Pivot Lightly-Supervised Training for Statistical Machine Translation

Matthias Huck and **Hermann Ney**
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{huck,ney}@cs.rwth-aachen.de

Abstract

In this paper, we investigate large-scale lightly-supervised training with a pivot language: We augment a baseline statistical machine translation (SMT) system that has been trained on human-generated parallel training corpora with large amounts of additional unsupervised parallel data; but instead of creating this synthetic data from monolingual source language data with the baseline system itself, or from target language data with a reverse system, we employ a parallel corpus of target language data and data in a pivot language. The pivot language data is automatically translated into the source language, resulting in a trilingual corpus with unsupervised source language side. We augment our baseline system with the unsupervised source-target parallel data.

Experiments are conducted for the German-French language pair using the standard WMT newstest sets for development and testing. We obtain the unsupervised data by translating the English side of the English-French 10^9 corpus to German. With careful system design, we are able to achieve improvements of up to +0.4 points BLEU / -0.7 points TER over the baseline.

1 Introduction

Pivot language approaches for statistical machine translation are typically applied in scenarios where no bilingual resources to build a translation system between a source and a target language exist. For many under-resourced language pairs, no human-generated parallel data of source and target texts is available. There may however still be bitexts at hand between both the source language and a third *pivot* language as well as the same pivot language and the target language. Pivot translation employs such bitext to bridge from source to target across the pivot language, thus effectively providing source-to-target translation (Utiyama and Isahara, 2007; Wu and Wang, 2009). The method may also be advantageous in cases where many translation systems between a large number of languages are to be built. To save time and cost, it may be convenient to resort to a pivot approach and to set up $2(n-1)$ systems between each of $n-1$ languages and a common pivot language, instead of setting up $n(n-1)$ systems between all pairs of n languages (Koehn et al., 2009).

We utilize the pivot translation paradigm with a different motivation in this work. We investigate a source-target language combination—German and French—that does not suffer from a lack of resources. A noticeable amount of parallel training data and monolingual target language data exists for this language pair, from which we build a well-performing German→French baseline SMT system. We however argue that our system could be improved if we were able to also deploy the extensive amount of parallel resources of both of these languages with third languages, in particular with English. English-German and English-French parallel corpora may contain additional information that is not present in the German-French data. We take a pivot lightly-supervised training approach to make our German→French setup learn from English-German and English-French resources. From English-German corpora, we set up

an English→German translation system. We run this system on the English side of a large English-French parallel corpus. The German output of this translation step and the French side of the English-French parallel corpus constitute an unsupervised bitext which can be used as supplementary training material for our German→French system.

2 Related Work

Our method combines techniques from two topics: pivot translation and lightly-supervised training.

Many pivot translation approaches have been proposed in the past. The vast literature on pivot translation cannot be discussed in detail here due to space contraints. Wu and Wang (2009) and Utiyama and Isahara (2007) provide good overviews of the field. More recent publications are e.g. (Cettolo et al., 2011), (Leusch et al., 2010) and (Koehn et al., 2009), to mention some. The *synthetic method* (Wu and Wang, 2009) comes closest to what is done by us. We would like to particularly point to the work by Cohn and Lapata (2007) and by Callison-Burch et al. (2006). Cohn et al. adopt the *pivot translation by triangulation* method to improve existing baselines. This idea is quite similar to our pivot lightly-supervised training approach, which employs synthetic data. Callison-Burch et al. suggest an interesting paraphrasing technique for SMT that rests upon parallel data with a pivot language. The effect should de facto be comparable.

