Analyzing the Use of Character-Level Translation with Sparse and Noisy Datasets

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

Preslav Nakov
Qatar Computing Research Institute
Qatar Foundation, P.O. box 5825
Doha, Qatar
pnakov@qf.org.qa

Abstract

This paper provides an analysis of character-level machine translation models used in pivot-based translation when applied to sparse and noisy datasets, such as crowdsourced movie subtitles. In our experiments, we find that such character-level models cut the number of untranslated words by over 40% and are especially competitive (improvements of 2-3 BLEU points) in the case of limited training data. We explore the impact of character alignment, phrase table filtering, bitext size and the choice of pivot language on translation quality. We further compare cascaded translation models to the use of synthetic training data via multiple pivots, and we find that the latter works significantly better. Finally, we demonstrate that neither word- nor character-BLEU correlate perfectly with human judgments, due to BLEU's sensitivity to length.

1 Introduction

Statistical machine translation (SMT) systems, which dominate the field of machine translation today, are easy to build and offer competitive performance in terms of translation quality. Unfortunately, training such systems requires large parallel corpora of sentences and their translations, called *bitexts*, which are not available for most language pairs and textual domains. As a result, building an SMT system to translate directly between two languages is often not possible. A common solution to this problem is to use an intermediate, or *pivot* language to bridge the gap in training such a system.

A typical approach is a *cascaded translation* model using two independent steps of translating from the source to the pivot and then from the pivot to the target language. A special case is where the pivot is closely related to the source language, which makes it possible to train useful systems on much smaller bitexts using *character-level translation* models. This is the case we will consider below, translating Macedonian to English via related languages, primarily Bulgarian.

Our main contribution is the further analysis of such a setup. We show that character-level models can cut the number of untranslated words almost by half since translation involves many transformations at the sub-word level. We further explore the impact of character alignment, phrase table pruning, data size, and choice of pivot language on the effectiveness of character-level SMT models. We also study the use of character-level translation for the generation of synthetic training data, which significantly outperforms all cascaded translation setups. Finally, we present a manual evaluation showing that neither word- nor character-BLEU correlate perfectly with human judgments.

The remainder of the paper is organized as follows: Section 2 presents related work. Section 3 discusses technical details about using character-level SMT models. Section 4 describes the experiments, and Section 5 discusses the results. Section 6 concludes with directions for future work.

2 Related Work

SMT using pivot languages has been studied for several years. Cohn and Lapata (2007) used *triangulation* techniques for the combination of phrase tables. The lexical weights in such an approach can be estimated by bridging word alignments (Wu and Wang, 2007; Bertoldi et al., 2008).

Proceedings of Recent Advances in Natural Language Processing, pages 676–684,
Hissar, Bulgaria, 7-13 September 2013.

Cascaded translation via pivot languages is used by various researchers (de Gispert and Mariño, 2006; Koehn et al., 2009; Wu and Wang, 2009). Several techniques are compared in (Utiyama and Isahara, 2007; de Gispert and Mariño, 2006; Wu and Wang, 2009). Pivot languages can also be used for paraphrasing and lexical adaptation (Bannard and Callison-Burch, 2005; Crego et al., 2010). None of this work exploits the similarity between the pivot and the source/target language.

The first step in our pivoting experiments involves SMT between closely related languages, which has been handled using word-for-word translation and manual rules for a number of language pairs, e.g., Czech–Slovak (Hajič et al., 2000), Turkish–Crimean Tatar (Altintas and Cicekli, 2002), Irish–Scottish Gaelic (Scannell, 2006), Cantonese–Mandarin (Zhang, 1998). In contrast, we explore statistical approaches that are potentially applicable to many language pairs.

Since we combine word- and character-level models, a relevant line of research is on combining SMT models of different granularity, e.g., Luong et al. (2010) combine word- and morpheme-level representations for English–Finnish. However, they did not assume similarity between the two languages, neither did they use pivoting.

Another relevant research combines bitexts between related languages with little or no adaptation (Nakov and Ng, 2009; Marujo et al., 2011; Wang et al., 2012; Nakov and Ng, 2012). However, that work did not use character-level models.

