739 in vocabulary (types) for Romanian and German, respectively. After applying the parallel corpus generated by the ro-en-de-80% system, the training corpus contains 674K sentence pairs. With around 20% additional data, the OOVs in running words are reduced to 358 in Romanian and 668 in German, and the OOVs in vocabulary are reduced to 295 in Romanian and 574 in German, respectively. Adding the corpus generated by the de-en-ro system, the training corpus contains 701K sentence pairs with reductions of OOVs both in running words and in vocabulary, too. As we add the generated corpus by the ro-en-de-80% system and by the de-en-ro system, the training corpus contains 811K sentences pairs, with a size of around 60% more than that of the baseline. The OOVs are further reduced to 313/495 and 263/432 in Romanian/German both in running words and vocabulary, respectively. We can see that it effectively reduces the number of unknown words by including our generated data into the training.

ID	system	BLEU[%]
1	baseline	9.41
5	direct translation	11.6
6	+ro-en-de-all	11.5
2	+ro-en-de-80%	11.9
3	+de-en-ro-all	11.4
4	+combined	12.4

Table 2: Translation performance on ro→de in BLEU[%] using additional virtual corpus obtained by different ways.

3.3 Evaluation Results

As shown in Table 2, in the baseline system of Romanian to German, the BLEU score is very low, 9.41% due to the out-of-domain and small scaled available training data. In general we can directly translate the Romanian test set into English then translate this English output into German using the baseline ro-en and en-de systems respectively. The BLEU score is 11.6% using this approach. We apply the parallel corpus obtained by translating the English text in the Romanian-English parallel corpus of SETimes, and the BLEU score is increased from 9.41% to 11.5%. By including the parallel corpus obtained from the other direction to the baseline system, de-en-ro, the performance is improved

system	BLEU[%]
baseline	8.87
+combined	11.99

Table 3: Translation performance on de→ro in BLEU[%] using the virtual corpus in system 4 in Table 2.

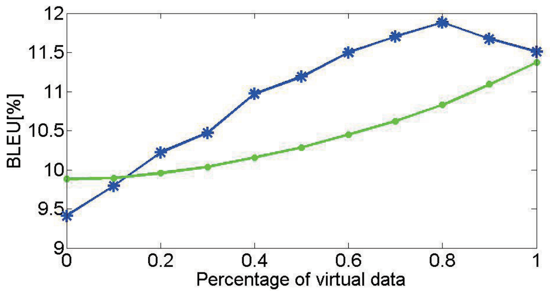

Figure 5: Balance of precision and recall: BLEU[%] (*) and the phrase table size (o) vs. percentage of virtual data included in the training.

from 9.41% to 11.4% in the BLEU score. However, this does not outperform the direct translation approach. Therefore, we select better sentences using the log-linear model of IBM model-1 in normal and inverse directions as described in Section 2.4. This leads to an improvement on the translation performance over the direct approach namely 11.9% vs. 11.6% in the BLEU score as well as a reduction on the training data size. The compatibility computation reduced training time and enhanced the translation quality at the same time. If we put parallel corpra acquired both from ro-en-de-80% and de-en-ro directions into the training, we achieve a result of 12.4% in the BLEU score, which is more than 30% relative improvement over the baseline system and nearly 10% over the direct translation system.

We also compile the virtual corpus used in system 4 in Table 2 to translate from German to Romanian. The translation result on the baseline system is evaluated as 8.87% in the BLEU score. Applying the virtual corpus enhances the translation quality significantly, i.e. the BLEU score is increased to 11.99%.

3.4 Balance of Precision and Recall

In data selection, sentence pairs are ranked after the compatibility cost, an example is shown in Table 4. Then we set a threshold to control how many percent of the best ranked sentence pairs to be included into the training. As discussed in (Deng et al., 2008), a larger phrase table does not always lead to a better translation quality, and the optimal translation performance and efficiency can be achieved through balancing the precision and the recall. Figure 5 shows the percentage of the virtual data applied into the training versus the translation performance and phrase table size. The translation output obtained using the baseline system is evaluated as 9.41% in the BLEU score, as presented in Table 2 system ID 1. The BLEU score increases with adding more virtual data into the training, until the peak is reached by adding 80% of the data. Then the curve falls down to the value that we receive using the system ID 2, where all data is applied. This observation tells us that the translation performance can be significantly improved by including the generated virtual corpus. However, this corpus contains noisy sentence pairs, and selecting clean sentence pairs helps on one side further improve the translation performance, and on the other side, shrink the training data and the phrase table size for efficiency.

