Character-Based Pivot Translation for Under-Resourced Languages and Domains

Jörg Tiedemann
Department of Linguistics and Philology
Uppsala University, Uppsala/Sweden
jorg.tiedemann@lingfil.uu.se

Abstract

In this paper we investigate the use of character-level translation models to support the translation from and to under-resourced languages and textual domains via closely related pivot languages. Our experiments show that these low-level models can be successful even with tiny amounts of training data. We test the approach on movie subtitles for three language pairs and legal texts for another language pair in a domain adaptation task. Our pivot translations outperform the baselines by a large margin.

1 Introduction

Data-driven approaches have been extremely successful in most areas of natural language processing (NLP) and can be considered the main paradigm in application-oriented research and development. Research in machine translation is a typical example with the dominance of statistical models over the last decade. This is even enforced due to the availability of toolboxes such as Moses (Koehn et al., 2007) which make it possible to build translation engines within days or even hours for any language pair provided that appropriate training data is available. However, this reliance on training data is also the most severe limitation of statistical approaches. Resources in large quantities are only available for a few languages and domains. In the case of SMT, the dilemma is even more apparent as parallel corpora are rare and usually quite sparse. Some languages can be considered lucky, for example, because of political situations that lead to the production of freely available translated material on a large scale. A lot of research and development would not have been possible without the European Union and its language policies to give an example.

One of the main challenges of current NLP research is to port data-driven techniques to under-resourced languages, which refers to the majority of the world's languages. One obvious approach is to create appropriate data resources even for those languages in order to enable the use of similar techniques designed for *high-density* languages. However, this is usually too expensive and often impossible with the quantities needed. Another idea is to develop new models that can work with (much) less data but still make use of resources and techniques developed for other well-resourced languages.

In this paper, we explore pivot translation techniques for the translation from and to resource-poor languages with the help of intermediate resource-rich languages. We explore the fact that many poorly resourced languages are closely related to well equipped languages, which enables low-level techniques such as character-based translation. We can show that these techniques can boost the performance enormously, tested for several language pairs. Furthermore, we show that pivoting can also be used to overcome data sparseness in specific domains. Even high density languages are under-resourced in most textual domains and pivoting via in-domain data of another language can help to adapt statistical models. In our experiments, we observe that related languages have the largest impact in such a setup.

The remaining parts of the paper are organized as follows: First we describe the pivot translation approach used in this study. Thereafter, we dis-

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 141–151,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

cuss character-based translation models followed by a detailed presentation of our experimental results. Finally, we briefly summarize related work and conclude the paper with discussions and prospects for future work.

2 Pivot Models

Information from pivot languages can be incorporated in SMT models in various ways. The main principle refers to the combination of source-to-pivot and pivot-to-target translation models. In our setup, one of these models includes a resource-poor language (source or target) and the other one refers to a standard model with appropriate data resources. A condition is that we have at least some training data for the translation between pivot and the resource-poor language. However, for the original task (source-to-target translation) we do not require any data resources except for purposes of comparison.

We will explore various models for the translation between the resource-poor language and the pivot language and most of them are not compatible with standard phrase-based translation models. Hence, triangulation methods (Cohn and Lapata, 2007) for combining phrase tables are not applicable in our case. Instead, we explore a cascaded approach (also called "transfer method" (Wu and Wang, 2009)) in which we translate the input text in two steps using a linear interpolation for rescoring N-best lists. Following the method described in (Utiyama and Isahara, 2007) and (Wu and Wang, 2009), we use the best n hypotheses from the translation of source sentences s to pivot sentences p and combine them with the top m hypotheses for translating these pivot sentences to target sentences t:

$$\hat{t} \approx \underset{t}{argmax} \sum_{k=1}^{L} \alpha \lambda_k^{sp} h_k^{sp}(s,p) + (1-\alpha)\lambda_k^{pt} h_k^{pt}(p,t)$$

where h_k^{xy} are feature functions for model xy with appropriate weights λ_k^{xy}.[1] Basically, this means that we simply add the scores and, similar to related work, we assume that the feature weights can be set independently for each model using minimum error rate training (MERT) (Och, 2003). In our setup we added the parameter α that can be used to weight the importance of one model over the other. This can be useful as we do not consider the entire hypothesis space but only a small subset of N-best lists. In the simplest case, this weight is set to 0.5 making both models equally important. An alternative to fitting the interpolation weight would be to perform a global optimization procedure. However, a straightforward implementation of pivot-based MERT would be prohibitively slow due to the expensive two-step translation procedure over n-best lists.

