The TALP-UPC Neural Machine Translation System for German/Finnish-English Using the Inverse Direction Model in Rescoring

Carlos Escolano, Marta R. Costa-jussà and **José A. R. Fonollosa**
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
carlos.escolano@tsc.upc.edu, {marta.ruiz,jose.fonollosa}@upc.edu

Abstract

In this paper, we describe the TALP-UPC participation in the News Task for German-English and Finish-English. Our primary submission implements a fully character to character neural machine translation architecture with an additional rescoring of a n-best list of hypothesis using a forced back-translation to the source sentence. This model gives consistent improvements on different pairs of languages for the language direction with the lowest performance while keeping the quality in the direction with the highest performance.

Additional experiments are reported for multilingual character to character neural machine translation, phrase-based translation and the additional Turkish-English language pair.

1 Introduction

Neural Machine Translation (MT) has been proven to reach state-of-the-art results in the last couple of years. The baseline encoder-decoder architecture has been improved by an attention-based mechanism citebahdanau:2015, subword units (Sennrich et al., 2016b), character-based encoders (Costa-jussà and Fonollosa, 2016) or even with generative adversarial nets (Yang et al., 2017), among many others.

Despite its successful beginnings, the neural MT approach still has many challenges to solve and improvements to incorporate into the system. However, since the system is computationally expensive and training models may last for several weeks, it is not feasible to conduct multiple experiments for a mid-sized laboratory. For the same reason, it is also relevant to report negative results on NMT.

In this system description, we describe our participation on German-English and Finnish-English for the News Task. Our system is a fully character-to-character neural MT (Lee et al., 2016) system with additional rescoring from the inverse direction model. In parallel to our final system, we also experimented with multilingual character-to-character system using German, Finnish and Turkish on the source side and English on the target side. Unfortunately, these last experiments did not work. All our systems are contrasted with a standard phrase-based system built with Moses (Koehn et al., 2007).

2 Char-to-char Neural MT

Our system uses the architecture from (Lee et al., 2016) where a character-level neural MT model maps the source character sequence to the target character sequence. The main difference in the encoder architecture respect to the standard neural MT model from (Bahdanau et al., 2015) is the use of a segmentation-free fully character-level network that extends initial character-based approaches like (Kim et al., 2015; Costa-jussà and Fonollosa, 2016). In the encoder, the network architecture includes character embeddings, convolution layers, max pooling and highway layers. The resulting character-based representation is then used as input to a bidirectional recurrent neural network. The main difference in the decoder architecture is that the single-layer feedforward network computes the attention score of next target character (instead of word) to be generated with every source segment representation. And afterwards, a two-layer character-level decoder takes the source context vector from the attention mechanism and predicts each target character.

3 Rescoring with inverse model

The motivation behind this technique is the idea that a good translation of a sentence has to be able to produce the original sentence with high probability when it is back-translated to the original source. We expect to be able to produce the source sentence from the translation with high probability only if the information of the source sentence is preserved.

In this approach, the first direct NMT decoder uses the standard beam search algorithm to generate an n-best list of translation hypothesis with its corresponding score

The list of translation outputs and the source sentence are then fed to the inverse *forced decoder* to calculate the probability of generating the original source sentence using each of them as input.

At this point, for each translation candidate we have two probabilities: the one obtained at the first translation step and the one obtained from the inverse *forced decoding*. A simple linear combination of scores is then used to rerank and select the best translation. Specifically, for this decision task, we used the rescoring tools provided by *Moses* that allow us to create a weighted model (using a validation set). For each sentence its final score is calculated as $w1 \cdot s1 + w2 \cdot s2$, where $w1$ and $s1$ are the weight and score (logarithm of the probability) of the translation model, while $w2$ and $p2$ are the weight and score (logarithm of the probability) provided by the forced decoder in the inverse direction. The hypothesis with the highest score is then returned as the final translation.

4 System description

In this section we detail experimental corpora, architecture and parameters that we used to build our WMT 2017 submissions. We report additional details from contrastives systems that we used internally to compare our submissions.

As mentioned earlier, our submissions use a char-to-char neural MT architecture for German-English and Finnish-English. Additional contrastive submissions that we did not present in the WMT evaluation include: a standard phrase-based MT system built with Moses (Koehn et al., 2007) and a multilingual char-to-char neural MT system from the same paper (Lee et al., 2016), where we train different source languages to the same target language. The main difference with the multilingual architecture is that the number of convolutional filters varies. We built contrastive submissions on the phrase-based system for German-English, Finnish-English and we also built it for a language pair that we did not present in the evaluation which was Turkish-English. Multilingual char-to-char was only built for German,Finnish and Turkish to English.

