A Recipe for Low-Resource NMT

Eryk Wdowiak( )

Arba Sicula, Brooklyn, NY, USA
eryk@wdowiak.me
http://www.arbasicula.org

Abstract. Incorporating theoretical information into the dataset, tokenization and subword splitting improves translation quality in low-resource settings. Previous research has shown that one can train a reasonably good translation model by training a model with small subword vocabularies and high dropout parameters. And backtranslation and multilingual translation further improve translation quality. But just as a textbook helps a student learn a language, it also helps a machine learn a language. Theoretical information allows us to make more efficient use of a given dataset and train a better model.

AQ1

Keywords: Neural machine translation · Subword-splitting · Low-resource languages · Sicilian language

1 Introduction

Last year, several researchers began an important discussion about large language models [2]. This paper shows what one can accomplish with small language models. Just as Transformers [16] scale upwards, they also scale downwards providing meaningful models that serve people in their preferred language.

Our innovation is to use existing methods more efficiently. Instead of scaling upwards to achieve performance gains, we achieved good translation quality by incorporating theoretical information into the dataset, the tokenization and the subword splitting. Given our experience, this paper proposes *modeling the language* to make more efficient use of a given dataset and to offer the promise of language models to all the world's people.

Our goal was to create a neural machine translator for the Sicilian language. Sicilian provides a good case study in low-resource machine translation for several reasons. First, the language has been continuously recorded since the Sicilian School of Poets joined the imperial court of Frederick II in the 13th century.

And in our times, Arba Sicula has spent the past 43 years translating Sicilian literature into English (among its numerous activities to promote the Sicilian language). In the course of their work with the many dialects of Sicilian, the organization established a "Standard Sicilian," a single form of the language.

To train our translator, we had to make better use of limited amounts of parallel text than previous researchers had. Just a few years ago, [10] calculated

K. Arai (Ed.): SAI 2022, LNNS 507, pp. 1–8, 2022.
https://doi.org/10.1007/978-3-031-10464-0_50

learning curves for English-to-Spanish translation. At 377,000 words, their neural machine translation model only achieved a BLEU score of 1.6.

More recently, [13] improved upon their results by using subword-splitting [12] to train a neural German-to-English model that scored 16.6 on a 100,000 word dataset.

And we improved upon their results by incorporating theoretical information into our modeling strategy. With just 16,945 translated sentence pairs containing 266,514 Sicilian words and 269,153 English words, our *Tradutturi Sicilianu* achieved a BLEU score of 25.1 on English-to-Sicilian translation and 29.1 on Sicilian-to-English.

Then we augmented our dataset with backtranslation [11] and multilingual translation [9], which further increased our BLEU scores to 35.0 on English-to-Sicilian and to 36.8 on Sicilian-to-English.

That's a good result for a small amount of parallel text. It shows what one can accomplish by using theory to model the language.

The next section describes our data sources (Sect. 2). The section on subword splitting (Sect. 3) explains our method of biasing subwords towards theoretical stems and desinences. Then our "recipe (Sect. 4)" describes our method of training a translator on little parallel text. Finally, the section on multilingual translation (Sect. 5) explains how incorporating a trilingual "bridge" [6] of textbook exercises into our dataset further improves translation quality. And the last section concludes (Sect. 6).

AQ2

2 Data Sources

Our first ingredient is high-quality parallel text. Standard Sicilian provided the consistency necessary to create a high-quality corpus of Sicilian-English parallel text. With that good start, we avoided the dialect challenges faced by [8], who note that variations in pronunciation coupled with the lack of a written standard cause extreme inconsistency in spelling.

Consistent spelling increases word frequencies, enabling us to train a neural machine translation model on a small corpus of parallel text.

To seed this project, Arthur Dieli kindly provided 34 translations of Giuseppe Pitrè's *Sicilian Folk Tales* and lots of encouragement. And Arba Sicula, which has been translating Sicilian literature into English since 1979, contributed its bilingual journal of Sicilian history, language, literature, art, folklore and cuisine.

