Journal of Cheminformatics

RESEARCH ARTICLE Open Access

# Neural machine translation of chemical nomenclature between English and Chinese

Tingjun Xu*, Weiming Chen, Junhong Zhou, Jingfang Dai, Yingyong Li and Yingli Zhao

**Abstract**

Machine translation of chemical nomenclature has considerable application prospect in chemical text data processing between languages. However, rule based machine translation tools have to face significant complication in rule sets building, especially in translation of chemical names between English and Chinese, which are the two most used languages of chemical nomenclature in the world. We applied two types of neural networks in the task of chemical nomenclature translation between English and Chinese, and made a comparison with an existing rule based machine translation tool. The result shows that deep learning based approaches have a great chance to precede rule based translation tools in machine translation of chemical nomenclature between English and Chinese.

## Introduction

Chemical names are primitive representations of chemicals, widely used by chemists in research articles, patents, data materials to describe chemical substances. Names accorded with chemical nomenclatures of IUPAC and CAS are exact expressions of molecular structures [1, 2], therefore those names can be used as identifiers of substances in chemical databases, and can be recognized by machine easily for converting names to conventional chemical structure representations [3, 4]. English and Chinese are the two most used languages of chemical nomenclature in the world, according to number of results found by Google for searching a chemical name in different languages [5]. However, the linguistic differences of English and Chinese chemical names have limited exchanges between users on both sides [6]. Therefore, machine translation of chemical nomenclature would be more applicable than manual translation in chemical data processing. For example, data sets of compound names would be more valuable when derived from chemical named entity recognition systems of Chinese text-mining materials, because bulk translation of Chinese chemical names into English by machine make it possible for the names to be converted into connection tables of chemical structures, owing to the fact that the vast majority of "name to structure" tools are only for English nomenclature [7–9].

Unfortunately, it still has a lot of work to be done in machine translation of chemical nomenclature beyond existing researches [10–12], especially in translation of chemical names between English and Chinese [13, 14]. There has significant complication in the task of analyzing Chinese chemical names, and make it difficult to set up a sophisticated machine translation rule set for conversion of various Chinese chemical names to or from English [13, 14]. For example, the Chinese chemical name of "ethyl acetate" is "乙酸乙酯", there is not only no word boundaries (spaces) in Chinese chemical names, but the order of words is also reversed, "ethyl" for "乙酯", "acetate" for "乙酸", and there is a special case of "ethyl" translated into "乙酯", because it would be "乙基" or "乙" in other names such as "ethyl alcohol" (乙醇). The word "ol" of hydroxyl in English organic compound names is often confused in Chinese names, it would be "醇" when the OH group bond with an aliphatic parent, for example "methanol" is translated into "甲醇", but it would be "酚" when the OH group bond with an aromatic ring, for example "benzene-1,2,4-triol" is translated into "

*Correspondence: xutingjun@sioc.ac.cn
Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 345 LingLing Road, Shanghai 200032, China

苯-1,2,4-三酚". Similar cases exist widely in translation of chemical nomenclature between English and Chinese.

Therefore, rule based machine translation of chemical nomenclature between English and Chinese requires specialized expertise, and need a formally trained chemist who is fluent in English and Chinese languages to build a perfect set of rules [8, 13, 14]. Machine Learning (ML) based approaches may be more applicable for the task of chemical nomenclature translation, because there is no need for building complex rule set, and in view of that there already have various applications in the field of neural machine translation [15–17]. We herein describe two Deep Learning (DL) based approaches of neural machine translation of chemical nomenclature between English and Chinese, and make a comparison with an existing rule based machine translation tool.

## Materials and methods

ML based approaches ideally require large data sets from which to learn, and quality of the data is critical. Obtaining such data set of chemical names is a significant challenge, as many are maintaining in one language (English or Chinese), and others are below-standard quality for ML use. Furthermore, translation of chemical nomenclature between English and Chinese is not inherent symmetry, that means an English chemical name have a translation of Chinese name which may be not suitable for back-translation. For example, "β-phenethylol" is translated into "β-苯乙醇" in Chinese, but back-translation of "β-phenylethyl alcohol" seems more appropriate. That is to say, the reversed version of training data for English to Chinese translation model may be not suitable for Chinese to English translation model.