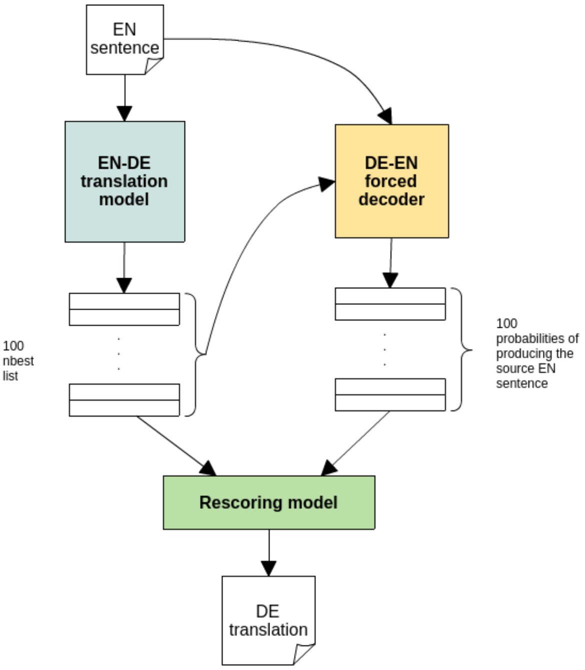

Figure 1: Overview of the architecture. In the image applied to a english-german translation

4.1 Data and Preprocess

For the three language pairs that we experimented with, we used all data parallel data available in the evaluation[1]. For German-English, we used: *europarl v.7*, *news commentary v.12*, *common crawl* and *rapid corpus of EU press releases*. We also used automatically back-translated in-domain monolingual data (Sennrich et al., 2016a). For Finnish-English, we used *europarl v.8*, *wiki headlines* and *rapid corpus of EU press releases*. For Turkish-English, we used *setimes2*. All our systems falled into the constrained category. Also note that only for German-English we took advantage of the monolingual corpus provided.

Preprocessing consisted in cleaning empty sentences, limiting sentences up to 50 words, tokenization and truecasing for each language using tools from Moses (Koehn et al., 2007). Table 1 shows details about the corpus statistics after preprocessing. For German and Finnish pairs

[1] http://www.statmt.org/wmt17/translation-task.html

LP	L	Set	S	W	V
DeEn	De	Train	9659106	203634165	1721113
		Dev	2999	62362	12674
		Test	2169	44085	9895
		Eval	3004	60965	12763
	En	Train	9659106	210205446	954387
		Dev	2999	64503	9506
		Test	2169	46830	7871
		Eval	3004	64706	9434
FiEn	Fi	Train	2468673	37755811	863898
		Dev	3000	47779	16236
		Test	2870	43069	15748
		Eval	3002	45456	16239
	En	Train	2468673	52262051	240625
		Dev	3000	63519	9059
		Test	2870	60149	8961
		Eval	3002	62412	8956
TuEn	Tu	Train	200290	4248508	158276
		Dev	1001	16954	6463
		Test	3000	54128	15898
		Eval	3007	55293	15264
	En	Train	299290	4713025	73906
		Dev	1001	22136	4318
		Test	3000	66394	9503
		Eval	3007	67839	9181

Table 1: Corpus Statistics. Number of sentences (S),words (W), vocabulary (V). M stands for millions and K stands for thousands.

the evaluation set is news2016 challenge test and the test set is the news2015 test. For Turkish news2016 developement and test set were employed.

Table 2 shows the total vocabulary size in characters (characters) for each language. We also show the limited vocabulary size that we used to train (vocabulary) and the coverage of this limited vocabulary (coverage).

4.2 Parameters and Training Details

- **Moses**. We used the following parameters: grow-diag-final word alignment symmetrization, lexicalized reordering, relative frequencies (conditional and posterior probabilities) with phrase discounting, lexical weights, phrase bonus, accepting phrases up to length 10, 5-gram language model with kneser-ney smoothing, word bonus and MERT optimisation (Koehn et al., 2007).

- **Char-to-char neural MT**. For the embedding of the source sentence, we use set of convolutional layers which number kernels are (200-200-250-250-300-300-300-300) and their lengths are (1-2-3-4-5-6-7-8) respectively. Additionally 4 highway layers are employed. And a bidirectional LSTM layer of 512 units for encoding. The maximum souce sentence's length is 450 during training and 500 for decoding both during training and sampling.

- **Multilingual char-to-char neural MT**. As proposed in the original work (Lee et al., 2016), we implement this model with slightly more convolutional filters than the char-to-char model, namely (200-250- 300-300-400-400-400-400). Also the maximum sentence lenght used for training is 400 for this model. The other parameters of the network are set to the same values than in the bilingual models.

4.3 Results

Table 3 shows results for the systems that we trained in this evaluation: phrase-based, char-to-char neural MT with and without inverse model rescoring and multilingual char-to-char neural MT. We submitted the best systems from Table 3 for German-English and Finnish-English, which is the char-to-char neural MT with rescoring of the inverse model. We computed statistical signficance based on (Clark et al., 2011). Our proposed method obtains a better BLEU score with $> 95\%$ statistical significance.

4.3.1 German $\longleftrightarrow$ English

This language pair was trained for 1.000.000 of updates (batches). We generated a 100 n-best list and did rescoring using force decoding over the inverse direction.

4.3.2 Finnish $\longleftrightarrow$ English

This model trained for 900.000 updates (batches) for both language pairs. Rescoring is applied to the 100 n-best list using the force decoded probabilities obtained from the inverse model.

