NeuralNetworkModelsForTranslatingInput Sequences To Output Sequences

HETA DESAI

Abstract

Deep Neural Networks (DNNs) have shown strong performance on complex learning tasks, especiallywhenlargelabeleddatasetsareavailable.However,theystrugglewithmappinginput sequences directly to output sequences. In this paper, the authors propose a general,
end-to-end method for sequence learning that relies on minimal assumptions about the structure of sequences. Their approach uses a multi-layer Long Short-Term Memory (LSTM) networktoencodeaninputsequenceintoafixed-sizevector,followedbyanotherdeepLSTM network that decodes this vector into an output sequence.

AkeyresultoftheirstudyisthatthisLSTM-basedmodelachievedaBLEUscoreof34.8onthe fulltestsetforEnglish-to-FrenchtranslationusingtheWMT'14dataset,despitebeingpenalized for generating out-of-vocabulary words. Notably, the model performed well even on long sentences. For comparison, a phrase-based Statistical Machine Translation (SMT) system scored 33.3 on the same dataset.

FurtherimprovementwasobservedwhentheLSTMwasusedtorerankthetop1000translation hypotheses from the SMT system, raising the BLEU score to 36.5—close to the best result reported at the time.The LSTM also learned meaningful representations of phrases and sentences that were sensitive to word order and robust to changes like converting between active and passive voice.

Interestingly, performance improved significantly when the order of words in the source sentenceswasreversed(whilekeepingtargetsentencesinnormalorder).Thiscreatedmore short-term dependencies between source and target sequences, which helped make the training process more effective.

1. Introduction

Deep Neural Networks (DNNs) are highly effective machine learning models that have demonstrated strong results on complex tasks like speech and visual object recognition. Their strengthliesintheirabilitytocarryoutcomplexparallelcomputationswithinalimitednumberof steps. For example, it has been shown that a neural network with just two hidden layers of quadratic size can sort N N-bit numbers—highlighting the expressive power of DNNs. Though DNNs are related to traditional statistical models, they are capable of learning complex computations. When trained with supervised backpropagation and provided with sufficientlabeleddata,DNNscanlearnparametersettingsthatallowthemtoexcelattasksthatare relatively easy for humans to perform.

However,amajorlimitationofDNNsisthattheytypicallyrequireinputandoutputtobe fixed-length vectors. This restricts their application to tasks where the structure can be
compactlyrepresentedinthisway.Manyreal-worldproblems—likemachinetranslation,speech recognition,andquestionanswering—involvesequencesofvaryinglengths,makingitdifficultto apply traditional DNN architectures directly. As such, a general-purpose method that can learn to map sequences of arbitrary lengthto other sequences would be highly beneficial.

Sequences are challenging for standard DNNs because they demand a flexible input-output dimensionality. This paper presents a straightforward yet powerful approach to handling such tasks using the Long Short-Term Memory (LSTM) architecture. The approach involves using oneLSTMnetworktoreadandencodetheinputsequenceintoafixed-sizevector,andanother LSTMtodecodethevectorbackintoanoutputsequence.ThesecondLSTMfunctionssimilarly to a language model but is conditioned on the encoded input.

Thismethodisespeciallysuitedforproblemswithlong-rangedependenciesbetweeninputsand outputs, thanks to LSTM's ability to remember information over extended time steps. Prior work has explored similar goals. For example, Kalchbrenner and Blunsom were among the first to encode an input sentence into a fixed vector, and Cho et al. applied similar ideas for re-ranking translation hypotheses. Additionally, Graves proposed a differentiable attention mechanism that inspired more advanced translation systems like that of Bahdanau et al. Another technique, Connectionist Sequence Classification (CTC), has also been used for sequence learning, though it relies on a monotonic alignment between input and output sequences.

