Extending the Vocabulary of Fictional Languages using Neural Networks

Thomas Zacharias *
Electrical Engineering
Tel Aviv University
thomasz@mail.tau.ac.il

Ashutosh Taklikar *
Electrical Engineering
Tel Aviv University
ashutosht@mail.tau.ac.il

Raja Giryes
Electrical Engineering
Tel Aviv University
raja@tauex.tau.ac.il

Abstract

Fictional languages have become increasingly popular over the recent years appearing in novels, movies, TV shows, comics, and video games. While some of these fictional languages have a complete vocabulary, most do not. We propose a deep learning solution to the problem. Using style transfer and machine translation tools, we generate new words for a given target fictional language, while maintaining the style of its creator, hence extending this language vocabulary.

Introduction

Languages can be broadly classified into natural and constructed languages. Fictional languages are constructed languages designed for a particular fictional setting - a book, movie, television show, or video game. There is a large demand for fictional languages in the entertainment industry with many of these languages being used in blockbusters, video games, comic books, and novels.

Some of these languages, such as Na'vi and Dothraki, are constructed by professional linguists called conlangers. Yet, there are plenty of fictional works where the artists are unable to employ conlangers and need to design a language on their own. This is particularly true for fictional books, which are typically the creation of a single person. While there are exceptional authors like J.R.R. Tolkein who created multiple complete fictional languages such as Quenya and Sindarin for The Lord of the Rings, most authors usually create a limited dictionary for their fictional language consisting of a few hundred words with their translations to English.

Using neural networks based style transfer techniques as a word generation tool, we demonstrate the ability to extend the vocabulary of fictional languages, i.e. given a limited dictionary of a few hundred words and their translations, our network is able to extrapolate the vocabulary of the language while maintaining the style of the creator.

Methods

For style transfer, we employ the work of Shen et al. [2017] that performs a refined alignment of sentence representation across text corpora. While originally used on sentences, we achieve style transfer at the word level by representing each word as a sequence of characters. For training the network, datasets with words written in two different languages are used.

To measure network performance, we tested the network on natural languages where English acted as our fictional language. During training, the network was fed with 850 English words (target language) [Norman, 2019] and their translations to another natural language (source language) so that it could learn the underlying style and content representations. The trained network can generate new words

*T. Z. and A. T. contributed equally to this work.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

in the target language by inputting words from the source language that the network never saw. On generating 100 new English words, we saw that the network was able to produce 20 words that are already a part of the English vocabulary (the network did not see these words during training). The network also proposed 67 unique words that currently do not belong to the English dictionary. In total, on asking the network to translate 100 words, the network proposed 87 new English words. It is important to note that we worked with hundreds of words during training which is approximately the number of words one would find in a fictional dictionary. A thorough quantitative and qualitative analysis can be found in the supplementary material.

We also used machine translation models like seq2seq, RNN, and transformers for the novel problem of vocabulary generation. These networks act as a baseline (see supplementary) comparison to our proposed model of using style transfer as a vocabulary generation tool.

Results

To extend the vocabulary of a fictional language, we chose the "Ancient Language", created by Christopher Paolini for the Inheritance cycle. He invented 425 words in the "Ancient language" which defined the size of our training dataset. Once the network was trained, we gave the network English words that were never defined in the Ancient Language and asked the network to propose translations. Table 1 shows some of our results. By comparing the Paolini's definitions and our networks proposals, we can see that the network is clearly able to generate words while maintaining the style of Christopher Paolini's original vision for the language. See supplementary for more results.

Christopher Paolini's definitions		Our network's proposals			
English	A. L.	English	A. L.	English	A. L.
water	adurna	pokemon	pokenor	baseball	blaskalí
celebration	agaetí	computer	pumortor	soccer	sotvert
blood	blödh	artificial	arifinaskil	basketball	blaskelblaka
birth	burthr	intelligence	iletningr	aeroplane	areltarín
day	dag	overleaf	torelvarí	train	traina
deaf	daufr	python	dornth	missile	misilves
ladies	ementyr	conference	fornengr	electricity	eletringa
forward	fram	deep	deyy	television	ethelingr
sorrow	harmr	learning	elvarning	religion	relthinga
true	ilumaro	president	redrisn	professor	proesors
stop	letta	democracy	demdratory	domino	dumothrime
lock	laesa	earth	eathrí	cap	aka
mother	menoa	internet	tentring	bat	bata
orchid	niernen	laptop	latoka	keyboard	kebraoda
warp	orpin	smartphone	smathringa	violin	vilino

Table 1: **Ancient language (A. L.).** English words and their translation into the Ancient Language. The left side corresponds to words that Christopher Paolini defined. The right side contains our network's proposals for translations of English words that Christopher Paolini never defined.