3.5 Output Examples

In Table 4, we present examples of German sentences generated using the ro-en-de-80% system. The first column shows the compatibility score calculated using Equation 1 as a cost. Romanian and English sentences are obtained from the parallel corpus of SETimes. We translate the English text into German using the SMT system described in Section 3.1. The Romanian-German sentence pairs are ranked after their costs. Worse translations are filtered out based on the percentage of the number of sentence pairs to be used divided by the total number of sentence pairs in the generated data. From Table 4, we can see that the generated Romanian-German sentence pairs are good translations to each other, as the cost is low, such as in the first line. As the cost value increases, the translations get worse, such as in the last line, the German sentence contains English words ”wounded” and ”shootout” due to unknown words in the English-German translation system.

Table 5 shows two translation examples from the MT output. We list the source sentence in Romanian, single reference sentence in German, the translation output by the baseline system and the translation output by the system ro-en-de-80%. As can be seen from Table 5, the translation quality of source sentences is greatly improved using the system ro-en-de-80% over the baseline system. Translations of words and word orderings are more adequate using the system ro-en-de-80%.

4 Related Work

We introduced the idea of compatibility and generated large-scale bilingual resources through a third language. Currently, for MT application, there are

cost	Romanian	German	per English
12.4	Rezultatele nu sunt surprinzătoare.	Die Ergebnisse sind nicht überraschend.	The results are not surprising.
23.2	Acum situaţia este mai bună.	Jetzt ist es eine bessere Situation.	Now it's a better situation.
35.1	Luli nu a mai fost văzut de atunci.	Luli nicht gesehen, da.	Luli has not been seen since.
44.1	Un suspect a fost rănit în schimbul de focuri.	Ein Verdächtiger wurde in der shootout wounded.	One suspect was wounded in the shootout.

Table 4: Examples of generated German corpus using Romanian-English parallel data and English-German MT.

a)	source	Universităţile sunt centrele de putere ale generării cunoaşterii.
	reference	Universitäten sind Motoren der Erzeugung von Wissen.
	baseline	Die Universitäten und werden von der Gewinnung dürfte
	ro-en-de-80%	Hochschulen sind die Macht der Entstehung des Wissens.
b)	source	Televiziunea este sursa noasträ primarä de informare şi divertisment.
	reference	Das Fernsehen ist die primäre unsere Informations- und Unterhaltung .
	baseline	Fernsehen die primäre stellte unsere Informations- und Unterhaltung gefördert werden.
	ro-en-de-80%	Fernsehen ist unsere Hauptquelle von Information und Unterhaltung.

Table 5: Examples of translation output by the baseline system and by the ro-en-de-80% system.

two approaches relating to our work: self-training and translation via bridge languages. However, these approaches are different from ours. The former one has been mainly focused on data exploitation from the available bilingual information, while the linguistic resources from a third language has been seldom applied. The latter one has been focused more on correcting the existing word alignment and phrase models rather than discovering new word, phrase or even sentence level translations through bridge languages. Our approach can be considered as a self-training with bridge language. We generate, instead of explore or gather, parallel data via bridge language, and the linguistic knowledge between the source-bridge and bridge-target languages are applied to learn translations between source and target languages.

Callison-Burch (2002) presented a co-training method for SMT, the agreement of multiple translation systems is explored to find the best translation for re-training. We applied compatibility instead of agreement based approach, detailed description on the difference between compatibility and agreement is referred to Section 1 and Section 2.3. Ueffing et al. (2009) explored model adaptation methods to use the monolingual data from the source language, while their learning and application are constrained in a bilingual way without introducing any information from a third language.