A general condition for the pivot approach is to assume independent training sets for both translation models as already pointed out by (Bertoldi et al., 2008). In contrast to research presented in related work (see, for example, (Koehn et al., 2009)) this condition is met in our setup in which all data sets represent different samples over the languages considered (see section 4).[2]

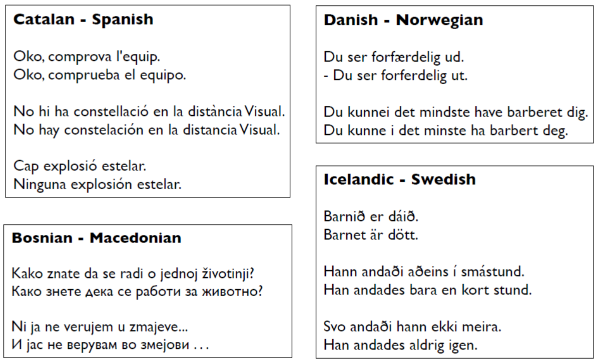
3 Character-Based SMT

The basic idea behind character-based translation models is to take advantage of the strong lexical and syntactic similarities between closely related languages. Consider, for example, Figure 1. Related languages like Catalan and Spanish or Danish and Norwegian have common roots and, therefore, use similar concepts and express them in similar grammatical structures. Spelling conventions can still be quite different but those differences are often very consistent. The Bosnian-Macedonian example also shows that we do not have to require any alphabetic overlap in order to obtain character-level similarities.

Regularities between such closely related languages can be captured below the word level. We can also assume a more or less monotonic relation between the two languages which motivates the idea of translation models over character N-grams treating translation as a transliteration task (Vilar et al., 2007). Conceptually it is straightforward to think of phrase-based models on the character level. Sequences of characters can be used instead of word N-grams for both, translation and language models. Training can proceed with the same tools and approaches. The basic task is to

[1] Note, that we do not require the same feature functions in both models even though the formula above implies this for simplicity of representation.

[2] Note that different samples may still include common sentences.



Catalan - Spanish

Oko, comprova l'equip.
Oko, comprueba el equipo.

No hi ha constellació en la distància Visual.
No hay constelación en la distancia Visual.

Cap explosió estelar.
Ninguna explosión estelar.

Bosnian - Macedonian

Kako znate da se radi o jednoj životinji?
Како знете дека се работи за животно?

Ni ja ne verujem u zmajeve...
И јас не верувам во змејови . . .

Danish - Norwegian

Du ser forfærdelig ud.
- Du ser forferdelig ut.

Du kunnei det mindste have barberet dig.
Du kunne i det minste ha barbert deg.

Icelandic - Swedish

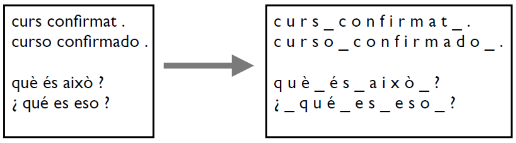
Barnið er dáið.
Barnet är dött.

Hann andaði aðeins í smástund.
Han andades bara en kort stund.

Svo andaði hann ekki meira.
Han andades aldrig igen.

Figure 1: Some examples of movie subtitle translations between closely related languages (either sharing parts of the same alphabet or not).

prepare the data to comply with the training procedures (see Figure 2).



curs confirmat .
curso confirmado .

què és això ?
¿ qué es eso ?

c u r s _ c o n f i r m a t _ .
c u r s o _ c o n f i r m a d o _ .

q u è _ é s _ a i x ò _ ?
¿ _ q u é _ e s _ e s o _ ?

Figure 2: Data pre-processing for training models on the character level. Spaces are represented by '_' and each sentence is treated as one sequence of characters.

3.1 Character Alignment

One crucial difference is the alignment of characters, which is required instead of an alignment of words. Clearly, the traditional IBM word alignment models are not designed for this task especially with respect to distortion. However, the same generative story can still be applied in general. Vilar et al. (2007) explore a two-step procedure where words are aligned first (with the traditional IBM models) to divide sentence pairs into aligned segments of reasonable size and the characters are then aligned with the same algorithm.

An alternative is to use models designed for transliteration or related character-level transformation tasks. Many approaches are based on transducer models that resemble string edit operations such as insertions, deletions and substitutions (Ristad and Yianilos, 1998). Weighted finite state transducers (WFST's) can be trained on unaligned pairs of character sequences and have been shown to be very effective for transliteration tasks or letter-to-phoneme conversions (Jiampojamarn et al., 2007). The training procedure usually employs an expectation maximization (EM) procedure and the resulting transducer can be used to find the Viterbi alignment between characters according to the best sequence of edit operations applied to transform one string into the other. Extensions to this model are possible, for example the use of many-to-many alignments which have been shown to be very effective in letter-to-phoneme alignment tasks (Jiampojamarn et al., 2007).