Conclusion

We proposed a deep learning based approach to enable the completion of fictional language vocabularies. We showed how style transfer can be used as a vocabulary generation tool; proposed ways to analyze the outputs; and successfully demonstrated the extension of a fictional vocabulary.

While we relied on the technique from Shen et al. [2017] to perform the new words generation, one may use other textual style transfer frameworks in a similar way [Zhao et al., 2017, Li et al., 2018, Zhang et al., 2018, Li et al., 2019, Akama et al., 2017]. Apart from the obvious applications in the entertainment industry, our proposed methodology could also be relevant for translating modern words such as internet, modem, etc, to existing languages in an automated way that takes into account the style of the target language.

References

Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. Generating stylistically consistent dialog responses with transfer learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–412, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-2069.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. Domain adaptive text style transfer. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/d19-1325. URL http://dx.doi.org/10.18653/v1/d19-1325.

Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1169. URL https://www.aclweb.org/anthology/N18-1169.

David Norman. 1000 most common english words spoken in united states of america, https://gist.github.com/deekayen/4148741, 2019.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841, 2017.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. Learning sentiment memories for sentiment modification without parallel data. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi: 10.18653/v1/d18-1138. URL http://dx.doi.org/10.18653/v1/d18-1138.

Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized autoencoders, 2017.

Extending the Vocabulary of Fictional Languages using Neural Networks - Supplemental

Thomas Zacharias *
Electrical Engineering
Tel Aviv University
thomasz@mail.tau.ac.il

Ashutosh Taklikar *
Electrical Engineering
Tel Aviv University
ashutosht@mail.tau.ac.il

Raja Giryes
Electrical Engineering
Tel Aviv University
raja@tauex.tau.ac.il

Methods

As a baseline comparison we used Seq2Seq, RNN, and transformer models to analyze network performance.

Word generator using Seq2Seq. We use seq2seq models for the novel problem of vocabulary generation by training the model on a pair of languages using parallel word translations. We use an encoder-decoder architecture implemented using GRUs [Cho et al., 2014]. The input is given as a sequence of characters to the encoder that assigns attention based weight to each character using the attention model in [Bahdanau et al., 2014]. This results in the encoder output and hidden state which are inputted to the decoder along with the decoder start token. The decoder then returns a prediction along with the decoder hidden state which is passed back to the model. We use teacher-forcing, where the target character is passed iteratively to the decoder, to make it converge faster and train better.

Word generator using an RNN baseline. We also trained a character level RNN and analyzed its outputs. We trained the RNN model on all the words available in the target language. The trained RNN was then given the first character of the desired new target word and was required to generate an output. For making a dictionary-like model, where an input source word is mapped to the target language, the initial character of the target language was based on a combination of the characters of the source word to be translated (see illustration in Fig. 1). We used the mean of the character embeddings of the source word, but any other indicator could be used.

We note a few disadvantages of the above method: (i) Indicator selection: By using the mean of the characters embeddings as an indicator, there is a high chance that multiple source words would have the

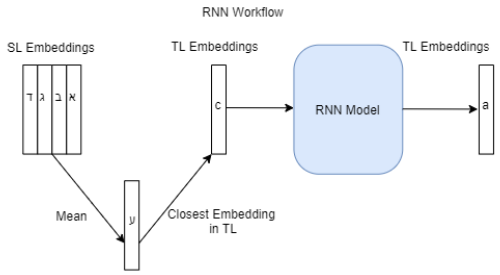

Figure 1: **RNN workflow.** The mean of the source language word embeddings is taken and the closest match in the target language domain is found. This embedding is then passed as the initial character to the RNN model.