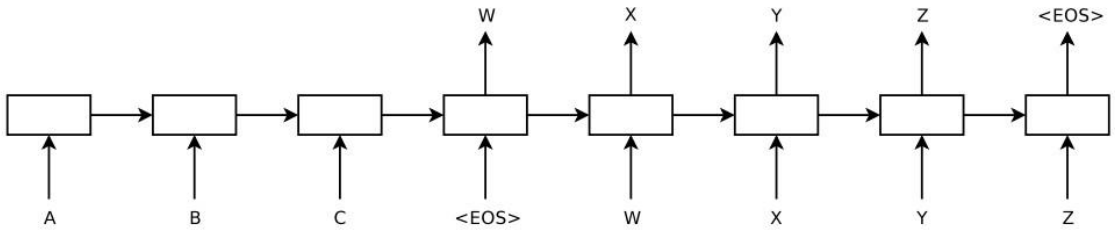

Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

The key finding of this study is that, on the WMT'14 English-to-French translation task, the model achieved a BLEU score of **34.81** using an ensemble of five deep LSTM networks (each with384millionparametersand8,000-dimensionalstates).Translationsweregeneratedusinga straightforward left-to-right beam search decoder. This score represents the highest result for directtranslationusinglargeneuralnetworkstodate.Forcomparison,astandardphrase-based Statistical Machine Translation (SMT) system scored **33.30** on the same task. Notably, the LSTM model used a fixed vocabulary of 80,000 words, so any word not in this vocabulary led toapenaltyintheBLEUscore.DespitethislimitationtheLSTMoutperformedtheSMTsystem, indicating that even a relatively basic neural architecture with room for improvement can surpass traditional systems.

Additionally,theLSTMwasusedto**rescore**the1,000-besthypothesesgeneratedbytheSMT system,pushing the BLEU score up to **36.5**—an improvement of 3.2 points over the baseline and nearly matching the previous best published score of **37.0**.

Interestingly, the LSTM handled **long sentences** well, even though similar architectures often strugglewiththis.Thesuccessislargelyattributedtoasimplebuteffectivestrategy:**reversing the order of words** in the source sentences during training and testing (while keeping the target sentences unchanged). This introduced short-term dependencies that made learning easier and allowed the LSTM to optimize more effectively using stochastic gradient descent (SGD).This reversal trick is considered a key technical innovation of the paper.

AnotherstrengthoftheLSTMmodelisitsabilitytorepresentvariable-lengthinputsentencesas **fixed-size vectors**. Since translation requires capturing the **meaning** of a sentence, the model is encouraged to produce embeddings where semantically similar sentences are close together in this learned space. Qualitative analysis shows that the model is sensitive to **word order** and robust to variations in sentence structure, such as **active vs. passive voice**.l

2. Themodel

The Recurrent Neural Network (RNN) [31, 28] is a natural generalization of feedforward neural networkstosequences.Givenasequenceofinputs(x1,...,xT),astandardRNNcomputesa sequence of outputs (y1, . . . , yT) by iterating the following equation:

$$\begin{aligned} h_t &= \text{sigm}\left(W^{\text{hx}}x_t + W^{\text{hh}}h_{t-1}\right) \\ y_t &= W^{\text{yh}}h_t \end{aligned}$$

RecurrentNeuralNetworks(RNNs)arewell-suitedformappingsequencestosequenceswhen the alignment between the input and output is already known. However, they become challenging to apply when the input and output sequences are of different lengths and the relationship between them is complex or non-monotonic.

A straightforward approach for handling general sequence-to-sequence problems is to use one RNN to encode the input sequence into a fixed-size vector, and then use a second RNN to decodethatvectorintotheoutputsequence—anideaalsoexploredbyChoetal.[5].Whilethis setup theoretically provides the necessary information for the model to perform the task, it is difficult to train due to the long-term dependencies involved (see figure 1) [14, 4, 16, 15]. However,LongShort-TermMemory(LSTM)networks[16]arespecificallydesignedtohandle long-range temporal dependencies, making them a promising solution for this problem.

The main objective of the LSTM in this context is to model the conditional probability p(y1,...,yT|x1,...,xT)p(y_1,...,y_{T'}|x_1,...,x_T)p(y1,...,yT|x1,...,xT),where(x1,...,xT)(x_1, ...,x_T)(x1,...,xT)istheinputsequenceand(y1,...,yT)(y_1,...,y_{T'})(y1,...,yT)isthetarget sequence,whichmaydifferinlength.Theprocessinvolvestwosteps:first,theLSTMencodes the input sequence into a fixed-length vector vvv, derived from its final hidden state. Then, a second LSTM acts as a language model, using vvv as its initial hidden state to generate the probabilitydistributionovertheoutputsequencey1,...,yTy_1,...,y_{T'}y1,...,yT.

$$p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, \ldots, y_{t-1})$$

Inthisapproach,eachprobabilityp(yt|v,y1,...,yt−1)p(y_t|v,y_1,...,y_{t-1})p(yt|v,y1,...,yt−1)is calculated using a softmax function applied over the entire vocabulary. The LSTM architecture follows the design from Graves [10]. Importantly, each sentence must end with a special end-of-sentencetoken(“<EOS>”),allowingthemodeltodefineaprobabilitydistributionover sequences of any length. Figure 1 illustrates this idea, where the LSTM processes the sequence “A”, “B”, “C”, “<EOS>” to generate a fixed representation, which is then used to predict the sequence “W”, “X”, “Y”, “Z”, “<EOS>”.