Character-level models were used for transliteration (Matthews, 2007; Tiedemann and Nabende, 2009) and for SMT between closely related languages (Vilar et al., 2007; Tiedemann, 2009a; Nakov and Tiedemann, 2012). Tiedemann (2012a) used pivoting with character-level SMT.

3 Character-level SMT Models

Closely related languages largely overlap in vocabulary and exhibit strong syntactic and lexical similarities. Most words have common roots and express concepts with similar linguistic constructions. Spelling conventions and morphology can still differ, but these differences are typically regular and thus can easily be generalized.

These similarities and regularities motivate the use of character-level SMT models, which can operate at the sub-word level, but also cover mappings spanning over words and multi-word units. Character-level SMT models, thus combine the generality of character-by-character transliteration and lexical mappings of larger units that could possibly refer to morphemes, words or phrases, to various combinations thereof.

One drawback of character-level models is their inability to model long-distance word reorderings. However, we do not assume very large syntactic differences between closely related languages. Another issue is that sentences become longer, which causes an overhead in decoding time.

In our experiments below, we use phrase-based SMT, treating characters as words, and using a special character for the original space character. Due to the reduced vocabulary, we can easily train models of higher order, thus capturing larger context and avoiding generating non-word sequences: we opted for models of order 10, both for the language model and for the maximal phrase length (normally, 5 and 7, respectively).

One difficulty is that training these models requires the alignment of characters in bitexts. Specialized character-level alignment algorithms do exist, e.g., those developed for character-to-phoneme translations (Damper et al., 2005; Jiampojamarn et al., 2007). However, Tiedemann (2012a) has demonstrated that standard tools for word alignment are in fact also very effective for character-level alignment, especially when extended with local context. Using character n-grams instead of single characters improves the expressive power of lexical translation parameters, which are one of the most important factors in standard word alignment models. For example, using character n-grams increases the vocabulary size of a 1.3M tokens-long Bulgarian text as follows: 101 single characters, 1,893 character bigrams, and 14,305 character trigrams; compared to 30,927 words. In our experiments, we explore the impact of increasing n-gram sizes on the final translation quality. We can confirm that bigrams perform best, constituting a good compromise between generality and contextual specificity.

Hence, we used GIZA++ (Och and Ney, 2003) to generate IBM model 4 alignments (Brown et al., 1993) for character n-grams, which we symmetrized using the *grow-diag-final-and* heuristics. We then converted the result to character alignments by dropping all characters behind the initial one. Finally, we used the Moses toolkit (Koehn et al., 2007) to build a character-level phrase table.

We tuned the parameters of the log-linear SMT model by optimizing BLEU (Papineni et al., 2002). Computing BLEU scores over character sequences does not make much sense, especially for small n-gram sizes (usually, $n \leq 4$). Therefore, we post-processed the character-level n-best lists in each tuning step to calculate word-level BLEU. Thus, we optimized word-level BLEU, while performing character-level translation.

4 Experiments and Evaluation

We used translated movie subtitles from the freely available OPUS corpus (Tiedemann, 2009b). The collection includes small amounts of parallel data for Macedonian-English (MK-EN), which we use as our test case. There is substantially more data for Bulgarian (BG), our main pivot language. For the translation between Macedonian and Bulgarian, there is even less data available. See Table 1.

dataset	# sentences	# words
MK-EN	160K	2.2M
MK-BG	102K	1.3M
BG-EN	10M	152M
MK-mono	536K	4M
BG-mono	16M	136M
EN-mono	43M	435M

Table 1: Size of the datasets.

The original data from OPUS is contributed by on-line users with little quality control and is thus quite noisy. Subtitles in OPUS are checked using automatic language identifiers and aligned using time information (Tiedemann, 2009b; Tiedemann, 2012b). However, we identified many misaligned files and, therefore, we realigned the corpus using `hunalign` (Varga et al., 2005). We also found several Bulgarian files misclassified as Macedonian and vice versa, which we addressed by filtering out any document pair for which the BLEU score exceeded 0.7 since it is likely to have large overlapping parts in the same language. We also filtered out sentence pairs where the Macedonian/Bulgarian side contained Bulgarian/Macedonian-specific letters.

From the remaining data we selected 10K sentence pairs (77K English words) for development and another 10K (72K English words) for testing; we used the rest for training. We used 10K pairs because subtitle sentences are short, and we wanted to make sure that the dev/test datasets contain enough words to enable stable tuning with MERT and reliable final evaluation results.