One advantage of the edit-distance-based transducer models is that the alignments they predict are strictly monotonic and cannot easily be confused by spurious relations between characters over longer distances. Long distance alignments are only possible in connection with a series of insertions and deletions that usually increase the alignment costs in such a way that they are avoided if possible. On the other hand, IBM word alignment models also prefer monotonic alignments over non-monotonic ones if there is no good reason to do otherwise (i.e., there is frequent evidence of distorted alignments). However, the size of the vocabulary in a character-level model is very small (several orders of magnitude smaller than on the word level) and this may cause serious confusion of the word alignment model that very much relies on context-independent lexical translation probabilities. Hence, for character alignment, the lexical evidence is much less reliable without their context.

It is certainly possible to find a compromise between word-level and character-level models in order to generalize below word boundaries but avoiding alignment problems as discussed above. Morpheme-based translation models have been explored in several studies with similar motivations as in our approach, a better generalization from sparse training data (Fishel and Kirik, 2010; Luong et al., 2010). However, these approaches have the drawback that they require proper morphological analyses. Data-driven techniques exist even for morphology, but their use in SMT still needs to be shown (Fishel, 2009). The situation is comparable to the problems of integrating linguistically motivated phrases into phrase-based SMT (Koehn et al., 2003). Instead we opt for a more general approach to extend context to facilitate, especially, the alignment step. Figure 3 shows how we can transform texts into sequences of bigrams that can be aligned with standard approaches without making any assumptions about linguistically motivated segmentations.

cu ur rs so o_ _c co on nf fi ir rm ma ad do o_ _. ._

¿_ _q qu ué é_ _e es s_ _e es so o_ _? ?_

Figure 3: Two Spanish sentences as sequences of character bigrams with a final '_' marking the end of a sentence.

In this way we can construct a parallel corpus with slightly richer contextual information as input to the alignment program. The vocabulary remains small (for example, 1267 bigrams in the case of Spanish compared to 84 individual characters in our experiments) but lexical translation probabilities become now much more differentiated.

With this, it is now possible to use the alignment between bigrams to train a character-level translation system as we have the same number of bigrams as we have characters (and the first character in each bigram corresponds to the character at that position). Certainly, it is also possible to train a bigram translation model (and language model). This has the (one and only) advantage that one character of context across phrase boundaries (i.e. character N-grams) is used in the selection of translation alternatives from the phrase table.[3]

3.2 Tuning Character-Level Models

A final remark on training character-based SMT models is concerned with feature weight tuning. It certainly makes not much sense to compute character-level BLEU scores for tuning feature weights especially with the standard settings of matching relatively short N-grams. Instead we would still like to measure performance in terms of word-level BLEU scores (or any other MT evaluation metric used in minimum error rate training). Therefore, it is important to post-process character-translated development sets before adjusting weights. This is simply done by merging characters accordingly and replacing the place-holders with spaces again. Thereafter, MERT can run as usual.

3.3 Evaluation

Character-level translations can be evaluated in the same way as other translation hypotheses, for example using automatic measures such as BLEU, NIST, METEOR etc. The same simple post-processing as mentioned in the previous section can be applied to turn the character translations into "normal" text. However, it can be useful to look at some other measures as well that consider near matches on the character level instead of matching words and word N-grams only. Character-level models have the ability to produce strings that may be close to the reference and still do not match any of the words contained. They may generate non-words that include mistakes which look like spelling-errors or minor grammatical mistakes. Those words are usually close enough to the correct target words to be recognized by the user, which is often more acceptable than leaving foreign words untranslated. This is especially true as many unknown words represent important content words that bear a lot of information. The problem of unknown words is even more severe for morphologically rich language as many word forms are simply not part of (sparse) training data sets. Untranslated words are especially annoying when translating languages that use different writing systems. Consider, for example, the following subtitles in Macedonian (using Cyrillic letters) that have been translated from Bosnian (written in Latin characters):

reference:	И чаша вино, како и секогаш.
word-based:	И **čašu vina**, како секогаш.
char-based:	И **чаша вино**, како секогаш.
reference:	Во староto светилиште.
word-based:	Во **starom svetilištu**.
char-based:	Во **стар светилиштето**.

The underlined parts mark examples of character-level differences with respect to the reference translation. For the pivot translation approach, it is important that the translations generated in the first step can be handled by the second one. This means, that words generated by a character-based model should at least be valid input words for the second step, even though they might refer to erroneous inflections in that context. Therefore, we add another measure to our experimental results presented below – the number of unknown words with respect to the input language of the second step. This applies only to models that are used as the first step in pivot-based translations. For other models, we include a string similarity measure based on the longest common subsequence ratio (LCSR) (Stephen, 1992) in order to give an impression about the "closeness" of the system

[3] Using larger units (trigrams, for example) led to lower scores in our experiments (probably due to data sparseness) and, therefore, are not reported here.

output to the reference translations.