However,theactualimplementationofourmodeldiffersfromthisbasicsetupinthreekeyways:

1. **Separate LSTMs for Encoding and Decoding**: We use one LSTM for encoding the input sequence and a separate LSTM for decoding the output. This not only increases thenumberofmodelparameterswithoutaddingsignificantcomputationalcost,butalso facilitates training on multiple language pairs at the same time [18].
2. **DeeperArchitecturesPerformBetter**:WediscoveredthatdeepLSTMsoutperform shallow ones. As a result, our model uses an LSTM with four layers.
3. **InputSentenceReversal**:Acrucialimprovementwasreversingthewordorderofthe input sentence. For instance, instead of training the model to map “a, b, c” to its translation“α,β,γ”,wetrainittomap“c,b,a”to“α,β,γ”.Thisreversalalignsthe beginningoftheinputwiththebeginningoftheoutput(e.g.,“a”iscloserto“α”),which helpsthemodellearnmoreefficiently.Thissimplechangesignificantlyboosted LSTM performance.

3. Experiments

We evaluated our approach on the WMT'14 English-to-French machine translation task using two main strategies. First, we used our model to translate input sentences directly, without relying on any traditional statistical machine translation (SMT) system. Second, we used our model to rescore the n-best candidate translations produced by a baseline SMT model. We reporttheperformanceofbothmethods,provideexampletranslations,andexplorethelearned sentence representations.

3.1 DatasetDetails

Fortraining,weusedtheWMT'14English-to-Frenchdataset.Ourtrainingsubsetincluded12 million sentence pairs, with approximately 348 million French words and 304 million English words. This dataset was chosen because it's a cleaned and preprocessed version made available by [29], which also provides tokenized training and test data alongwith 1000-best translation hypotheses from the SMT baseline.

Since neural language models rely on word embeddings, we used fixed-size vocabularies for each language. Specifically, we selected the 160,000 most frequent words for English (the sourcelanguage)andthe80,000mostfrequentwordsforFrench(thetargetlanguage).Words not found in these vocabularies were replaced with a special "UNK" (unknown) token.

3.2 DecodingandRescoring

Themainfocusofourexperimentswastrainingalarge,deepLSTMusingnumeroussentence pairs. The training goal was to maximize the log-likelihood of producing the correct target sentence T given the source sentence S. In other words, the model was optimized to increase the probability of generating accurate translations.

$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

whereSisthetrainingset.Oncetrainingiscomplete,weproducetranslationsbyfindingthe most likely translation according to the LSTM:

$$\hat{T} = \arg\max_{T} p(T|S)$$

To find the most likely translation, we use a straightforward left-to-right beam search decoder. This decoder keeps track of a limited number BBB of partial hypotheses at each step, where each partial hypothesis is a prefix of a potential translation. At every timestep, each hypothesis in the beam is extended by appending every possible word from the vocabulary. Since this generatesalargenumberofnewhypotheses,weretainonlythetopBBBhypothesesbasedon their log probability scores. Once a hypothesis includes the special end-of-sentence token <EOS>,itisremovedfromthebeamandaddedtothesetofcompletetranslations.Althoughthis method is not exact, it is easy to implement. Notably, even with a beam size of 1, our model performs well, and most of the beam search's benefits are retained with a beam size of just 2 (as shown in Table 1).

We also applied our LSTM to rescore the 1000-best candidate translations produced by the baselineSMTsystem[29].Foreachhypothesisinthelist,wecalculateditslog-likelihoodusing our LSTM and averaged this score with the one from the SMT system.

3.3 ReversingSourceSentences

Even though LSTMs are designed to handle long-range dependencies, we observed

significantlybettertrainingresultswhenwereversedtheorderofwordsinthesourcesentences (while keeping target sentences unchanged). With this reversal, test perplexity improved from 5.8to4.7,andtheBLEUscorerosefrom25.9to30.6.