We further used the Macedonian–English and the Bulgarian–English movie subtitles datasets from OPUS, which we split into dev/test (10K sentence pairs for each) and train datasets. We made sure that the dev/test datasets for MK-BG, MK-EN and BG-EN do not overlap, and that all dev/test sentences were removed from the monolingual data used for language modeling.

Table 2 shows our baseline systems, trained using standard settings for a phrase-based SMT model: Kneser-Ney smoothed 5-gram language model and phrase pairs of maximum length seven.

Task	BLEU	NIST	TER	METEOR
MK-EN	22.33	5.47	63.57	39.19
MK-BG	30.70	6.52	50.94	70.44
BG-MK	28.01	6.24	51.98	69.89
BG-EN	37.60	7.34	47.41	58.89

Table 2: Phrase-based SMT baselines.

4.1 Translating Between Related Languages

We first investigate the impact of character alignment on the character-level translation between related languages. For this, we consider the extension of the context using character n-grams proposed by Tiedemann (2012a).

Another direction we explore is the possibility of reducing the noise in the phrase table. Treating even closely related languages by transliteration techniques is only a rough approximation to the translation task at hand. Furthermore, during training we observe many example translations that are not literally translated from one language to another. Hence, the character-level phrase table will be filled with many noisy and unintuitive translation options. We, therefore, applied phrase table pruning techniques based on relative entropy (Johnson et al., 2007) to remove unreliable pairs.

		PT Size		BLEU (%)	
n	align	std	fltd	std	fltd
1	2.5	5.1	1.0	30.47	31.13
2	2.6	5.2	0.9	30.87	31.32
3	2.9	6.9	0.9	30.32	31.03
4	3.0	10.4	1.1	29.76	30.42
5	3.1	12.0	1.2	29.25	30.19
6	3.2	10.8	1.2	28.81	29.68
7	3.4	8.1	1.1	28.73	29.73
8	3.5	6.4	1.0	28.40	29.45
9	3.6	5.4	0.9	27.67	29.19
10	3.6	5.1	0.9	27.11	28.78

Table 3: MK-BG character alignment points, phrase table sizes (in million of entries) and BLEU scores before (std) and after phrase filtering (fltd).

Table 3 shows the phrase table sizes for different settings and alignment approaches. We can see that, in all cases, the size of the filtered phrase tables is less than 20% of that of the original ones, which yields significant boost in decoding performance. More importantly, we see that filtering also leads to consistently better translation quality in all cases. This result is somewhat surprising for us: in our experience (for word-level models), filtering has typically harmed BLEU. Finally, we see that both with and without filtering, the best BLEU scores are achieved for $n = 2$.

The numbers in the table imply that the alignments become noisier for n-grams longer than two characters; look at the increasing number of phrases that can be extracted from the aligned corpus, many of which do not survive the filtering.

4.2 Bridging via Related Languages

Our next task is to use character-level models in the translation from under-resourced languages to other languages using the related language as a pivot. Several approaches for pivot-based translations have been proposed as discussed earlier.

We will look at two alternatives: (1) cascaded translations with two separate translation models and (2) bridging the gap by producing synthetic training corpora. For the latter, we automatically translate the related language in an existing training corpus to the under-resourced language.

Cascaded Pivot Translation

We base our translations on the individually trained translation models for the source (Macedonian) to the pivot language (Bulgarian) and for the pivot language to the final target language (English, in our case). As proposed by Tiedemann (2012a), we rerank k-best translations to find the best hypothesis for each given test sentence. For both translation steps, we set k to 10 and we require unique translations in the first step.

	BLEU	NIST	TER	METEOR
Model	individually tuned			
word-level pivot	22.48	5.46	64.11	47.77
char-based pivot	25.67	5.91	60.45	54.61
word+char+MK-EN	25.00	5.86	61.47	50.19
Model	globally tuned			
word-level pivot	23.38	5.44	64.33	48.31
char-based pivot	25.73	5.81	61.91	52.47
word+char+MK-EN	26.36	5.92	60.85	53.39

Table 4: Evaluating cascaded translation: Macedonian to English, pivoting via Bulgarian.