4 Experiments

We conducted a series of experiments to test the ideas of (character-level) pivot translation for resource-poor languages. We chose to use data from a collection of translated subtitles compiled in the freely available OPUS corpus (Tiedemann, 2009b). This collection includes a large variety of languages and contains mainly short sentences and sentence fragments, which suits character-level alignment very well. The selected settings represent translation tasks between languages (and domains) for which only very limited training data is available or none at all.

Below we present results from two general tasks:[4] (i) Translating between English and a resource-poor language (in both directions) via a pivot language that is close related to the resource-poor language. (ii) Translating between two languages in a domain for which no in-domain training data is available via a pivot language with in-domain data. We will start with the presentation of the first task and the character-based translation between closely related languages.

4.1 Task 1: Pivoting via Related Languages

We decided to look at resource-poor languages from two language families: Macedonian representing a Slavic language from the Balkan region, Catalan and Galician representing two Romance languages spoken mainly in Spain. There is only little or no data available for translating from or to English for these languages. However, there are related languages with medium or large amounts of training data. For Macedonian, we use Bulgarian (which also uses a Cyrillic alphabet) and Bosnian (another related language that mainly uses Latin characters) as the pivot language. For Catalan and Galician, the obvious choice was Spanish (however, Portuguese would, for example, have been another reasonable option for Galician). Table 1 lists the data available for training the various models. Furthermore, we reserved 2000 sentences for tuning parameters and another 2000 sentences for testing. For Galician, we only used 1000 sentences for each set due to the lack of additional data. We were especially careful when preparing the data to exclude all sentences from tuning and test sets that could be found in any pivot or direct translation model. Hence, all test sentences are unseen strings for all models presented in this paper (but they are not comparable with each other as they are sampled individually from independent data sets).

language pair	*#sent's*	*#words*
Galician – English	–	–
Galician – Spanish	2k	15k
Catalan – English	50k	400k
Catalan – Spanish	64k	500k
Spanish – English	30M	180M
Macedonian – English	220k	1.2M
Macedonian – Bosnian	12k	60k
Macedonian – Bulgarian	155k	800k
Bosnian – English	2.1M	11M
Bulgarian – English	14M	80M

Table 1: Training data for the translation task between closely related languages in the domain of movie subtitles. Number of sentences (#sent's) and number of words (#words) in thousands (k) and millions (M) (averages of source and target language).

The data sets represent several interesting test cases: Galician is the least supported language with extremely little training data for building our pivot model. There is no data for the direct model and, therefore, no explicit baseline for this task. There is 30 times more data available for Catalan-English, but still too little for a decent standard SMT model. Interesting here is that we have more or less the same amount of data available for the baseline and for the pivot translation between the related languages. The data set for Macedonian – English is by far the largest among the baseline models and also bigger than the sets available for the related pivot languages. Especially Macedonian – Bosnian is not well supported. The interesting questions is whether tiny amounts of pivot data can still be competitive. In all three cases, there is much more data available for the translation models between English and the pivot language.

In the following section we will look at the translation between related languages with various models and training setups before we consider the actual translation task via the bridge languages.

[4] In all experiments we use standard tools like Moses, Giza++, SRILM, mteval etc. Details about basic settings are omitted here due to space constraints but can be found in the supplementary material. The data sets are available from here: http://stp.lingfil.uu.se/∼joerg/index.php?resources

	bs-mk		bg-mk		es-gl		es-ca	
Model	BLEU %	↑LCSR	BLEU %	↑LCSR	BLEU %	↑LCSR	BLEU %	↑LCSR
word-based	15.43	0.5067	14.66	0.6225	41.11	0.7966	62.73	0.8526
char – WFST$_{1:1}$	21.37^{++}	0.6903	13.33^{--}	0.6159	36.94	0.7832	73.17^{++}	0.8728
char – WFST$_{2:2}$	19.17^{++}	0.6737	12.67^{--}	0.6190	43.39^{++}	0.8083	70.64^{++}	0.8684
char – IBM$_{char}$	23.17^{++}	0.6968	14.57	0.6347	**45.21**$^{++}$	**0.8171**	73.12^{++}	0.8767
char – IBM$_{bigram}$	**24.84**$^{++}$	**0.7046**	**15.01**$^{++}$	**0.6374**	44.06^{++}	0.8144	**74.21**$^{++}$	**0.8803**

