Towards the First Machine Translation System for Sumerian Transliterations

Ravneet Punia
Delhi Technological University, India
ravneetpunia_bt2K16@dtu.ac.in

Niko Schenk*
Amazon
Berlin, Germany
nikosch@amazon.com

Christian Chiarcos
Goethe University
Frankfurt, Germany
chiarcos@cs.uni-frankfurt.de

Émilie Pagé-Perron
University of Toronto, Canada
epp@ucla.edu

Abstract

The Sumerian cuneiform script was invented more than 5,000 years ago and represents one of the oldest in history. We present the first attempt to translate Sumerian texts into English automatically. We publicly release high-quality corpora for standardized training and evaluation and report results on experiments with supervised, phrase-based, and transfer learning techniques for machine translation. Quantitative and qualitative evaluations indicate the usefulness of the translations. Our proposed methodology provides a broader audience of researchers with novel access to the data, accelerates the costly and time-consuming manual translation process, and helps them better explore the relationships between Sumerian cuneiform and Mesopotamian culture.

1 Introduction

Sumerian is the first recorded written language of mankind. A specific logo-syllabic script – Sumerian cuneiform – was used to record a variety of every-day events of ancient Mesopotamia, such as temple activities, business, trading or myths for a period of about 3,000 years. These texts were engraved on clay tablets using a reed stylus and are important for understanding the historical context of the Mesopotamian culture. An example is shown in Figure 1. Aside from great traditions in literature and mathematics that contributed to the foundations of modern religion and science alike, cuneiform languages provide a largely uninterrupted record of administrative and economic transactions for a period of approximately 3,000 years, and thus play an important role in the development and evaluation of modern theories of economy and historical sociology (Weber, 1976). Among cuneiform languages, Sumerian serves a particularly prominent role, as many aspects of the Sumerian language have been preserved in the writing of subsequent (Akkadian, Babylonian, Assyrian, Hittite) cultures. In particular, the use of Sumerograms (expressions in Sumerian) continued throughout the entire cuneiform tradition.

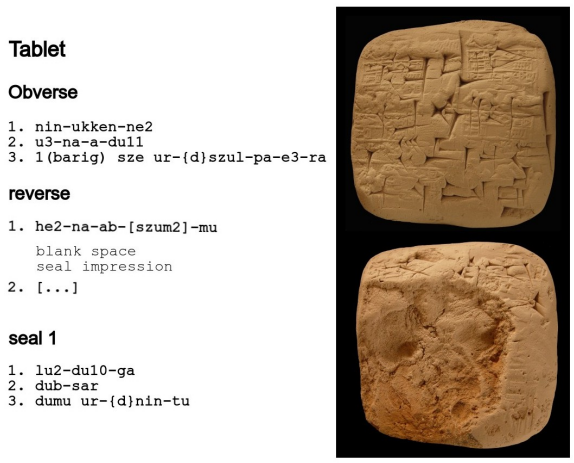

Figure 1: Example artifact of a cuneiform tablet with transliterated Sumerian text, Ur III period, Garshana, Mesopotamia (CDLI No. P323253), picture reproduced with kind permission by David I. Owen.

Here, we focus on a corpus from the limited time span (approx. 2100 - 2000 BCE) when Mesopotamia was united under rule of the Ur III dynasty – which established an extensive administrative apparatus and from which the majority

*Work was done prior to joining Amazon at Goethe University Frankfurt.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3454–3460
Barcelona, Spain (Online), December 8-13, 2020

of Sumerian documents originates. Overall, the Ur III corpus comprises 72,000 transcribed texts, out of which only 1,573 (2.2%) are available with translations.