One possibility is to just apply the models tuned for the individual translation tasks, which is suboptimal. Therefore, we also introduce a global tuning approach, in which we generate k-best lists for the combined cascaded translation model and we tune corresponding end-to-end weights using MERT (Och, 2003) or PRO (Hopkins and May, 2011). We chose to set the size of the k-best lists to 20 in both steps to keep the size manageable, with 400 hypotheses for each tuning sentence.

Another option is to combine (i) the direct translation model, (ii) the word-level pivot model, and (iii) the character-level pivot model. Throwing them all in one k-best reranking system does not work well when using the unnormalized model scores. However, global tuning helps reassign weights such that the interactions between the various components can be covered. We use the same global tuning model introduced above using a combined system as the blackbox producing k-best lists and tuning feature weights for all components involved in the entire setup. Using the three translation paths, we obtain an extended set of parameters covering five individual systems. Since MERT is unstable with so many parameters, we use PRO. Note that tuning gets slow due to the extensive decoding that is necessary (five translation steps) and the increased size of the k-best lists (400 hypotheses for each pivot model and 100 hypotheses for the direct translation model).

Table 4 summarizes the results of the cascaded translation models. They all beat the baseline: direct translation from Macedonian to English. Note that the character-level model adds significantly to the performance of the cascaded model compared to the entirely word-level one. Furthermore, the scores illustrate that proper weights are important, especially for the case of the combined translation model. Without globally tuning its parameters, the performance is below the best single system, which is not entirely surprising.

Synthetic Training Data

Another possibility to make use of pivot languages is to create synthetic training data. For example, we can translate the Bulgarian side of our large Bulgarian–English training bitext to Macedonian, thus ending up with "Macedonian"-English training data. This is similar to previous work on adapting between closely related languages (Marujo et al., 2011; Wang et al., 2012), but here we perform translation rather than adaptation.

Model	BLEU	NIST	TER	METEOR
BG word-syn.	26.01	5.82	61.49	50.31
BG char-syn.	28.17	6.17	58.97	55.27
BG w+c-syn.	28.62	6.25	58.52	55.75
BG w+c-syn.+MK-EN	29.11	6.30	58.27	56.64
SR word-syn.	25.39	5.72	63.26	41.15
SR char-syn.	27.25	6.14	60.31	47.32
SR w+c-syn.	29.05	6.29	59.73	49.18
SR w+c-syn.+MK-EN	30.39	6.51	58.08	50.91
SL word-syn.	24.78	5.58	64.04	39.34
SL char-syn.	24.03	5.67	63.11	44.76
SL w+c-syn.	27.30	6.11	60.46	47.69
SL w+c-syn.+MK-EN	28.42	6.26	59.57	49.06
CZ word-syn.	26.48	5.83	62.52	41.02
CZ char-syn.	23.74	5.51	64.96	44.96
CZ w+c-syn.	28.03	6.08	61.12	48.41
CZ w+c-syn.+MK-EN	29.24	6.29	59.39	49.60
ALL-syn.	36.25	7.24	53.06	61.74
ALL-syn.+MK-EN	36.69	7.28	52.83	62.26

Table 5: Macedonian-English translation using synthetic data (by translating X-EN to MK-EN).

For translating from Bulgarian to Macedonian, we experimented with a word-level and a character-level SMT model. We also combined the two by concatenating the resulting MK-EN bitexts. We relied on the single best translation in all cases. Here, the character-level model was our best performing one, when using a filtered phrase table based on bigram alignments. Using k-best lists would be another option, but those should be properly weighted when combined to form a new synthetic training set. In future work, we plan to try the more sophisticated bitext combinations from (Nakov and Ng, 2009).

Table 5 shows the overall results. Experiments with word-level and character-level SMT are shown in rows 1 and 2, respectively. The result from the model trained on a concatenation of synthetic bitexts is shown in the third row. Finally, we also added the original MK-EN bitext to the combination (e.g., row 4). We can see from the top four rows that synthetic data outperforms cascaded translation by 2-3 BLEU points. Here again, the character-level model is much more valuable than the word-level one, which is most probably due to the reduction in the number of out-of-vocabulary (OOV) words it yields.