Ueffing et al. (2007) introduced semi-supervised learning methods for the effective use of monolingual data in order to improve translation quality of SMT systems. Large-scale lightly-supervised training for SMT as we define it in this paper has been first carried out by Schwenk (2008). Schwenk translates a large amount of monolingual French data with an initial Moses (Koehn et al., 2007) baseline system into English. He uses the resulting unsupervised bitexts as additional training corpora to improve the baseline French→English system. With lightly-supervised training, Schwenk achieves improvements of around one point BLEU over the baseline. In a later work (Schwenk and Senellart, 2009) he applies the same method for translation model adaptation on an Arabic→French task with gains of up to 3.5 points BLEU. Li et al. (2011) present an approach that is very similar to lightly-supervised training. They conduct their experiments with a hierarchical phrase-based system and translate monolingual target language data into the source language. Lambert et al. (2011) investigate a large variety of lightly-supervised training settings on the French-English language pair in both directions. They draw some interesting conclusions, in particular that it is better to add automatically translated texts to the translation model training data which have been translated from the target to the source language (instead of from the source to the target language), and that using the word alignments that are produced by the decoder during the generation of the unsupervised data and using GIZA++ (Och and Ney, 2003) word alignments performs roughly equally well. Lambert et al. also propose to make use of an automatically contructed dictionary which provides unobserved morphological forms of nouns, verbs or adjectives. They achieve a gain of about 0.5 points BLEU over a competitive baseline.

Another technique we employ in our experiments is the combination of multiple phrase tables. Combining multiple phrase tables has e.g. been investigated for domain adaptation by Foster and Kuhn (2007) and Koehn and Schroeder (2007) before. Huck et al. (2011) combine phrase tables from human-generated data and from unsupervised data in the context of lightly-supervised training.

3 Lightly-Supervised Training

In previous lightly-supervised training scenarios, the baseline source-to-target SMT system is being augmented with additional unsupervised parallel data that is produced by automatically translating either source language monolingual data to the target language or target language monolingual data to the source language. The former is typically done with the baseline system itself, the latter with a reverse ("target-to-source") system. The reverse system can naturally only be trained on the same pre-existing parallel resources between source and target as the source-to-target baseline system. The language model data is not only composed of the respective side of the parallel resources and thus differs, though. The method does in fact go by the name of *lightly-supervised training* because the top-

ics that are covered in the monolingual corpora that are being translated may potentially also be covered by parts of the language model training data of the system which is used to translate them. This can be considered as a form of light supervision.[1]

The standard purpose of lightly-supervised training is adaptation: With sufficient amounts of reliable in-domain monolingual data, either in source or target language, a generic or out-of-domain baseline system can be trained towards aspects of topic and style of the domain under consideration. Extracting a translation model from the baseline parallel data plus the new unsupervised data results in new phrases (due to the phrase segmentation, choice of translation options and reordering performed by the system that is used for the production of the unsupervised data) and modified scores for phrases that have already been available before (due to the different number of occurrences). The vocabulary size remains unchanged. In the work of Schwenk (2008), a crucial ingredient of the lightly-supervised training pipeline is consequently the integration of a large supplementary bilingual dictionary with a high coverage, including morphological variants. The lightly-supervised training procedure enables the acquisition of phrases that contain words from this dictionary, where the input words would be out-of-vocabulary otherwise, and to learn reliable translation costs for phrase table entries which originate from the dictionary.

4 Pivot Lightly-Supervised Training

Pivot lightly-supervised training borrows the idea of improving an existing system with additional unsupervised parallel data from previous lightly-supervised training approaches. In contrast to these, the unsupervised data does not originate from monolingual data, but from parallel corpora of either source or target language with a pivot language. The pivot language data is being translated. This in turn resembles the synthetic method in pivot translation approaches. Existing pivot translation approaches however do not aim at improving systems, but at creating new ones from scratch for under-resourced language pairs.

A precondition for being able to perform pivot lightly-supervised training is the availability of a rich amount of multilingual data. A parallel corpus between source and target language is required in order to train the baseline system. We need parallel data between source or target (in our experiments: target) language and a pivot language which is used to produce the unsupervised data by automatically translating its pivot language side to target or source, respectively (in our experiments: source). The translation of pivot language data is done with a system that is trained on parallel data between pivot language and the language under consideration for the side of the unsupervised data that needs to be created automatically. Let's assume that a human-generated target-pivot corpus is at hand, as it is the case in our experiments. The pivot data then has to be translated into the source language. We thus need pivot-source human-generated parallel data to train a system that can conduct this translation.