Although we don't have a full explanation for why this works, we suspect it's because reversing the source sequence creates more short-term dependencies in the data. Typically, words in the source sentence are positioned far from their translated counterparts in the target sentence, leading to a high "minimal time lag" [17]. Reversing the source order doesn't change the average word-to-word distance, but it reduces the time lag for the initial parts of the sentences. This makes it easier for the model to learn connections between source and target sequences duringbackpropagation,whichlikelyaccountsforthenotableimprovementintranslationquality.

At first, we assumed that reversing the input sentences would only improve the model's confidencefortheinitialpartsofthetranslatedoutput,possiblyweakeningitspredictionstoward the end. However, we found that LSTMs trained on reversed input sequences performed significantly better on long sentences than those trained on the original order (see Section 3.7). This suggests that reversing the input helps the LSTM make more efficient use of its memory.

3.4 TrainingDetails

TrainingtheLSTMmodelsturnedouttoberelativelystraightforward.WeuseddeepLSTMswith 4layers,eachlayercontaining1,000LSTMunitsand1,000-dimensionalwordembeddings.The input vocabulary had 160,000 words, while the output vocabulary had 80,000. This means each sentence was represented using 8,000 values. Deep LSTMs showed a clear advantage over shallow ones, with each additional layer reducing perplexity by nearly 10%—likely due to the increased capacity of the deeper models. A standard softmax was used over the 80,000-word output vocabulary. The entire model had 384 million parameters, including 64 million dedicated to recurrent connections—32 million for the encoder and 32 million for the decoder.

Keytrainingsettingsincluded:

- InitializingallLSTMparametersusinganuniformdistributionbetween-0.08and0.08.
- Usingstochasticgradientdescent(SGD)withoutmomentum,startingwithafixed learningrateof0.7.After5epochs,wehalvedthelearningrateeveryhalf-epoch, training for a total of 7.5 epochs.
- Trainingwasdoneinbatchesof128sequences,andthegradientwasscaled accordingly.
- Although LSTMs typically avoid vanishing gradients, they can suffer from exploding gradients.Tomanagethis,weappliedgradientclipping:iftheL2normofthegradient $s=\|g\|2$ exceeded 5, we scaled it as $s >5$, we set $g = 5g/s$.
- Since sentence lengths vary—most are between 20 to 30 words, while some exceed 100—randomminibatchestendedtobeinefficient.Tosolvethis,wegroupedsentencesofsimilarlengthtogetherineachbatch,whichd oubledtrainingspeed.

3.5 Parallelization

Initially, our C++ implementation of the deep LSTM (with the configuration described above) couldprocessaround1,700wordspersecondonasingleGPU.Thiswasn't fastenough,sowe parallelized the model across an 8-GPU system. Each of the four LSTM layers was run on a separate GPU, and activations were passed to the next layer immediately after computation.

The remaining four GPUs handled the softmax computation, with each GPU responsible for multiplyinga1000×20,000matrix.Thisoptimizedsetupachievedaprocessingspeedof6,300 words per second (including both English and French words) with a minibatch size of 128.

Overall,trainingthemodeltookabouttendaysusingthisparallelizedsetup.

3.6 ExperimentalResults

Weassessedthetranslationqualityusingthe**casedBLEUscore**[24].Thiswascalculatedwith the multi-bleu.plscript on tokenized predicted translations and their referencecounterparts. This evaluation method aligns with those used in [5] and [2], and accurately reproducesthe33.3BLEUscorereportedin[29].Interestingly,whenweapplythesamemethod to evaluate the best-performing system from WMT'14 [9] (whose output isavailable at statmt.org/matrix), we obtain a BLEU score of **37.0**, which is higher than the **35.8** reported on the website.

ThemainresultsareshowninTables1and2.Thebestperformancecamefroman**ensemble of LSTM models**—each initialized differently and trained with different random minibatch orderings. Although this ensemble's direct translation results do not surpass the top WMT'14 system, it marks a significant achievement: this is the **first time a fully neural machine translation system has outperformed a phrase-based SMT baseline** on a large-scale machinetranslationtaskbyanotablemargin.Thiswasachieved**despitelimitations**,suchas not being able to translate out-of-vocabulary (OOV) words.

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	∼45

Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

WhentheLSTMmodelisusedto**rescore**the1000-bestlistgeneratedbytheSMTbaseline,it achieves a BLEU score within **0.5 points** of the top WMT'14 system—highlighting its strong performance.