Most of our data comes from *Arba Sicula* articles. Some parallel text comes from Dr. Dieli's translations of Pitrè's *Folk Tales*. And some comes from translations of the homework exercises in the *Mparamu lu sicilianu* [5] and *Introduction to Sicilian Grammar* [3] textbooks.

Although it only makes up a small portion of the dataset, adding the textbook examples yielded large improvements in translation quality on a test set drawn only from *Arba Sicula* articles. Just as a grammar book helps a human learn in a systematic way, it also helps a machine learn in a systematic way.

"Language models are few-shot learners" [4]. The textbook exercises provided the few examples of each grammatical element necessary to train a good model.

3 Subword Splitting

According to a recent case study of best practices for low-resource neural machine translation [13], neural models can achieve better translation quality than phrase-based statistical machine translation. In their best practices, the authors suggest using a smaller neural network with fewer layers, smaller batch sizes and a larger the dropout parameter.

Importantly, their largest improvements in translation quality (as measured by BLEU score) came from the application of a byte-pair encoding [12] that reduced the vocabulary from 14,000 words to 2000 words.

Our experience suggests that biasing the subword distribution toward theoretical stems and desinences further improves translation quality.

For example, the English present tense only has two forms – *speak* and *speaks* – while the Sicilian present tense has six – *parru*, *parri*, *parra*, *parramu*, *parrati* and *parranu*. But upon splitting them into subwords, *parr+* matches *speak+*, while the Sicilian verb endings (*+u*, *+i*, *+a*, *+amu*, *+ati* and *+anu*) match the English pronouns.

So subword splitting should allow us represent many different word forms with a much smaller vocabulary and should allow the translator to learn rare words and unknown words. For example, even if "jo manciu" ("I eat") does not appear at all in the dataset, but forms like "jo parru" ("I speak") and "iddu mancia" ("he eats") do appear, then subword splitting should allow the translator to learn "jo manciu" ("I eat").

In practice, achieving that effect required us to bias the learned subword vocabulary towards the stems and desinences one finds in a textbook. Specifically, we added a unique list of words from the *Dieli Dictionary* and the inflections of verbs, nouns and adjectives from *Chiù dâ Palora* to the Sicilian data.

Because each word was only added once, none of them affected the distribution of whole words. But once the words were split, they greatly affected the distribution of subwords, filling it with stems and suffixes. So the subword vocabulary that the machine learns is similar to the theoretical stems and desinences of a textbook. And the translation model learns to translate in a more theoretic manner, making it more generalizable to unseen data.

Within a given dataset, theoretical splitting increased our BLEU scores from 20.3 to 22.4 on English-to-Sicilian and from 21.4 to 24.1 on Sicilian-to-English.

4 A Recipe for Low-Resource NMT

Even though we only have a little parallel text, we can still develop a reasonably good neural machine translator. We just have to train a smaller model for the smaller dataset. As shown in Table 1, we trained models of three different sizes, all of which were smaller than the defaults provided by the Sockeye toolkit [7].

And just as we incorporated theoretical information into our dataset, we also incorporated theory into our modeling strategy. In this section, we incorporate insights from statistical theory because in a low-resource context, we must be careful to avoid over-fitting.

Table 1. Model sizes

	Defaults	Our models	Larger	Many-to-many
Layers	6	3	4	4
Embedding size	512	256	384	512
Model size	512	256	384	512
Attention heads	8	4	6	8
Feed forward	2048	1024	1536	2048

Training a large model on a small dataset is comparable to estimating a regression model with a large number of parameters on a dataset with few observations: It leaves you with too few degrees of freedom. The model thus becomes over-fit and does not make good predictions.

Reducing the vocabulary with subword-splitting, training a smaller network and setting high dropout parameters all reduce over-fitting. And self-attentional neural networks also reduce over-fitting because (compared to recurrent and convolutional networks) they are less complex. They directly model the relationships between words in a pair of sentences.

This combination of splitting, dropout and self-attention is an implementation of the best practices discussed above [13], but using the Transformer model [16] from the Sockeye toolkit [7].

It achieved a BLEU score of 25.1 on English-to-Sicilian translation and 29.1 on Sicilian-to-English with only 16,945 lines of parallel training data containing 266,514 Sicilian words and 269,153 English words.

