Preserving Personality through a Pivot Language

Low-Resource NMT of Ancient Languages

Annie K. Lamar & America Chambers
Department of Mathematics & Computer Science
University of Puget Sound
Tacoma, WA, USA
kalamar@pugetsound.edu

We investigate the effectiveness of neural machine translation in a low-resource setting, where particular attention is given to preserving the rhetorical style of a given persona. In particular, we attempt to complete Cicero's translation of Timaeus, a text originally written by Plato in Attic Greek. Due to the scarcity of parallel Attic Greek and Classical Latin text, we use English as a pivot language. Finally, we explore how additional texts, which may help improve the accuracy of the model, affect the preservation of a particular persona (i.e. Socrates) through translation into the pivot language and subsequently the target language.

Key words: Neural machine translation, low-resource, deep learning, persona, Greek, Latin

I. INTRODUCTION

In this paper, we investigate the effectiveness of neural machine translation in a low-resource setting, where particular attention is given to preserving the rhetorical style of a given persona. In the first century BCE, the Roman orator and senator Cicero began to translate into Classical Latin the philosophical work *Timaeus*, a text originally written by Plato in Attic Greek. However, Cicero completed only one-third of the translation before his death. Despite the recent advancements in the field of machine translation, completing the remainder of Cicero's translation remains a significant challenge due to (1) the scarcity of parallel Attic Greek and Classical Latin text and (2) the need to mimic not only Cicero's writing style but also the persona of Socrates, a character who occupies a major role within *Timaeus* and maintains a consistent personality in all of Plato's texts.

One approach in such a low-resource setting is to use outside texts [1]-- in our case, other texts that were written by Plato and Cicero but differ markedly from *Timaeus* in format and content. However, this approach creates a tension between accuracy and fidelity: additional texts help improve the accuracy of the model but may cause a "washing out" effect when trying to preserve a persona through translation. In this paper, we investigate this effect by examining how well the persona of Socrates is preserved in the face of additional data that differs in style, format, and content.

We like would to express our appreciation to the McCormick Student Research Grant in the Sciences for funding this research.

II. DATASETS

Neural machine translation (NMT) has been shown to surpass the performance of more conventional techniques (e.g. statistical machine translation) when a significant amount of training data exists between the language pairs [2, 3]. In this task, however, there are only 117 sentences in the translated portion of *Timaeus* -- much too little data for training. Although there is a scarcity of additional parallel Attic Greek and Classical Latin texts, we can use English as a pivot language [4, 5]. That is, we train separate Greek-to-English and English-to-Latin models and construct a separate training set for each.

For the Greek-to-English dataset, we use all of Plato's texts available through the Perseus Digital Library [6] which consists of 31,678 sentences. Of these 31,678 lines, approximately 16k are written in Socrates' voice while the remaining 15k are spoken by the narrator or other minor characters. We train our baseline model on just the 16k lines spoken by Socrates and a "noisy" model on the same 16k lines plus the additional 15k lines by other characters. Likewise, the English-to-Latin dataset contains most of Cicero's texts available through the Perseus Digital Library[1] and consists of 11,123 sentences.

The Greek-to-English dataset was manually aligned in its entirety by one of the authors. The English-to-Latin dataset was first aligned using Hunalign [7]. These alignments were then manually reviewed and corrected by one of the authors.

Note that this additional data differs in style, format, and content from *Timaeus* thus forming a type of transfer learning.

III. MODELS

Two pivot-based models were trained to investigate how the addition of less-relevant, "noisy" data to the dataset affects (1) the preservation of a particular persona through a pivot

[1] Plato wrote in Attic Greek, a dialect distinct enough from Koine Greek as to prevent the inclusion of biblical texts in the training set. Classical Latin is also distinct from the Latin used in the composition of biblical texts at the start of the first millennium. As such, we use the term "Greek" to refer to "Attic Greek" and "Latin" to refer to "Classical Latin."

language and target language and (2) the accuracy of the translation on the test set.

1. **Baseline**: The baseline model consists of two separate source-pivot models. The first source-pivot model was trained on only the subset of lines from the Greek-to-English dataset that were in the voice of Socrates (approximately 16k lines). The second source-pivot model was trained using the remaining lines spoken by the narrator or other minor characters (approximately 15k lines). For each sentence in the test set, the speaker is first identified and the sentence is routed to the correct model.
2. **Noisy:** The Noisy model was trained using the entirety of the Greek-to-English dataset where 50% of the training data is from Socrates' and 50% from other characters.

Note that both models use the same pivot-target model -- i.e., both models use the same English-to-Latin model.