Many pieces of Sumerian literature have been carefully edited and translated, but this material dates from periods when Sumerian was actually no longer a spoken language. Much of the material of our corpus on the other hand consists of short texts only, often of legal or administrative nature, e.g., about the transfer of goods and services. Specialists in Assyriology normally do not provide translations of such texts but work with the transliterated text directly. While such data might prove insightful for researchers from other areas, e.g., history or economy, it is largely *inaccessible* to non-specialists in the Sumerian language. There is thus a demand for the machine translation of Sumerian texts even beyond texts written in the language itself. Translating Sumerian is challenging on many levels because Sumerian is a linguistic isolate language with complex polysynthetic morphology. In a number of features, Sumerian is typologically different from any modern well-resourced language. This includes the extensive marking of semantic arguments by verbal morphology as well as the use of case morphology (case stacking) to mark syntactic phrase boundaries. Both phenomena are illustrated in the following verbal form:

...	bi2-in-ne2-sza-sze3							
...	bi-	i-	n-	e	-esz	-a	[-ak]	-sze
...	3-SG-NH	LOC2	3-SG-H-A	V-PL	3-PL	SUB	GEN	TERM
[... [	it_1	towards_2	he_3	speak	they_4	]	of_a	] on_b

'On_b (account) of_a (the fact) that they_{3+4} said that (they do not know about this_1)'

(CDLI no. P133620)

The example shows verbal agreement with three syntactic arguments (numbered 1,3,4) and one oblique argument (2) as well as nominalization of the verb (to express the meaning of a relative clause) and morphological marking for two cases, genitive (a, the case of the phrase itself) and terminative (b, the case of the morphological head of the verb), with phrase boundaries marked in the gloss. This form occurs as part of a legal text from the Ur-III corpus. As the example also shows, Sumerian uses a defective orthography that obfuscates certain (assumed) morphophonological processes (morphemes and syllabic characters do not align well, e.g., the prefixes *bi-* and *i-*, the verb *e* and the suffix *-esz*, and the suffixes *-a* and *-ak* are not orthographically separatable in the writing). Not all forms in the corpus exhibit this degree of morphological complexity, and in particular, most nominals tend to have a simpler structure, but overall, the corpus is sparse, and the rich morphology leads to a relatively low repetition rate in the data. Finally, many texts are missing information due to damaging or the decomposition of the tablets over time. In fact, a large corpus of transliterations is available, but unfortunately only a small subset is translated, which is part of our motivation for this project. The translation of these scripts is crucial in order to efficiently explore events related to the ancient civilisation (Crawford and Harriet, 2004).

In the past years, computer vision techniques were employed for the extraction of symbols, however, to date, no such system exists which tackles the challenging task of *translation* in an automated way. Recently, Pagé-Perron et al. (2017) described the concept for a system for Sumerian to English using character-based SMT. This system suffered massively from data sparsity and the approach has subsequently been abandoned by the authors. Our work fills this gap and, along with this paper, we publish the first machine translation pipeline for Sumerian–English. It fulfills the need to translate a large number of administrative texts by making them accessible to a *broader audience* beyond the closed circle of experts in Mesopotamian languages, including economists, historians, or linguists, as well as researchers working on ancient languages, for whom the manual translation of these texts is hardly possible.

2 Related Work

Aside from earlier work of the authors Pagé-Perron et al. (2017), we are not aware of any attempt to apply machine translation to cuneiform languages. However, the field does have a tradition with dictionary-based *glossing* of transliterated text. Similar to technologies commonly used in language documentation and linguistic typology (Robinson et al., 2007), the ORACC Lemmatizer (Robson, 2018; Liu et al., 2015) can provide word-by-word glosses along with a morphological analysis, albeit without contextual disambiguation, and without producing coherent text.

3 Data & Preprocessing

We work with the Ur-III corpus provided by the Cuneiform Digital Library Initiative[1] as part of the project Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC, 2017-2020). The Cuneiform Digital Library, founded in 1998, represents the central hub for digital philological data in Assyriology, and provides records for more than 340,000 cuneiform objects, out of which 120,000 come with transcriptions, 98,000 with images and 5,000 with translations. The Ur-III corpus only represents a fraction of this data, albeit a relatively homogeneous subset for a single language that thus represents a particularly promising area for the application of machine learning techniques.