Therefore, we built data sets on the basis of our chemical data analyzing system, the system has a corpus manually curated from scientific literatures and compiled by our data analysts. The corpus includes chemical names and their manual translations (English and Chinese), and the names are various from systematic compound names to trivial names. We extracted chemical names and their translations from the corpus, and made up two data sets, English names translated into Chinese names (En2Ch) batch and Chinese names translated into English names (Ch2En) batch. For names that have multiple translations, we chose the translations of chemical names that most used by our analysts in the data analyzing system. Eventually, we obtained the data sets of 30,394 records for En2Ch and 37,207 records for Ch2En both duplicate removed, and each data set is made up of source names and target names in two languages (Additional file 1). The data set has 100 unique characters in English name strings, and 2,056 unique characters in Chinese translated name strings.

We first applied a character-level sequence to sequence Convolutional Neural Network (CNN) based neural networks for translation of chemical nomenclature [18–20]. The neural networks conceptually consists of four elements: An encoder of three one-dimensional CNN layers encodes the input character sequence; A decoder of three one-dimensional CNN layers turns the target sequences into the same sequence but offset by one timestep in the future; Attention mechanism layers take outputs of the encoder and decoder; And a decoder of two one-dimensional CNN layers decodes the output character sequence, as shown in Fig. 1. The input chemical name strings are transformed into embedding sets of vectors. The number of vectors equals the number of unique characters in all input chemical names, and provided as an input to the encoder–decoder model with attention mechanism. The output strings are reversed from predicted sequences by re-embedding.

Recurrent Neural Networks (RNNs) and specifically Long Short Term Memory (LSTM), are also deep learning architectures well suited to the translation of variable-length sequences [15, 21, 22]. We then applied a LSTM based neural networks for translation of chemical nomenclature [15, 18, 23]. The LSTM based neural networks have an encoder of LSTM layers, the encoder turns input sequence to 2 state vectors, and a decoder of LSTM layers is trained to turn the target sequences into the same sequence but offset by one timestep in the future, and the decoder uses the state vectors from the encoder as initial state, this is a process called "teacher forcing", as shown in Fig. 2.

The neural networks were implemented in Python 3.7 [24] using Keras 2.3 [25] and TensorFlow backend [26]. The neural network models were trained on each data set of En2Ch and Ch2En. We split the data set for cross-validation at random, 80% for training set and 20% for validation set. The parameters of the neural networks were chosen according to the performances on the validation set: Batch size for training is 64, number of epochs is 100, latent dimensionality of the encoding and decoding space is 256 in CNN (Additional file 2) and LSTM (Additional file 3) based neural networks.

For the purpose of contrast analysis between DL based and rule based machine translation of chemical nomenclature, we used an existing rule based machine translation tool [27] on same data sets of the neural network models. Processes of the translation tool are conceptually consists of three steps: Disassembly, translation and reassembly. In a procedure of translation, a chemical name will be analyzed by the translation tool first, disassembled to word fragments, translated into target language, and then reassembled to a translated chemical name, all the

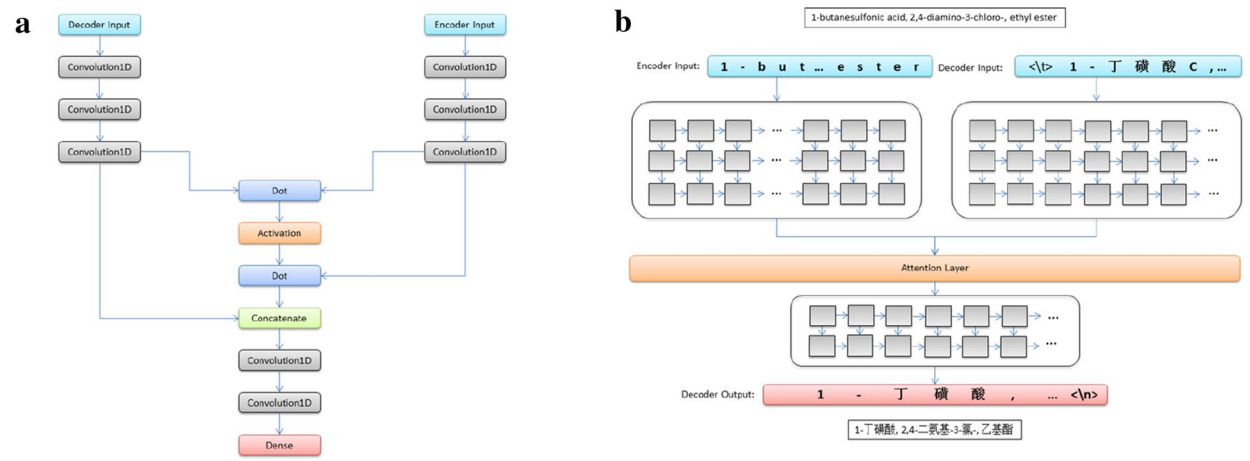


**Fig. 1** Architecture of the CNN based neural networks for machine translation of chemical nomenclature (**a**). Illustration of the CNN based neural networks for machine translation of chemical nomenclature in training mode (**b**)

processes are in compliance with a rule set, as shown in Fig. 3.