Just as lightly-supervised training, pivot lightly-supervised training may serve the purpose of adaptation. Adaptation is not its main goal, though. In our experiments, we will even be able to show that pivot lightly-supervised training yields improvements in a setting where the standard lightly-supervised training approach is not effective. This can easily be investigated empirically by doing a comparison with lightly-supervised training on the non-pivot side of the parallel corpus which is used for the creation of the unsupervised data. The crucial key to the effectiveness of an incorporation of synthetic source-target training data that results from translation of pivot data into the source language is the pivot→source translation system, i.e. mainly the pivot-source parallel data it is trained with. Standard lightly-supervised training without pivot language can merely benefit from high-quality monolingual resources to refine the phrase translation model. In the pivot approach, translation options and vocabulary of the source language with a pivot language can be bridged via the unsupervised data to the target language.[2] By adding the bridged unsupervised bi-

[1] In this paper, we loosely apply the term *lightly-supervised training* if we mean the process of utilizing a machine translation system to produce additional bitexts that are used as training data, but still refer to the automatically produced bilingual corpora as *unsupervised data*.

[2] Or in general: also with "source" and "target" interchanged in this statement. We restrict our presentation to the variant we

	French	German
Sentences	2.0M	
Running Words	53.1M	45.8M
Vocabulary	145.0K	380.4K

Table 1: Corpus statistics of the preprocessed parallel training data of the French→German setup. In the data, numerical quantities have been replaced by a single category symbol. Note that no compound splitting has been applied to the German target-side data.

	English	French
Sentences	17.4M	
Running Words	484.4M	573.8M
Vocabulary	1.4M	1.4M

Table 2: Corpus statistics of the preprocessed English-French WMT 10^9 data. Some noisy parts of the raw corpus have been removed beforehand. In the data, numerical quantities have been replaced by a single category symbol.

	English	German
Sentences	1.9M	
Running Words	50.6M	48.4M
Vocabulary	123.5K	387.6K

Table 3: Corpus statistics of the preprocessed parallel training data of the English→German setup. In the data, numerical quantities have been replaced by a single category symbol. Note that no compound splitting has been applied to the German target-side data.

texts, the source→target system does not only learn from the contents of the corpus the unsupervised data originates from, but also from bilingual information that is represented in the translation model of the pivot→source system.

5 Parallel Resources

We now specify the parallel training corpora we utilize for an empirical evaluation of pivot lightly-supervised training on a German→French translation task. The pivot language is English.

To train the German→French baseline system, we use 2.0M sentence pairs that are partly taken from the Europarl corpus (Koehn, 2005) and have partly been collected within the Quaero project.[3] Statistics of the preprocessed data can be found in the *direct* entry of Table 6 (first three lines). The preprocessing pipeline includes splitting of German compound words with the frequency-based method described in (Koehn and Knight, 2003). We apply compound splitting to German text whenever German is the source language, but not for setups where German is the target language.

The unsupervised data is produced by translating the English side of the English-French 10^9 corpus as provided for the translation task of the Workshop on Statistical Machine Translation (WMT).[4] Data statistics are given in Table 2. Some noisy parts of the raw corpus have been removed beforehand by means of an SVM classifier in a fashion comparable to the filtering technique described by Herrmann et al. (2011).

The English→German SMT system with which we translate the English side of the 10^9 corpus to German is trained with the English-German parallel resources that have been provided for the 2011 WMT shared translation task (constrained track). Statistics of the preprocessed corpus are given in Table 3.

6 Phrase-Based Translation System

We apply a phrase-based translation (PBT) system which is an in-house implementation of the state-of-the-art decoder as described by Zens and Ney (2008). A standard set of models is used, comprising phrase translation probabilities and lexical translation probabilities in both directions, word and phrase penalty, a distance-based distortion model, an n-gram target language model and three simple count-based binary features. Parameter weights are optimized with the downhill simplex algorithm (Nelder and Mead, 1965) on the word graph.

The language models in all our setups are 4-grams with modified Kneser-Ney smoothing and are trained with the SRILM toolkit (Stolcke, 2002) on large collections of monolingual data.