*T. Z. and A. T. contributed equally to this work.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

English word	Proposed Hebrew translations			Hebrew word	Proposed English translations		
	RNN	Seq2Seq	S.T.		RNN	Seq2Seq	S.T.
silver	ליפורצ	להתמודד	להכות	להתמודד	hour	ale	seaply
thank	טור	משחות	לאין	לשחות	ing	starv	wole
branch	זור	צמות	סרגמה	טווח	jout	poper	strice
match	חור	מול	משלה	מול	jon	ap	seppare
suffix	להיל	חשה	לשתתת	אשה	keat	the	merd
especially	ילהת	לא	קעלאוון	נעל	got	gurn	fillice
fig	זור	כתף	מי	כתף	joun	wn	rore
afraid	וליל	התפוכית	אתרנה	התפשטות	ing	seve	persice
huge	טורה	ססר	לשר	לסדר	ing	treat	beat
sister	לה	מפשור	ללמתח	מחנה	goter	ent	moun
steel	דיר	להמצה	לעות	להמציא	hour	mridy	are
discuss	כול	כותנה	התלאל	כותנה	ing	put	tear
forward	דיר	נפוד	תתפרה	נולד	hour	bubline	rotunt
similar	דיר	תיייי	לשמות	לקבוע	keat	wet	beet
guide	טור	רבעגורורורור	מפרה	רבע גלון	gat	win	wilk
experience	ילה	תחוע	קערורה	תשע	got	tet	st
score	דירה	משומיות	לעור	משאית	hour	twice	at
apple	טיר	רעש	אוור	רעש	fine	dide	brant

Table 1: **Proposed translations using RNN, Seq2Seq, and Style Tranfer (S.T.) methods.** (Column I) English words we want translated. Proposed Hebrew translations using (column II) RNN; (column III) Seq2Seq network; (column IV) Style Transfer network. (Column V) Hebrew words we want translated. Proposed English translations using (column VI) RNN; (column VII) Seq2Seq network; (column VIII) Style Transfer network. Words marked in green are preexisting English words that were generated but the network saw during training. Words marked in blue are preexisting English words that were generated but the network never saw. Words marked in red are generated words that occur more than once in the column.

same embedding resulting in the same words in the target language for different source words, which is clearly undesirable. (ii) Limited dataset: Since our end goal is to extrapolate fictional languages, we have to work with just hundreds of words. Most character level language models, however, work with more examples (on the order of magnitudes higher). These constraints demonstrate the limitations of using a simple RNN.

Word generator using transformers. Transformer based language models were also used but these performed poorly due to the limited size of the dataset and hence their results have been ommitted.

Experiments

To perform vocabulary generation, datasets with words written in the two different languages are required. During training, each network was fed with a pair of language datasets so that it could learn the underlying content representations. Once the networks are trained, we tested the network by inputting words in the source language that the network never saw and asking the network to "translate" the words to the target language.

To measure network performance, we first tested the network on natural languages. We used multiple language pairs. In this section, we focus on the results for the English-Hebrew pair. We selected the 850 most common English words spoken in the United States [Norman, 2019] and their translations in Hebrew. The networks are then trained over the English-Hebrew language pair. At inference, we simply provide an English word to obtain a "Hebrew translation", and vice versa. By limiting the dataset to hundreds of words, we simulate the approximate size of a fictional language vocabulary.

Table 1 displays English words with the proposed Hebrew translations and Hebrew words with their proposed English translations using the three networks. By analyzing the words generated, one can note that the Style Transfer and Seq2Seq networks display a wide range in character usage, word length, and have no repetitions. The RNN however performs relatively poorly demonstrating the need

English word we want translated	Proposed translations				
	Arabic	Hindi	Spanish	Amharic	Russian
silver	شرة	विड़का	silverian	ሆንባት	пресносать
thank	الصادة	पाड़ी	thanak	ረሩ	строл
branch	ذرس	प्रित्ध	brícancak	ተነገም	попрастить
match	قائرة	पूला	macta	ቅር	прость
suffix	متطفة	प्रसिजक	sufificar	መለሰለም	рестивитьс
especially	عيدية	कालाप्नीकीजिये	sespriclama	እንተለቀል	нстроммемнать
fig	موق	सिक	figa	አባ	меть
afraid	ماوق	प्रत्र	faricadiar	ተገሰኝ	постедиться
huge	جل	प्या	guego	ደጠሻ	вель
sister	حر	हिलाना	sistrecian	እንፈዳም	ресностьс
steel	قطيدة	हाज़ी	sentelo	ንግል	норани
discuss	تعرد	मेप्रनाने	discursina	እለረን	перерурустый
forward	كبمعة	सेतर्त	forricadar	አንገት	поставиться
similar	مواع	नाल्की	similarian	አንገድል	престостени
guide	عيم	मुक्य	guedido	አዋታ	дежить
experience	يصار	काप्ताकिनिके	experciende	እያንተለ	нстороменный
score	المن	खिबा	bisación	እንነት	проза
apple	المينة	पूर्का	sprepadía	ተቀሳል	порона

Table 2: **Proposed translations.** English words and their proposed translations to Arabic, Hindi, Spanish, Amharic, and Russian. Note that a different trained network was used for each target language.