Another huge advantage over the cascaded approaches presented above is the reduced decoding time. Now, the system behaves like a traditional phrase-based SMT engine. The only large-scale effort is the translation of the training corpus, which only needs to be done once and can easily be performed off-line in a distributed setup.

4.3 Learning Curves and Other Languages

Observing the success of bridging via related pivot languages leads to at least two additional questions: (1) How much data is necessary for training reasonable character-level translation models that are still better than a standard word-level model trained on the same data? and (2) How strongly related should the languages be so that it is beneficial to use SMT at the character level?

Size of the Training Data

We investigated the first question by translating from Macedonian to Bulgarian with increasing amounts of training data. For comparability, we kept the model parameters fixed.

The top-left plot in Figure 1 shows the learning curves for word- and character-level models for MK-BG. We can see that the character-level models clearly outperform the word-level ones for the small amounts of training data that we have: the abstraction at the character level is much stronger and yields more robust models.

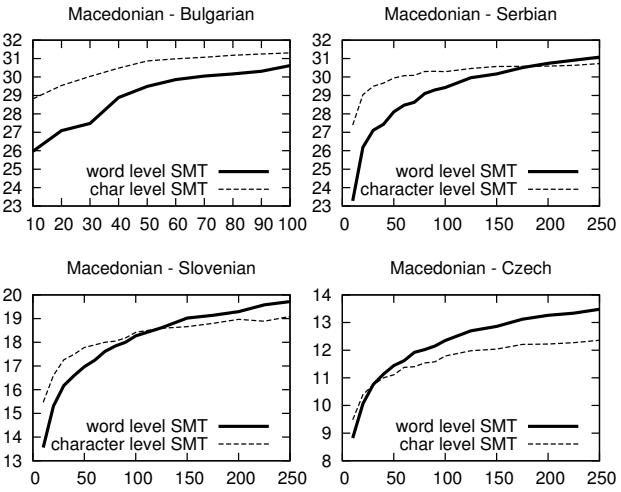

Figure 1: BLEU (in %) for word- and character-level SMT models with varying sizes of parallel training data (in thousands of sentence pairs).

Other Pivot Languages

We investigated the second question by experimenting with data from OPUS for two South-Slavic languages that are less related to Macedonian than Bulgarian. We selected Serbian and Slovenian from the Western group of the South-Slavic language branch (Bulgarian and Macedonian are in the Eastern group) from which Slovenian is the furthest away from Macedonian.

Note that while Bulgarian and Macedonian use Cyrillic, Slovenian and Serbian use the Latin alphabet (Serbian can also use Cyrillic, but not in OPUS). We have larger training datasets for the latter two: about 250-300 thousand sentence pairs.

To further contrast the relationship between South-Slavic languages (such as Bulgarian, Macedonian, Serbian, Slovenian) and languages from the Western-Slavic branch, we also experimented with Czech (about 270 thousand sentence pairs).

Note that we do not have the same movies available in all languages involved; therefore, the test and the development datasets are different for each language pair. However, we used the same amount of data in all setups: 10,000 sentence pairs for tuning and 10,000 pairs for evaluation.

Figure 1 also shows the learning curves for the three additional language pairs. For the South-Slavic languages, the character-level models make sense with sparse datasets. They outperform word-level models at least until around 100 thousand sentence pairs of training data.

Certainly, language relatedness has an impact on the effectiveness of character-level SMT. The difference between character- and word-level models shows that Slovenian is not the most appropriate choice for character-level SMT due to its weaker relation to Macedonian.

Furthermore, we can see that the performance of character-level models levels out at some point and standard word-level models surpass them with an almost linear increase in MT quality up to the point we have considered in the training procedures. Looking at Czech, we can see that character-level models are only competitive for very small datasets, but their performance is so low that they are practically useless. In general, translation is more difficult for more distant languages; this is also the case for word-level models.

Pivot-Based Translation

The final, and probably most important, question with respect to this paper is whether the other languages are still useful for pivot-based translation. Therefore, we generated translations of our Macedonian-English test set but this time via Serbian, Slovene and Czech. We used the approach that uses synthetic training data, which was the most successful one for Bulgarian, based on translation models trained on subsets of 100 thousand sentence pairs to make the results comparable with the Bulgarian case.