## Results and discussion

We toke the same testing data from the data set for DL and rule based machine translation of chemical nomenclature, 6,079 data records of English names translated into Chinese names (En2Ch), and 7,441 data records of Chinese names translated into English names (Ch2En). The results (Additional file 4) were analyzed in five fields as shown in Table 1, each field has values of performance at En2Ch and Ch2En. Success Rate: Percentage of translated names that each approaches successfully produced; String Matching Accuracy: Exactly string matching accuracy of translated names from each approaches and target names from the data sets; Data Matching Accuracy: Exactly string matching with data of translated names from the corpus which has multiple translations of one chemical name; Manual Spot Check: Manual spot check accuracy of 100 random names from the testing data sets for each result, and the checker was "blind" for which system generated the results; Running Time: Time of each approach running through the testing data sets in the same computing environment.

The two DL based approaches (CNN and LSTM) successfully produced translated names of all the testing data, owing to that the neural network models can always give an output unless there are characters absent from

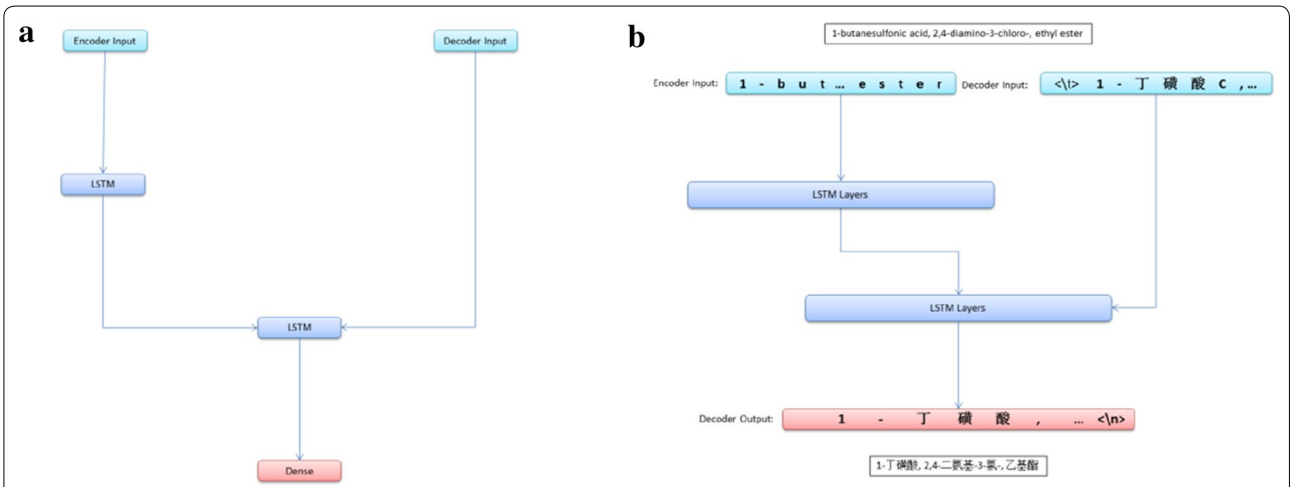


**Fig. 2** Architecture of the LSTM based neural networks for machine translation of chemical nomenclature (**a**). Illustration of the LSTM based neural networks for machine translation of chemical nomenclature in training mode (**b**)

the training data [28]. However, the rule based translation tool failed in producing translated names at a considerable proportion, especially on the Ch2En data set. For example, the translation of "羽扇豆醇棕榈酸酯", "异氧化前胡素", "原儿茶酸甲酯", etc. We believe failures in the processes of name disassembly and word fragments translation caused interrupts of the program, especially because no word boundaries (spaces) in these Chinese chemical names, leads to word segmentation errors, commonly happened in translation of long trivial names. Most trivial names of natural products are usually derived from the species names of their biological sources [29, 30], translation of these names may be irregular [6], some Chinese natural product names are even transliterations, and lead to failures of the rule based translation tool.