3.7 PerformanceonLongSentences

To our surprise, the LSTM model performed well even on longer sentences, as demonstrated bythequantitativeresultsinFigure3.Table3highlightssomeexamplesoflonginputsentences along with their corresponding translations.

3.8 ModelAnalysis

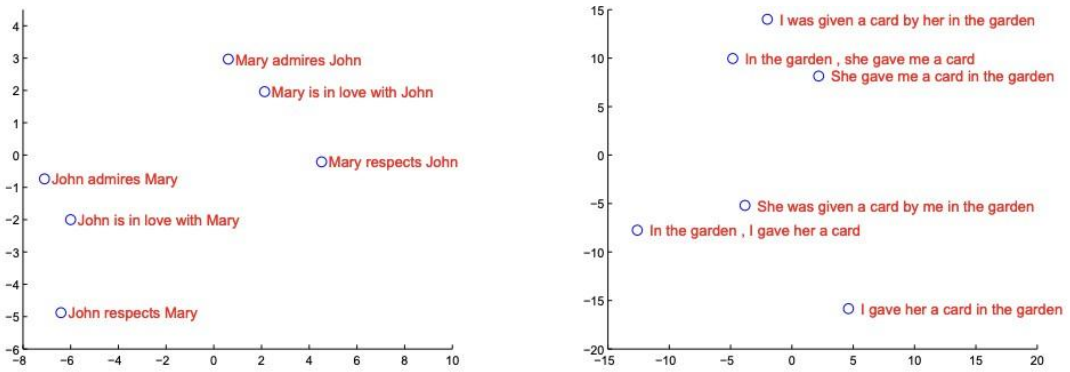

Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words model. Notice that both clusters have similar internal structure.

Anotablestrengthofourmodelisitscapabilitytoconvertasequenceofwordsintoafixed-size vector. Figure 2 illustrates several of these learned representations. The visualization reveals that the model captures the importance of word order, yet remainsrelatively unaffected by switching between active and passive voice. These two-dimensional views were generated using Principal Component Analysis (PCA).

Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

Table 3: A few examples of long translations produced by the LSTM alongside the ground truth translations. The reader can verify that the translations are sensible using Google translate.

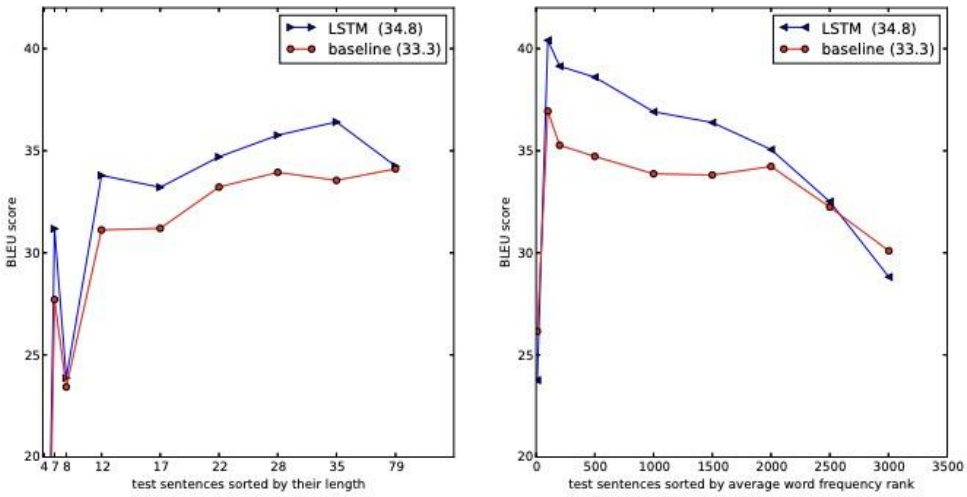

Figure 3: The left plot shows the performance of our system as a function of sentence length, where the x-axis corresponds to the test sentences sorted by their length and is marked by the actual sequence lengths. There is no degradation on sentences with less than 35 words, there is only a minor degradation on the longest sentences. The right plot shows the LSTM's performance on sentences with progressively more rare words, where the x-axis corresponds to the test sentences sorted by their "average word frequency rank".

4. RelatedWork

There has been substantial research on using neural networks for machine translation. Traditionally, the most straightforward and effective method has involved using RNN-based LanguageModels(RNNLMs)[23]orFeedforwardNeuralNetworkLanguageModels(NNLMs) [3]torescorethen-bestoutputsofastrongstatisticalmachinetranslation(SMT)system[22], which has consistently led to improvements in translation quality.