The unannotated Ur-III corpus comprises 1.5 million lines in transliteration in total, out of which researchers translated approx. 20,000 Sumerian-English phrases and provided them as parallel, phrase aligned data to the project.[2] Transliterated cuneiform tablets (cf. Figure 1) represent the primary source of information. Much of this data originates in the Ur III period (21^{st} century BC), and covers in particular many administrative texts. In later centuries, Sumerian was still being used, but ceased to be a spoken language, so we base our experiments on this particular subset, a relatively homogeneous and (by the standards of Assyriology) large data set. Before we trained our models, the transliterations were preprocessed and cleaned. We applied the following procedure:

- Phrases with missing parallel translations as well as duplicates were removed.
- All (sparse) numbers indicating quantities were normalized and replaced by the placeholders `NUMB`.
- Identical source phrases with different translations were also omitted from the data set.

The final corpus consists of 10,147 unique Sumerian–English phrase pairs divided into standardized training/development/test splits of 80/10/10% each. It contains $\approx$ 28k and 64k tokens, with vocabulary sizes $|V_S|$=4,126 and $|V_E|$=3,146 for Sumerian and English, respectively. The mean length of Sumerian and English phrases is rather short with 2.8 and 4.4 tokens, respectively.

4 Training MT Systems for Sumerian

Previous research pointed out that machine translation models suffer from issues related to polysemy and multiple word senses (Calvo et al., 2019; Huang et al., 2011). To tackle these, we experimented with embeddings which we trained on our own small domain of English translations, as well as different pretrained word embeddings. Different attention designs such as global and local attention networks (Luong et al., 2015) and multi-head attention networks (Hans and Milton, 2016) were also subject for experimentation in order to test the efficiency on different sequence lengths. Overall, we experimented with several neural machine translation models, incl. phrase-based MT and transfer learning and implemented: a *Base Translator* with custom in-domain trained embeddings, an *Extended Translator* using pretrained embeddings, and a *Transformer Translator* (Vaswani et al., 2017). We believe that the latter is beneficial regarding the out-of-vocabulary and polysemy issues described above, which is an inherent problem in the translation of sparse Sumerian fragments.

4.1 Base Translator

The architecture of the *Base Translator* is a standard sequence-to-sequence encoder-decoder model with attention (Bahdanau et al., 2015). In order to circumvent issues related to vanishing gradient problems during training (Hochreiter, 1998; Sherstinsky, 2018), we employed two stacked LSTM networks (Hochreiter and Schmidhuber, 1997) as basic building blocks in the proposed *Base Translator*. The inputs are the Sumerian source tokens and we used custom-trained English word vectors using word2vec (Mikolov et al., 2013) on all 1.5 million transliterations.

[1] `https://cdli.ucla.edu/`

[2] Throughout this paper, we use the term 'phrase' for a single, complete line in a document. In many cases, this will be a sentence or a clause, but it can also be a partial sentence, only. For training and evaluation, we exclude incomplete lines, so that the phrases of a single document do not necessarily constitute a complete text.

4.2 Extended Translator

The *Extended Translator* implements the same architecture as the *Base Translator* but instead of custom-trained embeddings for English on our small data set, we used pretrained embeddings from the much larger Wikipedia corpus (Pennington et al., 2014, GloVe). We used GloVe as initialization to the embedding layer in our model and experimented with different dimensionalities.