In their best practices study, the authors found that reducing the vocabulary to 2000 subwords yielded the largest improvements in translation quality. But their most successful training also occurred when they set high dropout parameters [13].

During training, dropout randomly shuts off a percentage of units (by setting it to zero), which effectively prevents the units from adapting to each other. Each unit therefore becomes more independent of the others because the model is trained as if it had a smaller number of units, thus reducing over-fitting [14].

Subword-splitting and high dropout parameters helped us achieve better than expected results with a small dataset. And the Transformer model pushed our BLEU scores into the double digits.

Compared to recurrent neural networks, the self-attention layers in the Transformer model more easily learn the dependencies between words in a sequence because the self-attention layers are less complex.

Recurrent networks read words sequentially and employ a gating mechanism to identify relationships between separated words in a sequence. By contrast, self-attention examines the links between all the words in the paired sequences and directly models those relationships. It's a simpler approach.

Table 2. Datasets and results

Dataset	Subwords	Lines	Word count (in tokens)			BLEU score	
			Sicilian	English	Italian	En-Sc	Sc-En
20	2,000	7,721	121,136	121,892	–	11.4	12.9
21	2,000	8,660	146.370	146,437	–	12.9	13.3
23	3,000	12,095	171,278	175,174	–	19.6	19.5
24	3,000	13,060	178,714	183,736	–	19.6	21.5
25	3,000	13,392	185,540	190,538	–	21.1	21.2
27	3,000	13,839	190,072	195,372	–	22.4	24.1
28	3,000	14,494	196,911	202,652	–	22.5	25.2
29	3,000	16,591	258,730	261,474	–	24.6	27.0
30	3,000	16,945	266,514	269,153	–	25.1	29.1
30	5,000	16,829	261,421	264,242	–	27.7	–
+back		+3,251	+92,141	–	–		
30	Sc: 5,000	16,891	262,582	266,740	–	19.7	26.2
Books	En: 7,500	32,804	–	929,043	838,152	35.1*	34.6*
+back	It: 5,000	+3,250	+92,146	–	–		
33		12,357	237,456	236,568	–		
Books		28,982	–	836,757	755,196	35.0*	36.8*
+back En/It-Sc	Sc: 5,000	+3,250	+92,146	–	–		
+back Sc-It	En: 7,500	+3,250	–	–	+84,657		
	It: 5,000	*4,660*	*30,244*	*35,173*	–	**It-Sc**	**Sc-It**
textbook		*4,660*	*30,244*	–	*29,855*	36.5†	30.9†
		4,660	–	*35,173*	*29,855*		

The *textbook* exercises form a trilingual "bridge," the strategy proposed by [6].

* larger model

† many-to-many model

Combining these three features – small subword vocabularies, high dropout parameters and self-attention – yields a trained model that makes relatively good predictions despite being trained on limited amounts of parallel text because they reduce over-fitting.

5 Multilingual Translation

Our discussion so far has focused on a dataset of Sicilian-English parallel text. This section augments our dataset with parallel text in other languages to enable multilingual translation [9] and improve translation quality.

In our case, we can obtain Sicilian-English parallel text from the issues of *Arba Sicula* but finding Sicilian-Italian parallel text is difficult.

Nonetheless, we trained a model to translate between Sicilian and Italian without any Sicilian-Italian parallel text at all (i.e. "zero shot" translation) by including Italian-English parallel text in our dataset. Then, to improve translation quality between Sicilian and Italian, we implemented a "bridging strategy" [6] by adding Sicilian-Italian-English homework exercises to our dataset.

It's an example of transfer learning. In our case, as the model learns to translate from Italian to English, it also learns to translate from Sicilian to English. And as the model learns to translate from English to Italian, it also learns how to translate from English to Sicilian.

More parallel text is available for some languages than others however, so [9] also studied the effect on translation quality and found that oversampling low-resource language pairs improves their translation quality, but at expense of quality among high-resource pairs.

Importantly however, the comparison with bilingual translators holds constant the number of parameters in the model. Training a larger model can improve translation quality across the board [1].