The performances of each approach at data matching with translated names from the corpus are all better than string matching with target names from the data set, for the reason that chemical names usually have more than one translation in the corpus. The two DL based approaches seem much more qualified for translation of chemical nomenclature, as observed from the result data, except the Ch2En batch produced by LSTM based neural networks. Although LSTM neural networks have fixed vanishing gradient problem of traditional RNNs [31], it seems still have deficiencies in learning long-term dependencies of Chinese characters, seeing that a lot of false results translated by the LSTM based neural networks appear at the end of chemical names, for example, "7-十八烯酸甲酯" (7-octadecenoic acid, methyl ester) was translated into "7-oleic acid, diethyl ester", "3-甲氧基苯乙酸" (3-methoxyphenylacetic acid) was translated into "3-methoxy cinnamal", etc.

**Table 1 The performances and comparisons between DL and rule based machine translation of chemical nomenclature**

| Field | CNN based | LSTM based | Rule based |
|---|---|---|---|
| Success Rate En2Ch | 100% | 100% | 75.97% |
| Success Rate Ch2En | 100% | 100% | 59.90% |
| String Matching Accuracy En2Ch | 82.92% | 89.64% | 39.81% |
| String Matching Accuracy Ch2En | 78.11% | 55.44% | 43.77% |
| Data Matching Accuracy En2Ch | 84.44% | 90.82% | 45.15% |
| Data Matching Accuracy Ch2En | 80.22% | 57.40% | 44.91% |
| Manual Spot Check En2Ch | 90.00% | 89.00% | 80.00% |
| Manual Spot Check Ch2En | 82.00% | 61.00% | 78.00% |
| Running Time En2Ch (s) | 1423 | 190 | 288 |
| Running Time Ch2En (s) | 1876 | 303 | 322 |

Manual spot check (Additional file 5) shows that performances presented by the two DL based approaches are close to string matching accuracy, but there is a large gap between performances presented by the rule based translation tool. We believe the main reason is multiple translations of one chemical name, but rule based translation tools are always constrained by

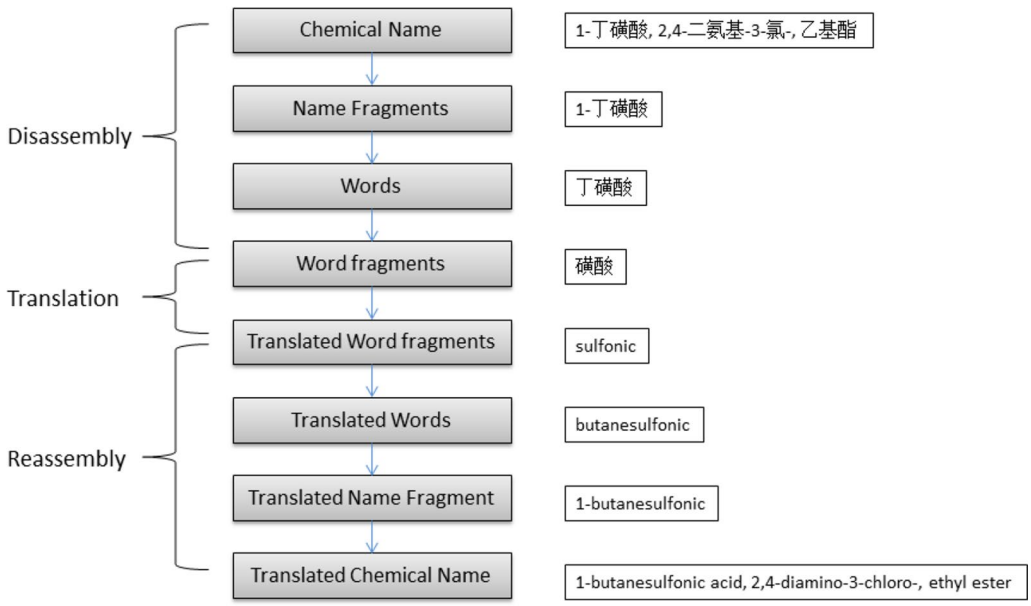


**Fig. 3** Illustration of the rule based machine translation tool of chemical nomenclature

**Table 2 The performances and comparisons of translating chemical names using different naming systems and having different length**

| Field | CNN based (%) | LSTM based (%) | Rule based (%) |
|---|---|---|---|
| IUPAC names | 92.00 | 62.00 | 80.00 |
| CAS names | 80.00 | 52.00 | 78.00 |
| Length not greater than 6 | 80.42 | 60.47 | 38.81 |
| Length greater than 6 | 74.38 | 47.31 | 49.33 |

relatively fixed rule sets, therefore the names produced by translation tools may not match with target names but they are also appropriate. For example, "p-toluene" (对甲苯) was translated into "p-甲苯", "ethenyl hexanoate" (己酸乙烯基酯) was translated into "己酸乙烯酯", etc.