4.3 Transformer Translator

Inspired by the latest research using multi-head self-attention mechanisms in encoder-decoder-based architectures (Vaswani et al., 2017), we propose another adapted implementation in the form of a *Transformer Translator*, with an encoder and decoder, both stacked with six identical layers along with pretrained embeddings in the same way as the *Extended Translator*. Based on best practices and in order to make the model aware of positional information of Sumerian and English tokens, a position-dependent signal is employed to each word embedding to assist the architecture in capturing the original order of words. Initially, in the encoding step, a representation is generated for each token in a Sumerian phrase, from its word embedding and positional encoding, which is then fed into a sequence of six stacked layers with multi-head attention where position-wise feed forward networks with residual connections are employed between every two sub-layers. Finally, the input to the decoder phase is the output embedding and the positional encoding using a similar grouping of stacks of multi-head self-attention layers. The decoder generates one word at a time greedily in a left-to-right fashion.

4.4 Phrase-Based Machine Translation

As a large portion of our raw data set is monolingual it seems plausible to employ methods of phrase-based machine translation (Lample et al., 2018). For the English monolingual data, we used the Europarl data set (Koehn, 2005), and first created a bilingual dictionary leveraging the independent monolingual data sets by aligning a monolingual word embedding space in an unsupervised way as described by Conneau et al. (2017). Using this bilingual dictionary we populated the phrase tables for Sumerian to English and English to Sumerian. Then, we trained n-gram language models for the Sumerian and English domain using the methods outlined in Heafield (2011). In a later step, we improved these translation models using iterative back-translation (He et al., 2016).

4.5 Transfer Learning

Supervised machine translation relies on massive amounts of data, hence typically performs poorly on low resource languages. The idea of transfer learning (Zoph et al., 2016) is to train a machine translation model in a high-resource language setting, e.g., from French to English as a parent model and then initializing the training constraints using the parent model and apply it to the child model. In our experimentation, we first trained a French to English model on the Europarl Corpus using transformers, then trained our child model from Sumerian to English. The training procedure for the French–English model is identical to the one outlined in Section 4.3.

5 Results & Evaluation

All supervised models and experiments described in this paper were implemented using OpenNMT[3] (Klein et al., 2017). For the phrase-based and transfer learning techniques, we used FairSeq (Ott et al., 2019). All translation models described in the previous section were trained, tuned, and evaluated on the same standardized training, development and test splits, respectively. First, we calculated BLEU scores (Papineni et al., 2002) for Sumerian translations against the gold data using various settings. The best results obtained are shown in the second column of Table 1. Moreover, in a qualitative evaluation, two experts in Sumerian rated 50 randomly chosen translations from each model, using the following scored ranking schema: *good* [3], *helpful* [2], *incorrect* [1] with exact definitions given in the supplementary material. All average ratings are shown in the last column of Table 1. A few important observations can be made:

[3] `http://opennmt.net/`

(1) The *Base Translator* is outperformed by the *Extended Translator* in both evaluation settings. Using pretrained embeddings can thus boost the performance significantly over custom-trained in-domain embeddings. We believe that the English translations alone are too sparse to induce qualitative word representations. **(2)** The *Extended Translator* is the best-performing model (cf. Figure 2 for an attention visualization) and the *Transformer Translator* performs slightly worse. This is most likely due to the large number of parameters and the sparse data domain it has been trained on. **(3)** The iterative back translation step incorporated in the phrase-based setting for the generation of the target to source sentence within the monolingual corpus seems problematic for Sumerian due to the short phrases and the inherent sparsity in the raw data. **(4)** Although we achieved a BLEU score of 36.9 for French to English, Sumerian is an isolated language and does not share any lexical similarity with modern languages which might explain why transfer learning could not improve overall performance.

Model Architecture	BLEU	Expert
1. *Base Translator*	19.6	1.7
2. *Extended Translator*	**21.6**	**2.2**
3. *Phrase-Based Translator*	8.2	1.1
4. *Transformer Translator*	20.9	2.0
5. *Transfer Learning*	15.3	1.4

Table 1: Comparison of different translation models by BLEU scores and expert ratings.