Word alignments are produced with GIZA++.[5]

tried in our experiments.

[3] http://www.quaero.org

[4] http://www.statmt.org/wmt12/translation-task.html. The 10^9 corpus is often also referred to as *WMT Giga French-English release 2*.

[5] http://code.google.com/p/giza-pp/

French→German	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT direct	15.8	69.8	15.1	70.2	15.4	68.1	15.0	70.2

Table 4: Translation performance of the French→German system (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

English→German	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT direct	14.7	68.7	14.7	68.3	15.8	64.7	14.8	67.6

Table 5: Translation performance of the English→German system (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

We train alignments in both directions and symmetrize them according to the refined method that was suggested by Och and Ney (2003).

7 Experiments

In our experiments, we work with the standard WMT newstest sets. These sets are multi-parallel corpora. Each of the sets exists in a version in each of the three languages that are of relevance to us: German, French, English. We employ newstest2009 as development set in all setups; newstest2008, newstest2010 and newstest2011 are held-out sets and used for testing. We evaluate in truecase with the BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) measures on a single reference translation.

7.1 Systems for Producing Unsupervised Data

We first measure the translation performance of two direct translation systems, a French→German system that we run on the French 10^9 data to produce pivot unsupervised data, and an English→German system that we run on the English 10^9 data to produce data for lightly-supervised training without pivot language for comparison with the pivot approach.

French→German The French→German system is based on the parallel data from Table 1. Translation results are shown in Table 4.

English→German The English→German system is based on the parallel data from Table 3. Translation results are shown in Table 5.

7.2 Human-Generated and Unsupervised Training Corpora

Table 6 contains statistics for the following German-French corpora:

direct The human-generated parallel data.

unsup. (non-pivot) The non-pivot unsupervised data produced with the French→German system.

unsup. (pivot) The pivot unsupervised data produced with the English→German system.

unsup. (non-pivot) + direct A concatenation of non-pivot unsupervised data and human generated parallel data.

unsup. (pivot) + direct A concatenation of pivot unsupervised data and human generated parallel data.

On the automatically generated German data, we indicate the overall number of running words, but also the number of running words without unknowns. Unknowns are words that result from source-side words being out-of-vocabulary to the translation system. These are carried over to the target side by means of an identity mapping, but are marked in a special way. We keep them when we extract phrases from the unsupervised data, but do not allow for the usage of phrase table entries that contain unknowns in search. We remove such entries from the phrase table. The German vocabulary of the system with which the unsupervised data is created is an upper bound for the vocabulary size without unknowns on the German side of the unsupervised data.

		German	French
direct	Sentences	2.0M	
	Running Words	47.3M	53.1M
	Vocabulary	196.3K	145.0K
unsup. (non-pivot)	Sentences	17.4M	
	Running Words	494.3M	573.8M
	Running Words (w/o Unknowns)	478.8M	–
	Vocabulary	1.3M	1.4M
	Vocabulary (w/o Unknowns)	123.0K	–
unsup. (pivot)	Sentences	17.4M	
	Running Words	450.9M	573.8M
	Running Words (w/o Unknowns)	434.2M	–
	Vocabulary	1.3M	1.4M
	Vocabulary (w/o Unknowns)	128.6K	–
unsup. (non-pivot) + direct	Sentences	19.4M	
	Running Words	541.6M	626.9M
	Running Words (w/o Unknowns)	526.1M	–
	Vocabulary	1.4M	1.4M
	Vocabulary (w/o Unknowns)	201.1K	–
unsup. (pivot) + direct	Sentences	19.4M	
	Running Words	498.2M	626.9M
	Running Words (w/o Unknowns)	481.5M	–
	Vocabulary	1.4M	1.4M
	Vocabulary (w/o Unknowns)	210.8K	–

Table 6: Corpus statistics of the preprocessed German-French (direct and unsupervised) parallel training data. In the data, numerical quantities have been replaced by a single category symbol. German compound words have been split.