The pivot–English data that we have translated to "Macedonian"-English is comparable for Slovene and Serbian (1 million sentence pairs) and almost double the size for Czech (1.9 million).

Table 5 summarizes the results for all pivot languages: Bulgarian (BG), Serbian (SR), Slovenian (SL), and Czech (CZ). We can see that Serbian, which is geographically adjacent to Macedonian, performs almost as well as Bulgarian, which is also adjacent, while Slovenian, which is further away, and not adjacent to any of the above, performs worse. Note that with a Slovenian pivot, the character-level model performs worse than the word-level model.

This suggests that the differences between Slovenian and Macedonian are not that much at the sub-word level but mostly at the word level. This is even more evident for Czech, which is geographically further away and which is also from a different Slavic branch. Note, however, that we had much more data for Czech-English than for any other X-EN bitext, which explains the strong overall performance of its word-level model.

Overall, we have seen that as the relatedness between the source and the pivot language decreases, so does the utility of the character-level model. However, in all cases the character-level model helps when combined with the word-level one, yielding 1.5-4 BLEU points of improvement. Moreover, using all four pivots yields seven additional BLEU points over the best single pivot.

5 Discussion

Finally, we performed a manual evaluation of word-level, character-level and combined systems translating from Macedonian to English using Bulgarian. We asked three speakers of Macedonian and Bulgarian to rank the English output from the eight anonymized systems in Table 6, given the Macedonian input; we used 100 test sentences.

Model	word BLEU	char BLEU	Avg. rank	">" score	Untr. words
reference	—	—	1.57	0.73	—
baseline	22.33	50.83	3.37	0.25	4,959
word-pivot	23.38	53.26	2.81	0.42	3,144
char-pivot	25.73	56.00	2.51	0.52	1,841
comb-pivot	26.36	56.39	2.63	0.46	1,491
word-synth.	26.01	55.59	2.77	0.43	3,258
char-synth.	28.17	58.21	2.31	0.58	1,818
comb-synth.	28.62	58.53	2.11	0.65	1,712

Table 6: Comparing word- and character-level BLEU to human judgments for MK-EN using BG.

The results are shown in Table 6. Column 4 shows the *average rank* for each system, and column 5 shows the *">" score* as defined in (Callison-Burch et al., 2012): the frequency a given system was judged to be strictly better than the rest divided by the frequency it was judged strictly better or strictly worse than the rest. We further include word- and character-level BLEU, and the number of untranslated words.

We calculated Cohen's kappas (Cohen, 1960) of 0.87, 0.86 and 0.83 between the pairs of judges, following the procedure in (Callison-Burch et al., 2012). This corresponds to almost perfect agreement (Landis and Koch, 1977), probably due to the short length of subtitles, which allows for few differences in translation and simplifies ranking.

The individual human judgments (not shown to save space) correlate perfectly in terms of relative ranking of (a) the three pivoting systems and (b) the three synthetic data systems. Moreover, the individual and the overall human judgments also correlate well with the BLEU scores on (a) and (b), with one notable exception: humans ranked *char-pivot* higher than *comb-pivot*, while word- and char-BLEU switched their ranks. A closer investigation found that this is probably due to length: the hypothesis/reference ratio for *char-pivot* is 1.006, while for *comb-pivot* it is 1.016. In contrast, for *char-synth.* it is 1.006, while for *comb-synth.* it is 1.002. Recent work (Nakov et al., 2012) has shown that the closest this ratio gets to 1, the better the BLEU score is expected to be.

Note also that word-BLEU and char-BLEU correlate perfectly on (a) and (b), which is probably due to tuning the two systems for word-BLEU.

Interestingly, the BLEU-based rankings of the systems inside (a) and (b) perfectly correlate with the number of untranslated words. Note the robustness of the character-level models: they reduce the number of untranslated words by more than 40%. Having untranslated words in the final English translation could be annoying since they are in Cyrillic, but more importantly, they could contain information that is critical for a human, or even for the SMT system, without which it could not generate a good translation for the remaining words in the sentence. This is especially true for content-baring long, low-frequency words. For example, the inability to translate the Macedonian лутам ('I am angry') yields "You don' лутам." instead of "I'm not mad at you."