Table 2: Translating from a related pivot language to the target language. Bosnian (bs) / Bulgarian (bg) – Macedonian (mk); Galician (gl) / Catalan (ca) – Spanish (es). *Word-based* refers to standard phrase-based SMT models. All other models use phrases over character sequences. The *WFST*$_{x:y}$ models use weighted finite state transducers for character alignment with units that are at most x and y characters long, respectively. Other models use Viterbi alignments created by IBM model 4 using GIZA++ (Och and Ney, 2003) between characters (*IBM*$_{char}$) or bigrams (*IBM*$_{bigram}$). LCSR refers to the averaged longest common subsequence ratio between system translations and references. Results are significantly better ($p < 0.01^{++}$, $p < 0.05^{+}$) or worse ($p < 0.01^{--}$, $p < 0.05^{-}$) than the word-based baseline.

	mk-bs		mk-bg		gl-es		ca-es	
Model	BLEU %	↓UNK	BLEU %	↓UNK	BLEU %	↓UNK	BLEU %	↓UNK
word-based	14.22	17.83%	14.77	5.29%	43.22	10.18%	59.34	3.80%
char – WFST$_{1:1}$	21.74^{++}	1.50%	16.04^{++}	**0.77**%	50.24^{++}	**1.17**%	62.87^{++}	**0.45**%
char – WFST$_{2:2}$	19.19^{++}	2.05%	15.32	0.96%	50.59^{++}	1.28%	59.84	0.47%
char – IBM$_{char}$	24.15^{++}	1.30%	17.12^{++}	0.80%	**51.18**$^{++}$	1.38%	64.35 $^{++}$	0.59%
char – IBM$_{bigram}$	**24.82**$^{++}$	**1.00**%	**17.28**$^{++}$	**0.77**%	50.70^{++}	1.36%	**65.14**$^{++}$	0.48%

Table 3: Translating from the source language to a related pivot language. UNK gives the proportion of unknown words with respect to the translation model from the pivot language to English.

4.1.1 Translating Related Languages

The main challenge for the translation models between related languages is the restriction to very limited parallel training data. Character-level models make it possible to generalize to very basic translation units leading to robust models in the sense of models without unknown events. The basic question is whether they provide reasonable translations with respect to given accepted references. Tables 2 and 3 give a comprehensive summary of various models for the languages selected in our experiments.

We can see that at least one character-based translation model outperforms the standard word-based model in all cases. This is true (and not very surprising) for the language pairs with very little training data but it is also the case for language pairs with slightly more reasonable data sets like Bulgarian-Macedonian. The automatic measures indicate decent translation performances at this stage which encourages their use in pivot translation that we will discuss in the next section.

Furthermore, we can also see the influence of different character alignment algorithms. Somewhat surprisingly, the best results are achieved with IBM alignment models that are not designed for this purpose. Transducer-based alignments produce consistently worse translation models (at least in terms of BLEU scores). The reason for this might be that the IBM models can handle noise in the training data more robustly. However, in terms of unknown words, WFST-based alignment is very competitive and often the best choice (but not much different from the best IBM based models). The use of character bigrams leads to further BLEU improvements for all data sets except Galician-Spanish. However, this data set is extremely small, which may cause unpredictable results. In any case, the differences between character-based alignments and bigram-based ones are rather small and our experiments do not lead to conclusive results.

4.1.2 Pivot Translation

In this section we now look at cascaded translations via the related pivot language. Tables 4 and 5 summarize the results for various settings.

As we can see, the pivot translations for Catalan and Galician outperform the baselines by a large margin. Here, the baselines are, of course, very weak due to the minimal amount of training data. Furthermore, the Catalan-English test set appears to be very easy considering the relatively high BLEU scores achieved even with tiny

Model (BLEU in %)	1x1	10x10
English – Catalan (baseline)	26.70	
English – (Spanish = Catalan)	8.38	
English – Spanish -word- Catalan	38.91^{++}	39.59^{++}
English – Spanish -char- Catalan	44.46^{++}	$\mathbf{46.82}^{++}$
Catalan – English (baseline)	27.86	
(Catalan = Spanish) – English	9.52	
Catalan -word- Spanish – English	38.41^{++}	38.65^{++}
Catalan -char- Spanish – English	40.43^{++}	$\mathbf{40.73}^{++}$
English – Galician (baseline)	—	
English – (Spanish = Galician)	7.46	
English – Spanish -word- Galician	20.55	20.76
English – Spanish -char- Galician	**21.12**	21.09
Galician – English (baseline)	—	
(Galician = Spanish) – English	5.76	
Galician -word- Spanish – English	13.16	13.20
Galician -char- Spanish – English	**16.04**	16.02

Table 4: Translating between Galician/Catalan and English via Spanish using a standard phrase-based SMT baseline, Spanish–English SMT models to translate from/to Catalan/Galician and pivot-based approaches using word-level models or character-level models (based on IBM_{bigram} alignments) with either one-best (1x1) or N-best lists (10x10 with $\alpha = 0.85$).