4.3.3 Turkish $\longleftrightarrow$ English

This model trained for 200.000 updates. For this model rescoring did not produce significative improvement in the results as seen in 3. Also analyzinfg the results obtained we came to the conclusion that the corpus employed of approximately 200.000 sentences was not big enough to train the char2char model specially when compared with the resuts obtained using the phrase based model.

4.3.4 Multilingual

This model trained for 1.200.000 updates using all parallel data provided for the competition in German-English, Finnish-English, Turkish-English. As we can see in 3 the results obtained by the bilingual models outperform the ones obtained by this model. It is also worth to mention the case performance in Turkish where 0 BLEU

Language	Pair	Characters	Vocabulary	Coverage(%)
German(DE)	DE-EN/EN-DE	2379	300	99
English(EN)	DE-EN/EN-DE	2540	300	99
Finnish(FI)	FI-EN/EN-FI	439	300	99
English(EN)	FI-EN/EN-FI	438	300	99
Turkish(TU)	TU-EN/EN-TU	140	140	100
English(EN)	TU-EN/EN-TU	160	160	100

Table 2: Characters, vocabulary size and coverage for each language.

System	**DeEn**		**EnDe**		**FiEn**		**EnFi**		**TuEn**		**EnTu**	
	test	eval	test	eval	test	eval	test	eval	test	eval	test	eval
Phrase	23.59	22.71	18.25	17.93	9.71	11.35	13.67	**15.62**	**11.10**	**9.77**	**7.25**	**8.33**
Char2Char	**28.63**	**32.07**	21.08	26.61	**14.75**	15.75	11.54	11.21	5.87	6.77	6.23	4.73
+Rescoring	**28.63**	**32.07**	**21.37**	**26.98**	**14.75**	15.75	**11.98**	**11.63**	5.87	6.77	6.23	4.73
Multilingual	24.91	29.81	-	-	12.66	13.06	-	-	0	0	-	-

Table 3: BLEU results. In bold, best results.

score was obtained. This may be also explained by the limited corpus used for this language pair compared to the ones employed for German and Finnish.

System	Examples
Truth	CHIO : ” goldene Sportpyramide ” fr Bernhard Langer
	der Grund war durchaus berzeugend .
Char2Char	CHIO : ” Golden Sport **Pyradid** ” fr Bernhard Langer .
	er **hatte** grndlich berzeugt
+Rescoring	CHIO : ” Golden Sport Pyramid ” fr Bernhard Langer .
	er war grndlich berzeugt .

Table 4: Examples of German translation with and without rescoring.

System	**Examples**
Truth	Louis Galicia sanoi , ett Frank ypyi alkuun hostelleissa, mutta ett skettin ” hnen asiansa sujuivat vihdoinkin hyvin ”.
	hn oli hyvntahtoinen ja hnell oli suuri sydn .
	kyll , tilaisuudessa kteltiin
Char2Char	Louis Galicia sanoi , ett Frank ji aluksi houkutteluihin, mutta hiljattain : ” asiat sujuivat lopultakin hyvin ”.hn oli ystvllinen henki.kyll annettiin ktti.hn oli
	ystvllinen henki.
	kyll annettiin ktti .
+Rescoring	Louis Galicia sanoi , ett Frank ji aluksi houkutteluihin, mutta hiljattain : ” asiat sujuivat vihdoin hyvin ”.
	hn oli ystvllinen ja suuri sydmen henki.
	kyll annettiin kdellisyytt .

Table 5: Examples of Finnish translation with and without rescoring.

Table 4 shows several translation output examples. The first example shows how the rescoring technique can help when a word has been incorrectly spelled. In the second example, we see the correction of a badly translated word.

Table 5 shows some examples of Finnish translations. The examples show how even if the rescoring is not able to generate the correct translation it is able to produce a more similar word than the model without rescoring.

5 Conclusions

In this paper, we have described the TALP-UPC participation in the News Task. Our system implements a char-to-char neural MT with rescoring of the inverse direction model. This model gives consistent improvements on different pairs of languages for the language direction with lowest performance while keeping invariant the language direction with highest performance.

Acknowledgements

This study has been funded by the Spanish *Ministerio de Economía y Competitividad* and the European Regional Development Fund, through the postdoctoral senior grant *Ramón y Cajal* and the contract TEC2015-69266-P (MINECO/FEDER, EU).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. volume abs/1409.0473. http://arxiv.org/abs/1409.0473.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT ’11, pages 176–181. http://dl.acm.org/citation.cfm?id=2002736.2002774.

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 357–361. http://anthology.aclweb.org/P16-2058.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2015. Character-aware neural language models. *CoRR* abs/1508.06615. http://arxiv.org/abs/1508.06615.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR* abs/1610.03017. http://arxiv.org/abs/1610.03017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891* .

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2017. Improving neural machine translation with conditional sequence generative adversarial nets. *CoRR* abs/1703.04887. http://arxiv.org/abs/1703.04887.