We train word alignments with GIZA++. Reusing the alignment of the unsupervised data given by the translation systems is not convenient for us because of a practical reason: We have to apply compound splitting on the unsupervised German data. German is going to be on source side in the systems that make use of the unsupervised data, and German compound splitting on source side typically improves the translation quality. We thus apply the compound splitting after having created the data and word-align the compound-split unsupervised German data with the corresponding French data from the 10^9 corpus. Note that the corpus statistics in Table 6 have been calculated after compound splitting has been applied.

7.3 German→French Experimental Results

We are now in a position to examine German→French translation quality based on human-generated training data, lightly-supervised training and pivot lightly-supervised training. The experimental results are presented in Table 7.

PBT direct The German→French baseline system is trained with the human-generated parallel data.

PBT transfer (En. intermediate) This setup applies the *pivot translation by transfer* scheme (Wu and Wang, 2009), which we additionally want to compare to. We set up direct systems for German→English and English→French translation in order to be able to conduct German→English→French transfer pivoting with English as intermediate language. The German→English translation system is trained on the data from Table 3, but with compound splitting on the German side. Its translation performance is indicated in Table 8. The English→French translation system is trained

German→French	newstest2008 BLEU	newstest2008 TER	newstest2009 BLEU	newstest2009 TER	newstest2010 BLEU	newstest2010 TER	newstest2011 BLEU	newstest2011 TER
PBT direct	19.2	66.6	18.9	66.6	20.3	65.9	19.6	65.6
PBT transfer (En. intermediate)	17.8	67.6	17.3	67.8	19.0	66.7	18.2	66.3
PBT unsup. (non-pivot)	17.4	69.3	16.7	70.0	17.8	69.2	17.8	68.3
PBT unsup. (pivot)	17.6	69.1	17.0	69.8	18.3	68.8	17.9	68.3
PBT unsup. (non-pivot) + direct								
— joint extraction	18.8	66.8	17.7	67.4	19.5	66.3	18.8	65.8
— joint extraction, direct lex.	18.8	66.6	17.9	67.2	19.7	66.0	19.1	65.5
— two phrase tables, direct lex.	19.3	67.7	18.8	67.7	19.7	67.0	19.6	66.5
PBT unsup. (pivot) + direct								
— joint extraction	18.7	67.0	18.2	67.5	19.6	66.5	18.9	66.1
— joint extraction, direct lex.	19.3	67.2	18.7	67.6	19.8	66.7	19.4	66.3
— two phrase tables, direct lex.	**19.4**	**66.2**	**19.0**	**66.3**	**20.7**	**65.3**	**19.9**	**65.0**

Table 7: Results for the German→French task (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

on the 10^9 data (Table 2). Its translation performance is indicated in Table 9. We translate from German to a single-best English intermediate hypothesis, which we feed into the English→French system to obtain a French output. Results are 1.3-1.6 points BLEU worse than direct translation.

PBT unsup. (non-pivot) This system is trained on the *unsup. (non-pivot)* corpus only, not on any human-generated data. Training a system on unsupervised parallel data only that has been automatically translated from a target-side monolingual corpus, results are 1.8-2.5 points BLEU worse than direct translation.

PBT unsup. (pivot) This system is trained on the *unsup. (pivot)* corpus only, not on any human-generated data. It resembles the *synthetic pivot translation* scheme (Wu and Wang, 2009). Results are 1.6-2.0 points BLEU worse than direct translation.

PBT unsup. (non-pivot) + direct These systems are based on lightly-supervised training without pivoting. They make use of both the baseline human-generated data and the unsupervised parallel corpus that has been automatically translated from target-side data.

PBT unsup. (pivot) + direct These systems are based on pivot lightly-supervised training. They make use of both the baseline human-generated data and the unsupervised parallel corpus that has been automatically translated from pivot language data.