IV. Encoder-Decoder RNNs

In this section, we provide a brief overview of the encoder-decoder recursive neural network (RNN) model [8] used in this paper for translation. Let $\mathbf{x} = [x_1\ x_2\ \dots\ x_n]$ be an input sentence. The "encoder" RNN incrementally builds a fixed-length encoding $\mathbf{h}$ of the input sentence given by

$$\mathbf{h} = g(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n) \text{ where } \mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t) \text{ for } t = 1...n \quad (1)$$

The encoding $\mathbf{h}$ is a function of the partial encodings ($\mathbf{h}_1$, $\mathbf{h}_2$, … $\mathbf{h}_n$) of the sentence where the partial encoding $\mathbf{h}_t$ is itself a function of two quantities: the content of the sentence up to this point (captured by $\mathbf{h}_{t-1}$) and the next word $\mathbf{x}_t$. The function g is commonly chosen to be $g(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n) = \mathbf{h}_n$, that is $\mathbf{h}_n$ is assumed to be a sufficient encoding of the entire sentence. One common choice for the function f is a long short-term memory (LSTM) model [9, 10] that keeps track of an internal state as a form of short-term memory. At each step, the new state is derived from the previous one in a process of forgetting and then updating in response to the new word $\mathbf{x}_t$.

Given the final encoding of the input sentence (which we denote as $\mathbf{h}_x$), the translation $\mathbf{y}=[y_1\ y_2\ \dots\ y_m]$ is generated using another "decoder" RNN. At each step, a word y_t is sampled conditioned on the previously generated word y_{t-1}, the partial encoding $\mathbf{h}_{t-1}$ of the translation, and the encoding of the input sentence $\mathbf{h}_x$. That is,

$$p(y_t \mid y_1...y_{t-1}, \mathbf{h}_x) = f(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{h}_x) \quad (2)$$

Note f must produce a valid probability distribution and is typically accomplished by passing the output of the LSTM through the softmax function.

In our experiments, we augment the model by using a bidirectional encoding (where we compute partial encodings from the beginning of the sentence forward *and* from the end of the sentence backward) and an attention mechanism (where the generation of each word y_t is now conditioned on a weighted sum of the partial encodings of $\mathbf{x}$ learned for each y_t rather than conditioning on $\mathbf{h}_x$) [2].

V. Experiments

We use the OpenNMT-py library [11]. All encoder RNNs are 2-layer bidirectional LSTMs with encoding dimension of 500 (i.e. the forward and backward encodings each have dimension 250). The decoder RNN is also a 2-layer bidirectional LSTM with dimension 500. We use stochastic gradient descent for optimization and a dropout probability of 0.3. We use a learning rate of 0.5 for all source-pivot models and a learning rate of 0.1 for the pivot-target (i.e. English-to-Latin) model. The validation sets were made by removing the last 100 lines from every 1000 lines, resulting in a validation set that was approximately 10% the size of the training set and which represented the variety of data within the given dataset. The models were trained until the performance on the validity stopped improving and saved every 50 iterations.

VI. Results

The models achieved the following BLEU [12] scores (where larger values are better):

Model		**Overall BLEU**	**1-gram**	**2-gram**	**3-gram**	**4-gram**
Baseline	***Greek-to-English***	0.34	11.70	1.60	0.10	0.00
	English-to-Latin	0.00	4.00	0.10	0.00	0.00
Noisy	***Greek-to-English***	1.60	26.70	4.40	0.50	0.10
	English-to-Latin	0.90	20.20	2.70	0.20	0.10

The Greek-to-English BLEU scores are calculated by comparing the English prediction with the actual English translation. The English-to-Latin BLEU scores are calculated by comparing the Latin prediction (which is translated from the English prediction) with the actual Latin translations. As a result, any errors in the English translation will affect the Latin translation.

As expected, the Noisy model attains higher BLEU scores than the Baseline model ostensibly due to the larger amount of training data (i.e. 31k sentences for the Greek-to-English Baseline model as opposed to only 16k sentences for the Noisy model).

When we examine the translations, we see the Noisy model was still able to preserve Socrates' personality despite 50% of the training data coming from other characters. Socrates' style of argument depends on asking questions in order to reveal a contradiction in another's logical assertion [13]. This most often takes the form of asking a negative question (e.g. "is it not so that…") and broadening the implications of the logical assertion to include not just some things, but all things (e.g. "in my city, in your city, in every city"). Below we show an example translation of a sentence spoken by Socrates under both models:

Greek:

ἆρ᾽ οὐ γυμναστικῇ καὶ μουσικῇ μαθήμασίν τε ὅσα προσήκει τούτοις, ἐν ἅπασι τεθράφθαι;

Correct English:

Were they not to be trained in gymnastics and in music and in all other learned things proper for such men?

Baseline Translations:

We must say that it is not easy to live and injustice in every case , and that in every case it is impossible to be born in every case

Non est id , quod in rebus et in re publica est

Noisy Translations:

it is not in respect of wrong and with music and music in all the studies of all the things in which it is in all respects

non modo est , et in omni rebus est , in quo rebus est est

In this example translation from the Baseline model, we see the inclusion of negative questioning ("it is not", "Non est id") and a broad generalization in the latter part of the sentence ("in every case", "et in re publica est"). In Cicero's texts, the Republic stands for the culmination of the most important and central aspects of life, morality, and culture; Cicero uses the phrase "in re publica" often to signal that a concept applies to all things. In this way, the inclusion of the Latin phrase "et in re publica est" shows a transfer of Socrates' Greek personality into Cicero's Roman mindset. These attributes of Socrates' rhetorical style still persist in the noisy model despite the addition of approximately 50% more non-relevant lines:

In the Noisy model we also observe an improved vocabulary. The Greek word μουσικῇ comes through correctly as "music" in the Noisy model, but does not appear in the translation in the Baseline model; likewise the Greek word μαθήμασίν is correctly translated as "studies" (literally, "things learned") in the Noisy model but not the Baseline model.