Character models are very robust with unknown morphological forms, e.g., a word-level model would not translate развеселам, yielding "I'm trying to make a развеселам.", while a character-level model will transform it to the Bulgarian развеселя, thus allowing the fluent "I'm trying to cheer you up." Note that this transformation does not necessarily have to pick the correct Bulgarian form, e.g., развеселя is a conjugated verb (1st person, singular, subjunctive), but it is translated as an infinitive, i.e., all Bulgarian conjugated forms would map to the same English infinitive.

Finally, it is also worth mentioning that character models are very robust in case of typos, concatenated or wrongly split words, which are quite common in movie subtitles.

6 Conclusion and Future Work

We have explored the use of character-level SMT models when applied to sparse and noisy datasets such as crowdsourced movie subtitles. We have demonstrated their utility when translating between closely related languages, where translation is often reduced to sub-word transformations. We have shown that such models are especially competitive in the case of limited training data (2-3 BLEU points of improvement, and 40% reduction of OOV), but fall behind word-level models as the training data increases. We have also shown the importance of phrase table filtering and the impact of character alignment on translation performance.

We have further experimented with bridging via a related language and we have found that generating synthetic training data works best. This makes it also straightforward to use multiple pivots and to combine word-level with character-level SMT models. Our best combined model outperforms the baseline by over 14 BLEU points, which represents a very significant boost in translation quality.

In future work, we would like to investigate the robustness of character-level models with respect to domain shifts and for other language pairs. We further plan a deeper analysis of the ability of character-level models to handle noisy inputs that include spelling errors and tokenization mistakes, which are common in user-generated content.

Acknowledgments

We would like to thank Petya Kirova and Veno Pacovski for helping us with the manual judgments. We also thank the anonymous reviewers for their constructive comments.

References

Kemal Altintas and Ilyas Cicekli. 2002. A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences*, ISCIS '02, pages 192–196, Orlando, FL.

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, MI.

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*, IWSLT '08, pages 143–149, Honolulu, HI.

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, WMT '12, pages 10–51, Montréal, Québec, Canada.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

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 for Computational Linguistics*, ACL '07, pages 728–735, Prague, Czech Republic.

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 '10, pages 232–240, Beijing, China.

Robert Damper, Yannick Marchand, John Marsters, and Alexander Bazin. 2005. Aligning text and phonemes for speech technology applications using an EM-like algorithm. *International Journal of Speech Technology*, 8(2):149–162.

Adriá de Gispert and José 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, pages 65–68, Genova, Italy.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, ANLP '00, pages 7–12, Seattle, WA.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1352–1362, Edinburgh, Scotland, United Kingdom.

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: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, NAACL '07, pages 372–379, Rochester, NY.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 967–975, Prague, Czech Republic.

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. Demonstration session*, ACL '07, pages 177–180, Prague, Czech Republic.

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

Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74.

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 7th Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 148–157, Cambridge, MA.

Luís Marujo, Nuno Grazina, Tiago Luís, Wang Ling, Luísa Coheur, and Isabel Trancoso. 2011. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, EAMT '11, pages 129–136, Leuven, Belgium.

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 6th Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 1358–1367, Singapore.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44:179–222.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '12, pages 301–305, Jeju Island, Korea.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING '12, pages 1979–1994, Mumbai, India.

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*, ACL '03, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, PA.

Kevin Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the LREC 2006 Workshop on Strategies for developing machine translation for minority languages*, pages 103–108, Genoa, Italy.

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 the 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.

Jörg Tiedemann. 2012a. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 141–151, Avignon, France.

Jörg Tiedemann. 2012b. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC '12, pages 2214–2218, Istanbul, Turkey.

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*, NAACL-HLT '07, pages 484–491, Rochester, NY.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '05, pages 590–596, Borovets, Bulgaria.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 33–39, Prague, Czech Republic.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2012. Source language adaptation for resource-poor machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 286–296, Jeju Island, Korea.

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*, ACL '07, pages 856–863, Prague, Czech Republic.

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 Association for Computational Linguistics and of the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL-IJCNLP '09, pages 154–162, Singapore.

Xiaoheng Zhang. 1998. Dialect MT: a case study between Cantonese and Mandarin. In *Proceedings of the 17th International Conference on Computational Linguistics*, COLING '98, pages 1460–1464, Montréal, Québec, Canada.