Mann and Yarowsky (2001) presented a method to induce translation lexicon based on transduction models of cognate pairs via bridge language. The cognate string edit distance was applied instead of a general MT system, so that the vocabulary learning is limited to mostly European languages. For bridge or pivot languages in MT, Kumar et al. (2007) described a method to improve word alignment quality using multiple bridge languages. In (Wu and Wang, 2007) and (Habash and Hu, 2009) phrase translation tables are improved using the phrase tables obtained from pivot languages in different ways, and in (Eisele et al., 2008) a hybrid method combining RBMT and SMT systems is introduced to fill up the data gap for pivot translation. Cohn and Lapata (2007) presented a method to obtain more reliable translation estimates from small data sets using multi-parallel data. Different from the previous approaches, we work on a black-boxed translation system, which means generation of the virtual data can be performed on any kind of translation systems including rule based, statistical based or even human translation. The approach introduced in (Leusch et al., 2010) can combine the translation output of a test set produced by any pivot MTs per different languages, however the individual systems are not improved and novel training data is not exploited. Bertoldi et al. (2008) evaluated several methods of pivot languages but did apply the global corpus filtering i.e. compatibility measure to control the quality of data. Our purpose is not only to improve the translation quality but also to provide useful linguistic resources for other NLP tasks.

5 Conclusion and Future Work

Thousands of human languages are recognized in the world, and building up millions of translation systems between these language pairs suffers greatly on the scarce resource, such as parallel data. We introduced the idea of compatibility, where all languages can be mapped to the same semantic meanings so that transferring between representations can benefit from resources of other representations. Individual system and linguistic resources can be improved using the result of compatibility measure and the virtual corpus. For machine translation application, we generate virtual parallel data per bridge language, and re-compiling on this corpus improves our machine translation performance by more than 30% relatively.

Despite of encouraging results, this method can be further improved by applying more refined algorithms to measure the compatibility in MT. Other areas of NLP can also be explored based on the compatibility centric concept.

References

N. Bertoldi, M. Barbaiani, M. Federico, and R. Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *Proceedings of IWSLT*, pages 143–149, Honolulu, Hawaii, USA.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

C. Callison-Burch. 2002. Co-training for statistical machine translation. Master's thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

T. Cohn and M. Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of ACL*, Prague, Czech Republic, June.

Y. Deng, J. Xu, and Y. Gao. 2008. Phrase table training for precision and recall: What makes a good phrase and a good phrase pair? In *Proceedings of ACL*, Columbus, OH, June.

A. Eisele, C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen. 2008. Hybrid machine translation architectures within and beyond the euromatrix project. In *Proceedings of EAMT*, Hamburg, Germany.

N. Habash and J. Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of SMT workshop*, Athens, Greece, March.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, Prague, Czech Republic, June.

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket Island, Thailand, September.

S. Kumar, F. Och, and W. Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of EMNLP*, Prague, Czech Republic, June.

G. Leusch, A. Max, J. M. Crego, and H. Ney. 2010. Multi-pivot translation by system combination. In *Proceedings of IWSLT workshop*, Paris, France, December.

G. S. Mann and D. Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL*, NAACL '01, pages 1–8, Pittsburgh, Pennsylvania, June. Association for Computational Linguistics.

D. Munteanu, A. Fraser, and D. Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of HLT-NAACL*, Boston, May.

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

K. A. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelphia, July.

SETIMES. 2011. SETimes multilingual corpus home page. http://opus.lingfil.uu.se/SETIMES.php.

J. R. Smith, C. Quirk, and K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of HLT-NAACL*, pages 403–411, Los Angeles, June.

SMT. 2011. Sixth workshop on statistical machine translation home page. http://www.statmt.org/wmt11/.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The JRCAcquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, Genova, Italy, May.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of InterSpeech*, pages 901–904, Denver, Colorado, September.

N. Ueffing, G. Haffari, and A. Sarkar. 2009. Semi-supervised learning for machine translation. In *Learning machine translation*, pages 237–256, Pittsburgh, Pennsylvania, February. The MIT Press.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING*, pages 836–841, Copenhagen, Denmark, August.

H. Wu and H. Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21:165–181, September.