Now we show a second example demonstrating the same persistence of Socrates' rhetorical style in both models.

Greek:

ἆρ᾽ οὖν οὐ τὸ τῶν γεωργῶν ὅσαι τε ἄλλαι τέχναι πρῶτον ἐν αὐτῇ χωρὶς διειλόμεθα ἀπὸ τοῦ γένους τοῦ τῶν προπολεμησόντων;

Correct English:

Did we not begin by dividing off the class of land-workers in it, and all other crafts, from the class of its defenders?

Baseline Translations:

And is not that which is the greatest of all the rulers *in the world* for the sake of the multitude

hoc est enim hoc est , quod est publicae , ut rei publicae populi Romani .

Noisy Translations:

It is not , then , in regard to the majority of the Greeks in the world from which the universe has in it in the world from which the rules of the universe is distinct from it .

Est igitur est , ut in omni genere publica , quod in omni genere publica , in quo in re publica est .

Notice in the second example that the English "universe" gets transmitted in Latin as "in re publica est." This phrase, roughly meaning "in the republic," show again Cicero's equating of the Republic and the Roman people as the most general category. In this second example, however, we do not observe the persistence of Socrates' style of negative questioning in either Latin translation. Because this error appears in both the Baseline and Noisy model, we attribute this to the limited amount of English-to-Latin data used to the train the English-to-Latin model.

VII. Conclusion

The translations produced by the Baseline model demonstrate that it is possible to preserve a persona when translating via a pivot language. We also demonstrated with the Noisy model that Socrates' rhetorical style is not diminished even when 50% of the training data is "noisy" -- i.e., contains sentences spoken by other characters. Two aspects of Socrates' personality, his negative framing of questions and his tendency to generalize a logical assertion, were maintained in both the Baseline and Noisy model. We also see higher BLEU scores for the Greek-to-English models than the English-to-Latin models, likely because the Greek-to-English models were trained with approximately three times as much data as the English-to-Latin models.

VIII. Future Work

We are currently experimenting with training additional models that have been pre-trained using 100k lines of target-side monolingual data; that is, the Greek-to-English models will be pre-trained with 100k English lines and the English-to-Latin models will be pre-trained with 100k Latin lines. Target-side monolingual data has been shown to be an effective method of improving translation results without changes to the network architecture [14]. We will experiment with using pre-trained models for both the Baseline and Noisy models, testing if pre-training these models impacts the extent to which Socrates' personality is preserved as the amount of "noisy" data increases. That is, how much "noisy" data can be added before Socrates' personality becomes washed out.

IX. Related Work

See [2, 8, 15] for an introduction to using recurrent neural networks for mapping sentences to sentences (e.g. for machine translation or dialogue generation). For RNN models that attempt to preserve personality, see [16]. The use of a pivot language for machine translation was introduced by [4] for statistical machine translation and has since become a standard technique in NMT [5].

REFERENCES

[1] B. Zoph, D. Yuret, J. May, and K. Knight, K. "Transfer learning for low-resource neural machine translation", arXiv preprint:1604.02201, [cs.CL].

[2] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv:1409.0474 [cs.CL].

[3] M. T. Luong, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation", arXiv:1410.8206 [cs.CL].

[4] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation", *Machine Translation*, vol. 21, no. 3, pp. 165-181, 2007.

[5] Y. Cheng, Q. Yang, Y. Liu, M. Sun, and W. Xu, "Joint training for pivot-based neural machine translation", In Proc. of the 26th Int'l Joint Conf. on A.I., 2017, pp. 3974-3980.

[6] G. R. Crane, Ed., Perseus Digital Library, Tufts University. [Online]. Available: http://www.perseus.tufts.edu/hopper/ Accessed: June 2018]M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy, "Parallel corpora for medium density languages", In Proc. of the Recent Advances in Natural Lang. Processing, 2005, pp. 590-596.Te

[8] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation", arXiv:1406.1078 [cs.CL].

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[10] C. Olah, "Understanding LSTM networks", August, 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed: August 31, 2018].

[11] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation," arXiv:1701.02810 [cs.CL].

[12] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation", In *Proc. of the 40th Annual Mtg. on Assoc. for Compt'l. Linguistics,* pp. 311-318, 2002.

[13] M. Caminada, "A formal account of Socratic-style argumentation", *Journal of Applied Logic*, vol. 6, no. 1, pp. 109-132, 2008.

[14] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data", arXiv preprint:1511.06709 [cs:CL]

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks", In Proc. of the 27th Int'l Conf. on Neural Information Processing Systems, 2014, pp. 3104-3112.

[16] J. Li et al., "A persona-based neural conversation model", In Proc. of the 54th Annual Mtg. of the Assoc. for Compt'l Linguistics, 2016, pp. 994-1003.