The CNN based approach cost significantly more time than the other two, it could be the complexity of neural network architecture that decides computing cost of running through the testing data sets. We also found that some names translated by the two DL based approaches are better than target names in the data set. For example, the target name of "1-methoxy-4-methyl-benzene" is "对甲基苯甲醚" in the data set of En2Ch batch, although the target name is an alternative representation of this specific compound, it is better to be translated into "1-甲氧基-4-甲基-苯" literally. That is probably one of the reasons why the manual spot check accuracy of the two DL based approaches are all slightly better than matching accuracy.

We picked out 100 target names using IUPAC and CAS naming systems from Ch2En batch for manual check, to find out how different naming systems affect the translation. The result (Additional file 6) listed in Table 2 shows that the performances presented by rule based approach are close, but the performances presented by the two DL based approaches are better on IUPAC names than CAS names. We also evaluated the three approaches when translate chemical names of different length in Ch2En batch, we took the average length of 6 (Chinese characters) as a line of demarcation, as shown in Table 2. The rule based approach did not seem to be affected by the length of chemical names, but may be more applicable for regular names, because there are more systematic names in long chemical names. The neural networks we applied here are more applicable for translation of short sentences, seeing that the two DL based approaches made better performances on short chemical names.

## Conclusion

After comparison between two neural machine translation approaches and one existing rule based translation tool, we found that DL based approaches may precede rule based translation tools in general, but DL based approaches highly depend on quality and quantity of training data, and rule based tools highly depend on perfection of rule sets. We can not guarantee correctness of all chemical names translated by the two types of neural networks, but they had showed high accuracy. However, the rule based translation tool made much lower success rate, but it had considerable accuracy in manual spot check too. Further more, combination of the two types of neural networks (CNN and LSTM) may have greater capability [22, 32, 33], and would improve performance of LSTM based neural networks on Chinese chemical name translation.

One of the most common applications for Chinese chemical names translating into English is in scientific publications, because most of chemical journals in China request authors to provide abstracts both in Chinese and English, and some of editors require English chemical names in main text of manuscripts. Moreover, the rule based translation tool we applied here has been used as online services more than one million times in recent years [27]. Therefore, we believe the neural machine translation of chemical nomenclature we studied has considerable application prospect, and can provide new solutions not only for chemical data processing but for common use as well.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s13321-020-00457-0.

**Additional file 1:** Training data set.

**Additional file 2:** Python code for CNN model.

**Additional file 3:** Python code for LSTM model.

**Additional file 4:** Testing and result data set.

**Additional file 5:** Manual spot check and result data set.

**Additional file 6:** Manual check for translation of chemical names using different naming systems.

### Abbreviations

ML: Machine learning; DL: Deep learning; CAS: Chemical Abstracts Service; CNN: Convolutional neural network; RNN: Recurrent Neural Network; LSTM: Long Short Term Memory; IUPAC: International Union of Pure and Applied Chemistry; En2Ch: English chemical names translated into Chinese chemical

names; Ch2En: Chinese chemical names translated into English chemical names.

**Acknowledgements**
The authors would like to thank Prof. Weiming Chen for his guidance.

**Authors' contributions**
TX: Original idea. WC, JZ, JD, YL, and YZ: Supervised and participated to this work. All authors read and approved the final manuscript.

**Funding**
This work was supported by National Natural Science Foundation of China (21805303), CSDB (XXH135) and SGST (18DZ2294000).

**Availability of data and materials**
The data set for training and validation, the python code for generating the neural network models, and the result data set are included in the Additional files.