More recently, efforts have been made to integrate source language information directly into theseneuralmodels.Forinstance,Aulietal.[1]enhancedanNNLMbycombiningitwithatopic model of the source sentence, leading to better rescoring performance. Devlin et al. [8] also incorporated an NNLM into the MT decoder, using alignment data to identify and feed the most relevant source words into the neural model—an approach that yielded substantial performance gains.

Our approach closely aligns with the work of Kalchbrenner and Blunsom [18], who firstproposedconvertinginputsentencesintofixed-lengthvectorsandthendecodingthembackinto output sentences. However, they used convolutional neural networks (CNNs), which tend tolose word order information. Similarly, Cho et al. [5] employed an LSTM-style RNN to perform sentence-to-vector and vector-to-sentence transformations, although their main focus was on enhancing SMT systems.

Bahdanauetal.[2]alsopursueddirectneuraltranslation,introducinganattentionmechanismto improve performance on longer sentences—an issue observed by Cho et al. [5]. Pouget-Abadie et al. [26] tackled the same challenge by translating sentence segments to produce smoother outputs, somewhat resembling a phrase-based model. We believe that training on reversed source sentences could provide them with similar benefits.

End-to-endtrainingisalsoexploredbyHermannetal.[12],whousedfeedforwardnetworksto map inputs and outputs to nearby points in a vector space. However, their system cannot generate translations independently—it requires either a lookup from a pre-existing sentence database or the rescoring of candidate sentences.

5. Conclusion

In this study, we demonstrated that a large deep LSTM model—with a limited vocabulary and minimal assumptions about the structure of the task—can outperform a traditional SMT-based system, even one with an unlimited vocabulary, on a large-scale machine translation task. The strongperformanceofoursimpleLSTM-basedmethodsuggestsitcouldbeeffectiveforawide range of sequence learning tasks, given sufficient training data.

Interestingly,wefoundthatreversingthewordsinthesourcesentencessignificantlyimproved performance. This led us to conclude that encoding problems to increase the number of short-termdependenciescansimplifythelearningprocess.Notably,whilewestruggledtotraina standard RNN on unreversed input (as shown in Figure 1), we believe that reversing the sourcesentenceswouldmakesuchtrainingfeasible,thoughwedidnottestthisexperimentally.

WewerealsosurprisedbytheLSTM'scapabilitytoaccuratelytranslatelongsentences.Initially, weassumedthatthemodel'slimitedmemorywouldhinderitsperformanceonlonger inputs—especially since similar models have shown poor results on such tasks [5, 2, 26]. However,ourLSTM,whentrainedonthereverseddataset,handledlongsentenceseffectively.

Mostimportantly,weshowedthatastraightforward,relativelyunrefinedapproachcouldsurpass anSMTsystem.Thisindicatesthatwithfurtherdevelopment,evenbettertranslationresultsare achievable. Overall, our findings suggest that this method holds strong potential for tackling other complex sequence-to-sequence learning problems.

6. Acknowledgments

WearegratefultoSamyBengio,JeffDean,MatthieuDevin,GeoffreyHinton,NalKalchbrenner, Thang Luong, Wolfgang Macherey, Rajat Monga, Vincent Vanhoucke, Peng Xu, Wojciech Zaremba, and the Google Brain team for their valuable feedback and insightful discussions.

References

[1] M.Auli,M.Galley,C.Quirk,andG.Zweig.Jointlanguageandtranslationmodelingwith recurrent neural networks. In EMNLP, 2013.

[2] D.Bahdanau,K.Cho,andY.Bengio.Neuralmachinetranslationbyjointlylearningtoalign and translate. arXiv preprint arXiv:1409.0473, 2014.

[3] Y.Bengio,R.Ducharme,P.Vincent,andC.Jauvin.Aneuralprobabilisticlanguagemodel.In Journal of Machine Learning Research, pages 1137–1155, 2003.

[4] Y.Bengio,P.Simard,andP.Frasconi.Learninglong-termdependencieswithgradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994.

[5] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phraserepresentationsusingRNNencoder-decoderforstatisticalmachinetranslation.InArxiv preprint arXiv:1406.1078, 2014.