Model (BLEU in %)	1x1	10x10
English – Maced. (baseline)	11.04	
English – Bosn. -word- Maced.	7.33^{--}	7.64
English – Bosn. -char- Maced.	9.99	10.34
English – Bulg. -word- Maced.	12.49^{++}	$\mathbf{12.62}^{++}$
English – Bulg. -char- Maced.	11.57^{++}	11.59^{+}
Maced. – English (baseline)	20.24	
Maced. -word- Bosn. – English	12.36^{--}	12.48^{--}
Maced. -char- Bosn. – English	18.73^{-}	18.64^{--}
Maced. -word- Bulg. – English	19.62	19.74
Maced. -char- Bulg. – English	21.05	**21.10**

Table 5: Translating between Macedonian (Maced) and English via Bosnian (Bosn) / Bulgarian (Bulg).

amounts of training data for the baseline. Still, no test sentence appears in any training or development set for either direct translation or pivot models. From the results, we can also see that Catalan and Galician are quite different from Spanish and require language-specific treatment. Using a large Spanish – English model (with over 30% BLEU in both directions) to translate from or to Catalan or Galician is not an option. The experiments show that character-based pivot models lead to better translations than word-based pivot models (in terms of BLEU scores). This reflects the performance gains presented in Table 2. Rescoring of N-best lists, on the other hand, does not have a big impact on our results. However, we did not spend time optimizing the parameters of N-best size and interpolation weight.

The results from the Macedonian task are not as clear. This is especially due to the different setup in which the baseline uses more training data than any of the related language pivot models. However, we can still see that the pivot translation via Bulgarian clearly outperforms the baseline. For the case of translating to Macedonian via Bulgarian, the word-based model seems to be more robust than the character-level model. This may be due to a larger number of non-words generated by the character-based pivot model. In general, the BLEU scores are much lower for all models involved (even for the high-density languages), which indicates larger problems with the generation of correct output and intermediate translations.

Interesting is the fact that we can achieve almost the same performance as the baseline when translating via Bosnian even though we had much less training data at our disposal for the translation between Macedonian and Bosnian. In this setup, we can see that a character-based model was necessary in order to obtain the desired abstraction from the tiny amount of training data.

4.2 Task 2: Pivoting for Domain Adaptation

Sparse resources are not only a problem for specific languages but also for specific domains. SMT models are very sensitive to domain shifts and domain-specific data is often rare. In the following, we investigate a test case of translating between two languages (English and Norwegian) with reasonable amounts of data resources but in the wrong domain (movie subtitles instead of legal texts). Here again, we facilitate the translation process by a pivot language, this time with domain-specific data.

The task is to translate legal texts from Norwegian (Bokmål) to English and vice versa. The test set is taken from the English–Norwegian Parallel Corpus (ENPC) (Johansson et al., 1996) and contains 1493 parallel sentences (a selection of European treaties, directives and agreements). Otherwise, there is no training data available in this domain for English and Norwegian. Table 6 lists the other data resources we used in our study.

As we can see, there is decent amount of training data for English – Norwegian, but the domain is strikingly different. On the other hand, there

Language pair	*Domain*	#sent's	#words
English–Norwegian	subtitles	2.4M	18M
Norwegian–Danish	subtitles	1.5M	10M
Danish–English	DGT-TM	430k	9M

Table 6: Training data available for the domain adaptation task. DGT-TM refers to the translation memories provided by the JRC (Steinberger et al., 2006)

is in-domain data for other languages like Danish that may act as an intermediate pivot. Furthermore, we have out-of-domain data for the translation between pivot and Norwegian. The sizes of the training data sets for the pivot models are comparable (in terms of words). The in-domain pivot data is controlled and very consistent and, therefore, high quality translations can be expected. The subtitle data is noisy and includes various movie genres. It is important to mention that the pivot data still does not contain any sentence included in the English–Norwegian test set.

Table 7 summarizes the results of our experiments when using Danish and in-domain data as a pivot in translations from and to Norwegian.