Network	Dataset	Quantiative analysis			Qualitative analysis	
	E-H words	E.W.N.S.	P.N.W.	T.U.W.	Mean	Median
Seq2Seq	200-200	145	668	813	2.6	3
Style transfer	200-8000	184	842	1026	3.1	3
Seq2Seq	850-850	217	740	957	3.4	4
Style transfer	850-8000	191	895	1086	3.3	3

Table 3: **Quantitative and Qualitative analysis.** (Column I) Networks used for word generation. (Column II) The English-Hebrew dataset network was trained on. (Column III) (E.W.N.S.) Generated preexisting English Words that were Not Seen during training. (Column IV) (P.N.W.) Proposed novel Words currently not in the English vocabulary. (Column V) (T.U.W) Total Unique Words generated. (Column VI) Mean score (out of 5) from user study. (Column VII) Median score (out of 5) from user study.

for the more advanced solutions we use. Notice that the RNN proposed a few English words that exist in the language, however, the network had already seen these words during training so these cannot be considered as novel generated words. Such words, that exist in the language, but the networks already saw, are marked in green in Table 1. Red color marks repetitions of words that have already occurred in the table and blue color marks preexisting English words that were generated but the network never saw. It is interesting to note that all networks generated words that are short and have vowels - qualities that are both desirable.

As an added measure, we performed a user study to have a qualitative comparison of the Seq2Seq and Style Transfer networks.

Discussion

User Study. For the user study, we generated words from the seq2seq and style transfer networks trained on different datasets (see table 3) and asked judges to rate them from 0 (cannot be an English word) to 5 (definitely an English word). Each network trained on the different datasets as seen in table 3 was asked to generate 1300 new words. We conducted a quantitative analysis on the generated words by analyzing the number of preexisting English words and the total number of unique words that the

English word we want translated	Proposed Hebrew translation
block	סוין
blockade	סובינון
throw	להיפור
threw	להיפר
throwing	להיינור
like	נובה
likely	נסובה
hear	לקרון
heard	להראון
hearing	להרגמי
task	מולה
tasking	למימובי
wise	נפר
wisdom	נופה
imagine	מאפרים
imagined	מהגאית
imagination	מהלאימון

Table 4: **Linguistic connections.** English words originating from the same root and their proposed Hebrew translations.

English word we want translated	Proposed Ancient language translation
understand	dunendarvald
understood	duendorva
understanding	dunendanvaldrina
fiction	fitingoting
fictional	fitingatilona
revolt	relthilg
revolution	reluthingo
revolutionary	eruthilnadorin
frown	frongh
frowned	ffrendhr
frowning	frönnghing
imagine	minandae
imagined	mindanaedí
imagination	mingainatoind
generate	engaerth
generated	geengardethin
generation	gengarinona
generating	gengarnathin

Table 5: **Linguistic connections.** English words originating from the same root and their proposed Ancient language translations.

network could generate. The results for the same can be viewed in table 3. It can be seen that the style transfer network for both datasets is able to generate more unique words, however the seq2seq network trained on the 850 word language pair performs slightly better while generating preexisting English words.

To analyze the quality of the output words, 48 judges took part in the user study and a total of 2400 words were rated- 600 each corresponding to the rows in table 3. Qualitatively, the seq2seq performs comparably well to the style transfer network when trained on the dataset with 850 English words. However, both quantitatively and qualitatively, style transfer easily outperforms Seq2Seq when it comes to the smaller dataset containing 200 English words. This is because the style transfer network is able to take advantage of the unpaired style of training unlike the seq2seq network. Since the size of fictional datasets are largely limited, using the style transfer network should be preferable.