Our experience was consistent with these findings. As shown in Table 2, holding model size constant reduced translation quality when we added the Italian-English subset of Farkas' *Books* data (from the OPUS project [15]) to our dataset. So to push our BLEU scores into the thirties, we trained a larger model – an appropriately sized model.

In a broader effort, another study developed a "bridging strategy" to collect data for and to train a model that can directly translate between 100 languages. To overcome the limitations of *English-centric* data, the authors strategically selected pairs to mine data for, based on geography and linguistic similarity. Their approach yielded large improvements in translation quality in non-English directions, while matching translation quality in English directions [6].

A similar strategy improved our translation quality between Sicilian and Italian. Taking a theoretic approach, we bridged Sicilian, English and Italian by translating 4,660 homework exercises from the *Mparamu lu sicilianu* [5] and *Introduction to Sicilian Grammar* [3] textbooks. As shown in Table 2, this technique yielded translation quality between Sicilian and Italian that's almost as good as translation quality between Sicilian and English, for which we have far more parallel text.

6 Conclusion

Our recipe for low-resource neural machine translation – theoretical subword-splitting, high dropout parameters and self-attention – yields a trained model that makes relatively good predictions. Adding backtranslation and multilingual translation improves translation quality even more. And we improved upon our zero-shot result by bridging the three languages with textbook exercises.

Most importantly, we achieved these good results by training a small model. Instead of scaling upwards, we used theory to make more efficient use of a dataset and help a small model learn a good set of translation rules.

We hope our experience encourages practitioners to *model the language* and to develop language models for all the world's people.

Acknowledgments. Arba Sicula, Gaetano Cipolla and Arthur Dieli developed the resources that made this project possible. I would like to thank them for their support and encouragement.

Prof. Cipolla helped me learn Sicilian and he also helped me develop this recipe for low-resource neural machine translation. We thought about the problem together. He encouraged me to incorporate theoretical information into the model and that's why we got good results.

Dr. Dieli seeded this project with his vocabulary list and translations of Pitrè's *Folk Tales*. He helped me get started. And he and his family gave me a lot of support and encouragement. This project is dedicated to his memory.

Finally, I would like to thank Arba Sicula for the language resources that we used to develop the dictionary and translator. And I would like to thank the organization and its members for their sponsorship and development of Sicilian language and culture. Their poetry made this project beautiful.

Grazzi!

References

1. Arivazhagan, N., et al.: Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv preprint arXiv:1907.05019 (2019)
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922
3. Bonner, J.K.: Introduction to Sicilian Grammar. Legas, Brooklyn (2001)
4. Brown, T.B., et al.: Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165 (2020)
5. Cipolla, G.: Learn Sicilian. In: Mparamu lu Sicilianu. Legas, Mineola (2013)
6. Fan, A., et al.: Beyond English-Centric Multilingual Machine Translation. arXiv preprint arXiv:2010.11125 (2020)
7. Hieber, F., et al.: Sockeye: A Toolkit for Neural Machine Translation. arXiv preprint arXiv:1712.05690 (2017)
8. Hollenstein, N., Aepli, N.: Compilation of a Swiss German dialect corpusand its application to PoS tagging. In: Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pp. 85–94 (2014). https://aclanthology.org/W14-5310.pdf
9. Johnson, M.S., et al.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv preprint arXiv:1611.04558 (2016)
10. Koehn, P., Knowles, R.: Six Challenges for Neural Machine Translation. arXiv preprint arXiv:1706.03872 (2017)
11. Sennrich, R., Haddow, B., Birch, A.: Improving Neural Machine Translation Models with Monolingual Data. arXiv preprint arXiv:1511.06709 (2015)
12. Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. arXiv preprint arXiv:1508.07909 (2016)
13. Sennrich, R., Zhang, B.: Revisiting Low-Resource Neural Machine Translation: A Case Study. arXiv preprint arXiv:1905.11901 (2019)

14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014). http://jmlr.org/papers/v15/srivastava14a.html
15. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (2012). https://opus.nlpl.eu/
16. Vaswani, A., et al.: Attention Is All You Need. arXiv preprint arXiv:1706.03762 (2017)