Model (task: English – Norwegian)	BLEU
(step 1) English –dgt– Danish	52.76
(step 2) Danish –subs$_{wo}$– Norwegian	29.87
(step 2) Danish –subs$_{ch}$– Norwegian	29.65
(step 2) Danish –subs$_{bi}$– Norwegian	25.65
English –subs– Norwegian (baseline)	7.20
English –dgt– (Danish = Norwegian)	9.44^{++}
English –dgt– Danish -subs$_{wo}$- Norwegian	17.49^{++}
English –dgt– Danish -subs$_{ch}$- Norwegian	**17.61**$^{++}$
English –dgt– Danish -subs$_{bi}$- Norwegian	14.07^{++}

Model (task: Norwegian – English)	BLEU
(step 1) Norwegian –subs$_{wo}$– Danish	30.15
(step 1) Norwegian –subs$_{ch}$– Danish	27.81
(step 1) Norwegian –subs$_{bi}$– Danish	28.52
(step 2) Danish –dgt– English	57.23
Norwegian –subs– English (baseline)	11.41
(Norwegian = Danish) –dgt– English	13.21^{++}
Norwegian –subs+dgtLM– English	13.33^{++}
Norwegian –subs$_{wo}$– Danish –dgt– English	25.75^{++}
(Norwegian –subs$_{ch}$– Danish –dgt– English	23.77^{++}
Norwegian –subs$_{bi}$– Danish –dgt– English	**26.29**$^{++}$

Table 7: Translating out-of-domain data via Danish. Models using in-domain data are marked with *dgt* and out-of-domain models are marked with *subs*. *subs+dgtLM* refers to a model with an out-of-domain translation model and an added in-domain language model. The subscripts *wo*, *ch* and *bi* refer to word, character and bigram models, respectively.

The influence of in-domain data in the translation process is enormous. As expected, the out-of-domain baseline does not perform well even though it uses the largest amount of training data in our setup. It is even outperformed by the in-domain pivot model when pretending that Norwegian is in fact Danish. For the translation into English, the in-domain language model helps a little bit (similar resources are not available for the other direction). However, having the strong in-domain model for translating to (and from) the pivot language improves the scores dramatically. The out-of-domain model in the other part of the cascaded translation does not destroy this advantage completely and the overall score is much higher than any other baseline.

In our setup, we used again a closely related language as a pivot. However, this time we had more data available for training the pivot translation model. Naturally, the advantages of the character-level approach diminishes and the word-level model becomes a better alternative. However, there can still be a good reason for the use of a character-based model as we can see in the success of the bigram model (–subs$_{bi}$–) in the translation from Norwegian to English (via Danish). A character-based model may generalize beyond domain-specific terminology which leads to a reduction of unknown words when applied to a new domain. Note that using a character-based model in step two could possibly cause more harm than using it in step one of the pivot-based procedure. Using n-best lists for a subsequent word-based translation in step two may fix errors caused by character-based translation simply by ignoring hypotheses containing them, which makes such a model more robust to noisy input.

Finally, as an alternative, we can also look at other pivot languages. The domain adaptation task is not at all restricted to closely related pivot languages especially considering the success of word-based models in the experiments above. Table 8 lists results for three other pivot languages.

Surprisingly, the results are much worse than for the Danish test case. Apparently, these models are strongly influenced by the out-of-domain translation between Norwegian and the pivot language. The only success can be seen with another closely related language, Swedish. Lexical and syntactic similarity seems to be important to create models that are robust enough for domain shifts in the cascaded translation setup.

Pivot=xx	*en–xx*	*xx–no*	*en–xx–no*
German	53.09	23.60	3.15^{--}
French	66.47	17.84	5.03^{--}
Swedish	52.62	24.79	**10.07**$^{++}$
Pivot=xx	*no–xx*	*xx–en*	*no–xx–en*
German	15.02	53.02	5.52^{--}
French	17.69	65.85	8.78^{--}
Swedish	19.72	59.55	**16.35**$^{++}$

Table 8: Alternative word-based pivot translations between Norwegian (no) and English (en).

5 Related Work

There is a wide range of pivot language approaches to machine translation and a number of strategies have been proposed. One of them is often called *triangulation* and usually refers to the combination of phrase tables (Cohn and Lapata, 2007). Phrase translation probabilities are merged and lexical weights are estimated by bridging word alignment models (Wu and Wang, 2007; Bertoldi et al., 2008). Cascaded translation via pivot languages are discussed by (Utiyama and Isahara, 2007) and are frequently used by various researchers (de Gispert and Mariño, 2006; Koehn et al., 2009; Wu and Wang, 2009) and commercial systems such as Google Translate. A third strategy is to generate or augment data sets with the help of pivot models. This is, for example, explored by (de Gispert and Mariño, 2006) and (Wu and Wang, 2009) (who call it the *synthetic method*). Pivoting has also been used for paraphrasing and lexical adaptation (Bannard and Callison-Burch, 2005; Crego et al., 2010). (Nakov and Ng, 2009) investigate pivot languages for resource-poor languages (but only when translating from the resource-poor language). They also use transliteration for adapting models to a new (related) language. Character-level SMT has been used for transliteration (Matthews, 2007; Tiedemann and Nabende, 2009) and also for the translation between closely related languages (Vilar et al., 2007; Tiedemann, 2009a).