6 Conclusion

We have described the first experiments using machine translation for transliterated Sumerian to English, experimented with various architectures and found that using pretrained word embeddings in sequence-to-sequence models with attention can achieve the best performance in our sparse data setting. In future research, we would like to focus on improving the quality of custom-trained embeddings, for both English and Sumerian, as we still see room for improvement in this regard, for instance, by consultation of external Sumerian corpora, e.g., literature (Robson, 1998). An evaluation of the translations suggested already promising results and our research will hopefully provide a broader audience access to the data, including academics from other disciplines apart from Assyriology. All corpora, translations, training, and evaluation procedures are publicly available[4].

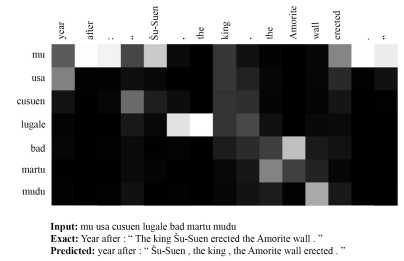

Figure 2: Sumerian-English attention weight visualization with `NUMB` placeholders for quantities.

Acknowledgments

The work described in this paper has been conducted in the context of the project 'Machine Translation and Automated Analysis of Cuneiform Languages' (MTAAC, 2017-2020), a collaborative project of the Universities of Toronto, Canada, the University of California, Los Angeles, US, and the Goethe-Unversity Frankfurt, Germany, funded as a Trans-Atlantic Platform Challenge Award by the National Endowment for the Humanities (NEH, US), the Social Sciences and Humanities Research Council (SSHRC, Canada) and the German Research Foundation (DFG, Germany).

We would like to thank collaborators of and contributors to the Cuneiform Digital Library Initiative (CDLI) that provides the data for our experiments and the umbrella for the activities of the MTAAC project. CDLI has been supported by the Google Summer of Code program, where aspects of machine translation have been addressed by several students since 2018. In particular, the first author conducted his work in the context of a Google Summer of Code project with the assistance of MTAAC project members and CDLI staff. The third author has been partially supported by the project 'Linked Open Dictionaries' (LiODi, 2015-2020), funded by the German Federal Ministry of Education and Research (BMBF) as an early carreer research group in eHumanities.

[4] `https://github.com/cdli-gh/Machine-Translation`

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hiram Calvo, Arturo P. Rocha-Ramirez, Marco A. Moreno-Armendáriz, and Carlos A. Duchanoy. 2019. Toward universal word sense disambiguation using deep neural networks. *IEEE Access*, 7:60264–60275.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Harriet Crawford and Crawford Harriet. 2004. *Sumer and the Sumerians*. Cambridge University Press.

Krupakar Hans and RS Milton. 2016. Improving the performance of neural machine translation involving morphologically rich languages. *arXiv preprint arXiv:1612.02482*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Fei Huang, Alexander Yates, Arun Ahuja, and Doug Downey. 2011. Language models as representations for weakly-supervised nlp tasks. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 125–134, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Yudong Liu, Clinton Burkhart, James Hearne, and Liang Luo. 2015. Enhancing Sumerian lemmatization by unsupervised named-entity recognition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1446–1451.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Émilie Pagé-Perron, Maria Sukhareva, Ilya Khait, and Christian Chiarcos. 2017. Machine translation and automated analysis of the Sumerian language. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–16, Vancouver, Canada, August. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Stuart Robinson, Greg Aumann, Steven Bird, et al. 2007. Managing fieldwork data with toolbox and the natural language toolkit. *Language Documentation and Conservation*, 1(1):44–57.

Eleanor Robson. 1998. *The electronic text corpus of Sumerian literature*. University of Oxford, Faculty of Oriental Studies.

Eleanor Robson. 2018. Lemmatising ATF files. `http://oracc.museum.upenn.edu/doc/help/lemmatising/`.

Alex Sherstinsky. 2018. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Max Weber. 1976. *The Agrarian Sociology of Ancient Civilizations*. London.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.