[6] D.Ciresan,U.Meier,andJ.Schmidhuber.Multi-columndeepneuralnetworksforimage classification. In CVPR, 2012.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and LanguageProcessing-SpecialIssueonDeepLearningforSpeechandLanguageProcessing, 2012.

[8] J.Devlin,R.Zbib,Z.Huang,T.Lamar,R.Schwartz,andJ.Makhoul.Fastandrobustneural network joint models for statistical machine translation. In ACL, 2014.

[9] NadirDurrani,BarryHaddow,PhilippKoehn,andKennethHeafield.Edinburgh's phrase-based machine translation systems for wmt-14. In WMT, 2014.

[10] A.Graves.Generatingsequenceswithrecurrentneuralnetworks.InArxivpreprint arXiv:1308.0850, 2013.

[11] A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification:labellingunsegmentedsequencedatawithrecurrentneuralnetworks.InICML, 2006.

[12] K.M.HermannandP.Blunsom.Multilingualdistributedrepresentationswithoutword alignment. In ICLR, 2014.

[13] G.Hinton,L.Deng,D.Yu,G.Dahl,A.Mohamed,N.Jaitly,A.Senior,V.Vanhoucke,P. Nguyen,T.Sainath,andB.Kingsbury.Deepneuralnetworksforacousticmodelinginspeech recognition. IEEE Signal Processing Magazine, 2012.

[14] S.Hochreiter.Untersuchungenzudynamischenneuronalennetzen.Master'sthesis,Institut fur Informatik, Technische Universitat, Munchen, 1991.

[15] S.Hochreiter,Y.Bengio,P.Frasconi,andJ.Schmidhuber.Gradientflowinrecurrentnets: the difficulty of learning long-term dependencies, 2001.

[16] S.HochreiterandJ.Schmidhuber.Longshort-termmemory.NeuralComputation,1997.

[17] S.HochreiterandJ.Schmidhuber.LSTMcansolvehardlongtimelagproblems.1997.

[18] N.KalchbrennerandP.Blunsom.Recurrentcontinuoustranslationmodels.InEMNLP, 2013.

[19] A.Krizhevsky,I.Sutskever,andG.E.Hinton.ImageNetclassificationwithdeep convolutional neural networks. In NIPS, 2012.

[20] Q.V.Le,M.A.Ranzato,R.Monga,M.Devin,K.Chen,G.S.Corrado,J.Dean,andA.Y.Ng. Building high-level features using large scale unsupervised learning. In ICML, 2012.

[21] Y.LeCun,L.Bottou,Y.Bengio,andP.Haffner.Gradient-basedlearningappliedto document recognition. Proceedings of the IEEE, 1998.

[22] T.Mikolov.StatisticalLanguageModelsbasedonNeuralNetworks.PhDthesis,Brno University of Technology, 2012.

[23] T.Mikolov,M.Karafi´at,L.Burget,J.Cernock`y,andS.Khudanpur.Recurrentneuralnetwork based language model. In INTERSPEECH, pages 1045–1048, 2010.

[24] K.Papineni,S.Roukos,T.Ward,andW.J.Zhu.BLEU:amethodforautomaticevaluation of machine translation. In

ACL, 2002.
[25] R.Pascanu,T.Mikolov,andY.Bengio.Onthedifficultyoftrainingrecurrentneural networks. arXiv preprint arXiv:1211.5063, 2012.
[26] J. Pouget-Abadie, D. Bahdanau, B. van Merrienboer, K. Cho, and Y. Bengio. Overcoming thecurseofsentencelengthforneuralmachinetranslationusingautomaticsegmentation.arXiv preprint arXiv:1409.1257, 2014.
[27] A.Razborov.Onsmalldepththresholdcircuits.InProc.3rdScandinavianWorkshopon Algorithm Theory, 1992.
[28] D.Rumelhart,G.E.Hinton,andR.J.Williams.Learningrepresentationsby back-propagating errors. Nature, 323(6088):533–536, 1986.
[29] H.Schwenk.Universitylemans.http://www-lium.univ-lemans.fr/˜schwenk/cslm_ joint_paper/, 2014. [Online; accessed 03-September-2014].
[30] M.Sundermeyer,R.Schluter,andH.Ney.LSTMneuralnetworksforlanguagemodeling.In INTERSPEECH, 2010.
[31] P.Werbos.Backpropagationthroughtime:whatitdoesandhowtodoit.Proceedingsof IEEE, 1990.