6 Conclusions and Discussion

In this paper, we have discussed possibilities to translate via pivot languages on the character level. These models are useful to support under-resourced languages and explore strong lexical and syntactic similarities between closely related languages. Such an approach makes it possible to train reasonable translation models even with extremely sparse data sets. Moreover, character level models introduce an abstraction that reduce the number of unknown words dramatically. In most cases, these unknown words represent information-rich units that bear large portions of the meaning to be translated. The following illustrates this effect on example translations with and without pivot model:

Example: Catalan – English (via Spanish)

Reference:	I have to grade these papers.
Baseline:	Tincque qualificar these exàmens.
$Pivot_{word}$:	Tincque qualificar these tests.
$Pivot_{char}$:	I have to grade these papers.

Example: Macedonian – English (via Bulgarian)

Reference:	It's a simple matter of self-preservation.
Baseline:	It's simply a question of себесочувување.
$Pivot_{word}$:	That's a matter of себесочувување.
$Pivot_{char}$:	It's just a question of yourself.

Leaving unseen words untranslated is not only annoying (especially if the input language uses a different writing system) but often makes translations completely incomprehensible. Pivot translations will still not be perfect (see example two above), but can at least be more intelligible. Character-based models can even take care of tokenization errors as the one shown above ("Tincque" should be two words "Tinc que"). Fortunately, the generation of non-word sequences (observed as unknown words) does not seem to be a big problem and no special treatment is required to avoid such output. We would still like to address this issue in future work by adding a word level LM in character-based SMT. However, (Vilar et al., 2007) already showed that this did not have any positive effect in their character-based system. In a second study, we also showed that pivot models can be useful for adapting to a new domain. The use of in-domain pivot data leads to systems that outperform out-of-domain translation models by a large margin. Our findings point to many prospects for future work. For example, we would like to investigate combinations of character-based and word-based models. Character-based models may also be used for treating unknown words only. Multiple source approaches via several pivots is another possibility to be explored. Finally, we also need to further investigate the robustness of the approach with respect to other language pairs, data sets and learning parameters.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-Based Statistical Machine Translation with Pivot Languages. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 143–149, Hawaii, USA.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic, June. Association for Computational Linguistics.

Josep Maria Crego, Aurélien Max, and François Yvon. 2010. Local lexical adaptation in machine translation through triangulation: SMT helping SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 232–240, Beijing, China, August. Coling 2010 Organizing Committee.

A. de Gispert and J.B. Mariño. 2006. Catalan-english statistical machine translation without parallel corpus: Bridging through spanish. In *Proceedings of the 5th Workshop on Strategies for developing Machine Translation for Minority Languages (SALTMIL'06) at LREC*, pages 65–68, Genova, Italy.

Mark Fishel and Harri Kirik. 2010. Linguistically motivated unsupervised segmentation for machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1741–1745, Valletta, Malta.

Mark Fishel. 2009. Deeper than words: Morph-based alignment for statistical machine translation. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics PacLing 2009*, Sapporo, Japan.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Stig Johansson, Jarle Ebeling, and Knut Hofland. 1996. Coding and aligning the English-Norwegian Parallel Corpus. In K. Aijmer, B. Altenberg, and M. Johansson, editors, *Languages in Contrast*, pages 87–112. Lund University Press.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. In *Proceedings of MT Summit XII*, pages 65–72, Ottawa, Canada.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 148–157, Cambridge, MA, October. Association for Computational Linguistics.

David Matthews. 2007. Machine transliteration of proper names. Master's thesis, School of Informatics, University of Edinburgh.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1367, Singapore, August. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Eric Sven Ristad and Peter N. Yianilos. 1998. Learning string edit distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 20(5):522–532, May.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of* *the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 2142–2147.

Graham A. Stephen. 1992. String Search. Technical report, School of Electronic Engineering Science, University College of North Wales, Gwynedd.

Jörg Tiedemann and Peter Nabende. 2009. Translating transliterations. *International Journal of Computing and ICT Research*, 3(1):33–41.

Jörg Tiedemann. 2009a. Character-based PSMT for closely related languages. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 12 – 19, Barcelona, Spain.

Jörg Tiedemann. 2009b. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April. Association for Computational Linguistics.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic, June. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic, June. Association for Computational Linguistics.

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