Three different settings have been tried for both the lightly-supervised and the pivot lightly-supervised approach. The word-based lexicon model used for phrase table smoothing and separate phrase tables for human-generated and unsupervised data have proven crucial for translation quality here.

joint extraction A single phrase table is extracted from the concatenation of unsupervised and human-generated data. Lexical scores are computed with a lexicon model which is likewise extracted from the word-aligned concatenated data.

joint extraction, direct lex. A single phrase table is extracted from the concatenation of unsupervised and human-generated data. Lexical scores are computed with a lexicon model which is extracted from the word-aligned human-generated data only.

two phrase tables, direct lex. Two separate phrase tables from the baseline human-generated data and from the unsupervised data are extracted and utilized by the decoder. On both of the phrase tables, lexical scores are computed with a lexicon model which is extracted from the word-aligned human-generated data only.

German→English	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT direct	21.4	63.3	21.2	62.4	23.1	60.7	21.0	62.5

Table 8: Translation performance of the German→English system (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

English→French	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT direct	23.1	62.8	25.5	59.2	27.1	56.6	29.5	53.6

Table 9: Translation performance of the English→French system (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage. The translation model of the system was trained with the 10^9 corpus.

OOV [%] with training data	newstest2008		newstest2009		newstest2010		newstest2011	
	German	French	German	French	German	French	German	French
direct	2.7	2.1	2.7	2.4	2.9	2.4	3.1	2.7
unsup. (non-pivot)	3.4	0.9	3.4	1.0	3.7	1.1	4.1	1.1
unsup. (pivot)	3.4	0.9	3.3	1.0	3.6	1.1	3.9	1.1
unsup. (non-pivot) + direct	2.7	0.9	2.7	0.9	2.9	1.1	3.1	1.0
unsup. (pivot) + direct	2.6	0.9	2.5	0.9	2.8	1.1	3.0	1.0

Table 10: Out-of-vocabulary (OOV) rates of the development and test sets with the vocabulary of each of the preprocessed German-French (direct and unsupervised) parallel training data settings. In the data, numerical quantities have been replaced by a single category symbol.

	entries	distinct source sides	avg. number of candidates
PBT direct	12.1M	198.3K	61
PBT unsup. (non-pivot)	24.7M	257.5K	96
PBT unsup. (pivot)	32.9M	245.4K	134
PBT unsup. (non-pivot) + direct	28.5M	274.1K	104
PBT unsup. (pivot) + direct	36.2M	266.6K	136

Table 11: Phrase table statistics for the German→French setups. All phrase tables have been filtered towards the German side of the four newstest sets and pruned to contain a maximum of 400 distinct translation candidates per source side.

OOV [%] with filtered phrase voc.	newstest2008	newstest2009	newstest2010	newstest2011
	French	French	French	French
PBT direct	2.6	2.9	3.0	3.1
PBT unsup. (non-pivot)	2.6	3.0	2.9	3.1
PBT unsup. (pivot)	1.8	2.0	2.0	2.0
PBT unsup. (non-pivot) + direct	2.5	2.8	2.8	3.0
PBT unsup. (pivot) + direct	1.7	1.9	1.9	2.0

Table 12: Out-of-vocabulary (OOV) rates of the French references of the development and test sets, measured with regard to the target side vocabulary of those phrase tables entries that can actually be used for the translation of each of the sets.

The best results are obtained with the third of these settings. In the third setting, lightly-supervised training without pivoting is in terms of BLEU exactly on the level of the *PBT direct* baseline system, but in terms of TER clearly worse. With pivot lightly-supervised training, we are able to outperform the baseline by up to +0.4 points BLEU / -0.7 points TER.

8 Analysis

An adaptation effect towards the domain of the newstest corpora by means of the unsupervised data from the 10^9 collection does not seem to exist, according to our (negative) results with non-pivot lightly-supervised training. We tried to analyze why pivot lightly-supervised training still yields improvements.