**Competing interests**
The authors declare that they have no competing interests.

Received: 5 August 2020 Accepted: 25 August 2020
Published online: 31 August 2020

**References**

1. McNaught A (2002) Chemical nomenclature and structure representation. Chem Int 24:12–14. https://doi.org/10.1515/ci.2002.24.2.12b
2. Chemical Abstracts Service (2007) Naming and indexing of chemical substances for chemical abstracts. Appendix IV of CA Index Guide
3. Ikutoshi, Matsuura (2005) Development of a system for translation of chemical name into 2D-structure. 28th symposium on chemical information and computer science, 29–32
4. Lowe DM, Corbett PT, Murray-Rust P, Glen RC (2011) Chemical name to structure: OPSIN, an open source solution. J Chem Inf Model 51:739–753. https://doi.org/10.1021/ci100384d
5. Google Inc (2020) Google. https://www.google.com/
6. China Chemical Society (2018) Nomenclature of organic compounds. SciencePress, Beijing
7. Vander Stouw GG, Elliott PM, Isenberg AC (1974) Automated conversion of chemical substance names to atom-bond connection tables. J Chem Doc 14:185–193. https://doi.org/10.1021/c160055a009
8. Cooke-Fox DI, Kirby GH, Rayner JD (1989) Computer translation of IUPAC systematic organic chemical nomenclature. 1. Introduction and background to a grammar-based approach. J Chem Inf Comput Sci 29:101–105. https://doi.org/10.1021/ci00062a009
9. Cooke-Fox DI, Kirby GH, Lord MR, Rayner JD (1990) Computer translation of IUPAC systematic organic chemical nomenclature. 4. Concise connection tables to structure diagrams. J Chem Inf Comput Sci 30:122–127. https://doi.org/10.1021/ci00066a004
10. Sayle R (2009) Foreign language translation of chemical nomenclature by computer. J Chem Inf Model 49:519–530. https://doi.org/10.1021/ci800243w
11. Summers L (1962) Machine translation of Russian organic chemical names into English by analysis and resynthesis of the component fragments. J Chem Doc 2:83–86. https://doi.org/10.1021/c160005a012
12. Garfield E (1961) Chemico-linguistics: computer translation of chemical nomenclature. Nature 192:192. https://doi.org/10.1038/192192a0
13. Chen B Chen W (2006) Study on machine translation of English compound name to Chinese. The 8th symposium on scientific database and information technology. Changsha, 2006.
14. Xu T, Chen W (2008) Study on machine translation of Chinese compound name to English. The 9th symposium on scientific database and information technology. Guilin, 2008.
15. Cho K, van Merrienboer B, Gulcehre G, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2017) Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), p 1724–1734. https://doi.org/10.3115/v1/D14-1179
16. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473. https://arxiv.org/abs/1409.0473
17. Luong T, Hieu P, Christopher DM (2015) Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, p 1412–1421. https://doi.org/10.18653/v1/D15-1166
18. Tanakitrungruang W (2017) Attention-based sequence-to-sequence in keras. https://wanasit.github.io/attention-based-sequence-to-sequence-in-keras.html
19. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. arXiv: 1409.3215. https://arxiv.org/abs/1409.3215
20. keras (2019) Sequence-to-sequence example in Keras (character-level). https://github.com/keras-team/keras/blob/master/examples/cnn_seq2seq.py
21. Sundermeyer M, Schlüter R, Ney H (2012) LSTM Neural Networks for Language Modeling. Interspeech. https://doi.org/10.1016/0165-6074(89)90269-X
22. Fooshee D, Mood A, Gutman E (2018) Deep learning for chemical reaction prediction. Mol Syst Des Eng 3:442–452. https://doi.org/10.1039/C7ME00107J
23. keras (2019) Trains a basic character-level sequence-to-sequence model. https://github.com/keras-team/keras/blob/master/examples/lstm_seq2seq.py
24. Python Software Foundation (2020) Python 3. https://www.python.org
25. Chollet F et al (2015) Keras. https://keras.io
26. Google Inc (2019) Tensorflow. https://github.com/tensorflow/tensorflow
27. Shanghai Institute of Organic Chemistry (2020) Machine translation tool for chemical nomenclature. https://www.organchem.csdb.cn/translate
28. Andrej K (2015) The unreasonable effectiveness of recurrent neural networks. https://karpathy.github.io/2015/05/21/rnn-effectiveness
29. Giles PM Jr (1999) Revised section F: natural products and related compound (IUPAC Recommendations 1999). Pure Appl Chem 71:587
30. Favre H, Powell W (2014) Nomenclature of organic chemistry: IUPAC recommendations and preferred names 2013. R Soc Chem. https://doi.org/10.1039/9781849733069
31. Christopher O (2015) Understanding LSTM Networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs
32. Sainath TN, Vinyals O, Senior A, Sak H (2015) Convolutional, long short-term memory, fully connected deep neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2015.7178838
33. Yoon K, Yacine J, David S, Alexander MR (2015) Character-aware neural language models. arXiv: 1508.06615. https://arxiv.org/abs/1508.06615

## Publisher's Note