Testing Style Transfer network on multiple languages. In addition to the in-depth analysis for the Hebrew-English pair, we also conducted multiple experiments with our style transfer network for different language pairs. For the convenience of readers around the world, we provide in Table 2 English words along with our network's proposed novel generated words in Arabic, Hindi, Spanish, Amharic, and Russian. In the two languages that we understand in the table, we observe the same resemblance to the target language that we had with the generation of words in English and Hebrew. We believe that the readers who understand the languages will observe the same when going through the results in Table 2. All these results demonstrate the capability of our proposed strategy in creating novel words in a given language that contains only a small number of existing words.

Linguistic connections. We also investigated linguistic connections between the translations of words from the same root. We selected English words originating from the same root and asked the trained style transfer network to propose Hebrew translations. Table 4 shows English words and their proposed translations. As can been seen in the table, the proposed Hebrew translations also share common word roots which strengthens our proposition of using style transfer for novel word generation. It is very interesting that the network was able to achieve this despite the non-parallel nature of training and the small number of words used from target language.

We also investigated linguistic connections between the proposed translations of words from the same root. Table 5 shows a few examples of English words belonging to the same root along with the

Christopher Paolini's definitions		Our network's proposals			
English	Ancient Language	English	Ancient Language	English	Ancient Language
create	aldanarí	artisan	rathans	torch	throch
elf	älfa	historian	istharina	constellation	nesthalona
halt	blöthr	modest	medts	racism	varsia
reduce	brakka	pony	ponn	forbid	frodhi
honor	celöbra	grimace	gramia	feign	feing
sage	chetowä	redundancy	renduvandí	slant	skant
brother	darmthrell	decisive	desviverna	ambition	maithinor
mists	datia	abundant	bandanthar	coerce	ceror
grow	eldhrimner	teenager	ternanga	aisle	iales
invoke	ethgrí	harsh	harsh	patience	vaethinva
marked	fódhr	fragment	franthandí	fisherman	frishvandr
sing	fyrn	pigeon	gieno	bathtub	bathhr
chant	gala	soprano	sorana	lend	lend
luck	guliä	freckle	frelka	rubbish	brisha
height	haedh	curl	völrr	shareholder	sharelhra
whale	hwal	continuation	onvithanda	decay	deyja
my	iet	housewife	shvaedhrin	deter	deetr
truth	ilumëo	stir	stirr	feminine	ferininve

Table 6: **Ancient language.** English words and their translation into the Ancient language. Left: Words that Christopher Paolini defined. Middle&Right: Our network's proposals for words that were never defined.

proposed translations. Notice how the proposed translations also have common word roots, which corroborates our hypothesis.

Ancient Language. Additional examples of words generated for the Ancient Language by the style transfer network can be viewed at table 6. These examples were generated by training on a dataset of 400 Ancient Language words and 8000 English words. We could use more English words since we could exploit the nature of non-parallel dataset of the style-transfer network to further improve the results.

Using non-parallel data. It is worth noting that by using the non-parallel training approach, we are not giving up linguistic information provided by parallel links. As shown by the linguist De Saussure [2011], meaning-to-form mapping is almost entirely arbitrary and there is no simple or logical relation between a certain word and its translation (for languages that are not phonologically related). The only constraints at play are the phonological and morphological rules of a language. By using style transfer, the aim is for the network to learn these phonological and morphological rules (the underlying style) which we believe that the network was able to achieve.

Conclusion

We proposed a deep learning based approach to enable the completion of fictional language vocabularies. We showed how style transfer can be used as a vocabulary generation tool; proposed ways to analyze the outputs; and successfully demonstrated the extension of a fictional vocabulary.

While we relied on the technique from Shen et al. [2017] to perform the new words generation, one may use other textual style transfer frameworks in a similar way [Zhao et al., 2017, Zhang et al., 2018, Li et al., 2019]. We also conducted baseline comparisons by using using RNN, seq2seq, and transformer models.

Apart from the obvious applications in the entertainment industry, our proposed methodology could also be relevant for translating modern words such as internet, modem, etc, to existing languages in an automated way that takes into account the style of the target language.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Ferdinand De Saussure. *Course in general linguistics*. Columbia University Press, 2011.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. Domain adaptive text style transfer. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/d19-1325. URL `http://dx.doi.org/10.18653/v1/d19-1325`.

David Norman. 1000 most common english words spoken in united states of america, https://gist.github.com/deekayen/4148741, 2019.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841, 2017.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. Learning sentiment memories for sentiment modification without parallel data. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi: 10.18653/v1/d18-1138. URL `http://dx.doi.org/10.18653/v1/d18-1138`.

Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized autoencoders, 2017.