Table 10 contains the out-of-vocabulary (OOV) rates of each of the newstest sets with regard to the vocabulary of the five training corpora from Table 6. The source-side OOV rates are barely reduced by adding unsupervised data. The target-side OOV rates are reduced considerably, but these numbers are overly optimistic as most of the words will correspond to unknowns on the source side. To obtain more insightful numbers, we had a look into the phrase tables of our systems. We filtered the phrase tables towards the German side of the four newstest sets and determined the total number of entries, the number of distinct source sides and the average number of candidates per source side. Note that our phrase tables are pruned to contain a maximum of 400 distinct translation candidates per source side. The phrase table statistics are presented in Table 11. Interestingly, the number of distinct source sides is slightly smaller with pivot lightly-supervised training than with non-pivot lightly-supervised training. The average number of translation candidates per source side is on the contrary about one third larger. This indicates that bilingual information that is represented in the translation model of the pivot→source system is in fact carried over to the source-target system via pivot lightly-supervised training. The richer choice of translation options pays off during search. Also, the target-side vocabulary that can actually be generated by the decoder is larger with pivot lightly-supervised training. To assess this, we filtered each phrase table towards the German side of each of the newstest sets individually. We then collected the French vocabulary present on the French side of the entries in each filtered phrase table and computed target-side OOV rates with respect to these filtered phrase vocabularies. The numbers are given in Table 12. The rates are roughly one third lower for pivot lightly-supervised training than for non-pivot lightly-supervised training and for the baseline.

9 Conclusion

We showed how a well-performing state-of-the-art SMT system can be improved by means of pivot lightly-supervised training. Pivot lightly-supervised training carries information which is present in additional resources that are parallel in source (or alternatively target) language and a third *pivot* language over to the source→target translation system. This is done via automatic generation of unsupervised source-target data. Gains in translation quality can even be achieved without a domain adaptation effect as in non-pivot lightly-supervised training.

Acknowledgments

This work was partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical machine translation using paraphrases. In *Proc. of the HLT-NAACL*, pages 17–24, New York City, New York, USA, June.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2011. Bootstrapping Arabic-Italian SMT through Comparable Texts and Pivot Translation. In *Proc. of the EAMT*, pages 249–256, Leuven, Belgium, May.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proc. of the ACL*, pages 728–735, Prague, Czech Republic, June.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June.

Teresa Herrmann, Mohammed Mediani, Jan Niehues, and Alex Waibel. 2011. The Karlsruhe Institute of Technology Translation Systems for the WMT 2011. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 379–385, Edinburgh, Scotland, UK, July.

Matthias Huck, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *Proc. of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, pages 91–96, Edinburgh, Scotland, UK, July.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proc. of the EACL*, pages 187–194, Budapest, Hungary, April.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic, June.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the ACL*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 Machine Translation Systems for Europe. In *Proc. of the MT Summit XII*, pages 65–72, Ottawa, Ontario, Canada, August.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the MT Summit X*, Phuket, Thailand, September.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on Translation Model Adaptation Using Monolingual Data. In *Proc. of the Sixth Workshop on Statistical Machine Translation (WMT)*, pages 284–293, Edinburgh, Scotland, UK, July.

Gregor Leusch, Aurélien Max, Josep Maria Crego, and Hermann Ney. 2010. Multi-Pivot Translation by System Combination. In *Proc. of the IWSLT*, pages 299–306, Paris, France, December.

Zhifei Li, Jason Eisner, Ziyuan Wang, Sanjeev Khudanpur, and Brian Roark. 2011. Minimum Imputed Risk: Unsupervised Discriminative Training for Machine Translation. In *Proc. of the EMNLP*, pages 920–929, Edinburgh, Scotland, UK, July.

John A. Nelder and Roger Mead. 1965. A Simplex Method for Function Minimization. *The Computer Journal*, 7:308–313.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Holger Schwenk and Jean Senellart. 2009. Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. In *Proc. of the MT Summit XII*, Ottawa, Ontario, Canada, August.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proc. of the IWSLT*, pages 182–189, Waikiki, Hawaii, USA, October.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of the AMTA*, pages 223–231, Cambridge, Massachusetts, USA, August.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, pages 901–904, September.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proc. of the ACL*, pages 25–32, Prague, Czech Republic, June.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. of the HLT-NAACL*, pages 484–491, Rochester, New York, USA.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proc. of the 47th Annual Meeting of the ACL and the 4th International Joint Conf. on Natural Language Processing of the AFNLP*, pages 154–162, August.

Richard Zens and Hermann Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *Proc. of the IWSLT*, pages 195–205, Waikiki, Hawaii, USA, October.