

# SEQ2SEQ-VIS : A Visual Debugging Tool for Sequence-to-Sequence Models

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush

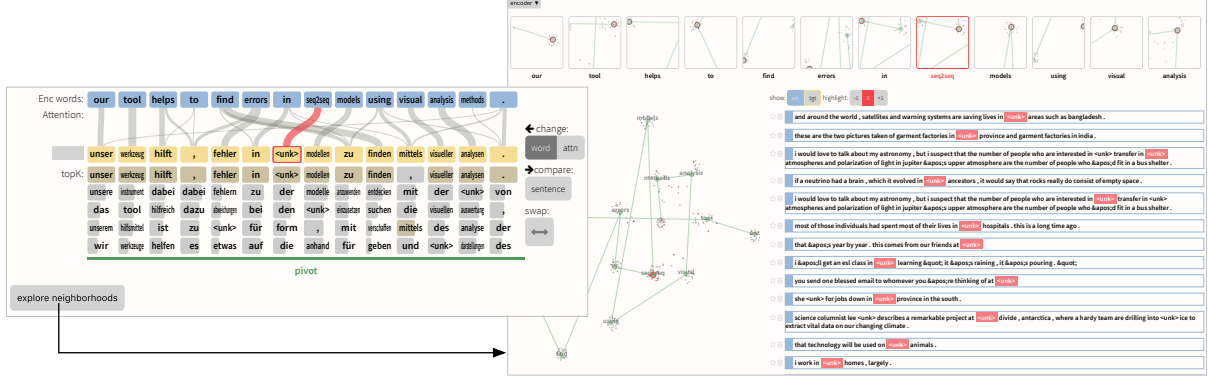


Fig. 1. Example of the Translation View (left) in Seq2Seq-VIS. The input sequence “our tool helps to find errors in seq2seq models using visual analysis methods.” is translated into the correct German sentence. The word “seq2seq” has correct attention between encoder and decoder (red highlight) but is not part of the language dictionary. When investigating the encoder neighborhoods (right), the user sees that “seq2seq” is close to other unknown words “(unk)”. The buttons enable user interactions for deeper analysis.

**Abstract**—Neural Sequence-to-Sequence models have proven to be accurate and robust for many sequence prediction tasks, and have become the standard approach for automatic translation of text. The models work in a five stage blackbox process that involves encoding a source sequence to a vector space and then decoding out to a new target sequence. This process is now standard, but like many deep learning methods remains quite difficult to understand or debug. In this work, we present a visual analysis tool that allows interaction with a trained sequence-to-sequence model through each stage of the translation process. The aim is to identify which patterns have been learned and to detect model errors. We demonstrate the utility of our tool through several real-world large-scale sequence-to-sequence use cases.

## 1 INTRODUCTION

Deep learning approaches based on neural networks have shown significant performance improvements on many artificial intelligence tasks. However, the complex structure of these networks often makes it difficult to provide explanations for their predictions. Attention-based *sequence-to-sequence models* (seq2seq) [3, 49], also known as encoder-decoder models, are representative of this trend. Seq2seq models have shown state-of-the-art performance in a broad range of applications such as machine translation, natural language generation, image captioning, and summarization. Recent results show that these models exhibit human-level performance in machine translation for certain important domains [10, 51].

Seq2seq models are powerful because they provide an effective supervised approach for processing and predicting sequences without requiring manual specification of the relationships between source and target sequences. Using a single model, these systems learn to do reordering, transformation, compression, or expansion of a source sequence to an output target sequence. These modifications are performed using a large internal state representation that encodes and then decodes the source sequence. With enough data, these models are seem to provide a general purpose mechanism for learning to predict sequences.

While the impact of seq2seq models has been clear, the added complexity and uncertainty of deep learning based models raises issues. These models often act as black-boxes during prediction, which makes

it difficult to track the source of any mistakes. The internal latent representations also make it difficult to analyze the model as it transforms the data far from the original sequence. While this property is shared with many other deep learning techniques, mistakes involving language are often very apparent to human readers. For instance, a widely publicized incident resulted from a seq2seq translation system mistakenly translating “good morning” into “attack them” leading to a wrongful arrest [12]. More common but worrying failures in seq2seq models might include a machine translation system greatly mistranslating a sentence, an image captioning system yielding an incorrect caption, or a speech recognition system producing an incorrect transcript.

Ideally, model developers wish to deploy systems that they fully understand and trust to produce correct results, but currently, this goal remains difficult with respect to deep learning models. In the meantime, we believe that the visual analytics community can contribute to this crucial challenge of better surfacing and visualizing the mistakes of seq2seq systems in a general and reproducible way.

We propose SEQ2SEQ-VIS, a visual analytics tool that satisfies this criteria by providing support for the following three goals:

- **Examine Model Decisions:** SEQ2SEQ-VIS allows users to understand, describe, and externalize model errors for all five stages of seq2seq models: encoder, decoder, attention, prediction, and beam search.
- **Connect Decisions to Samples:** SEQ2SEQ-VIS describes what a seq2seq model has learned based on transitions of latent states and their related neighbors from training data.
- **Test Alternative Decisions:** SEQ2SEQ-VIS provides sensible interactive methods that allow manipulation of model internals.

• H. Strobelt, A. Perer are with IBM Research and MIT-IBM Watson AI Lab.  
 • S. Gehrmann, A. Rush are with the Harvard NLP group.  
 • M. Behrisch and H. Pfister are with the Harvard Visual Computing group.

The full system is shown in Figure 1 (or larger in Fig. 7). It integrates visualizations for the components of the model (Fig 1 left) with internal representations from specific examples (Fig 1 middle) and nearest-neighbor lookups over a large offline corpus of precomputed examples (Fig 1 right).

Our paper begins with Sect. 2 by introducing important background, notations, and approaches to formalize our overall goal of seq2seq model debuggers. In Sect. 3 we present a guiding example that showcases how a typical model understanding- and debugging session looks like for an analyst. The subsequent “Goals and Task” Section 4 formalizes and highlights the context and enumerates necessary procedures for building seq2seq debuggers. Based on these guidelines, Sect. 5 and Sect. 6 introduce our visual and implementation design choices. Sect. 7 highlights in three real-world use cases how SEQ2SEQ-VIS guides the user through the visual analysis process. Sect. 8 puts our contributions in the context of related work for this research domain and Sect. 9 presents future work and reflections.

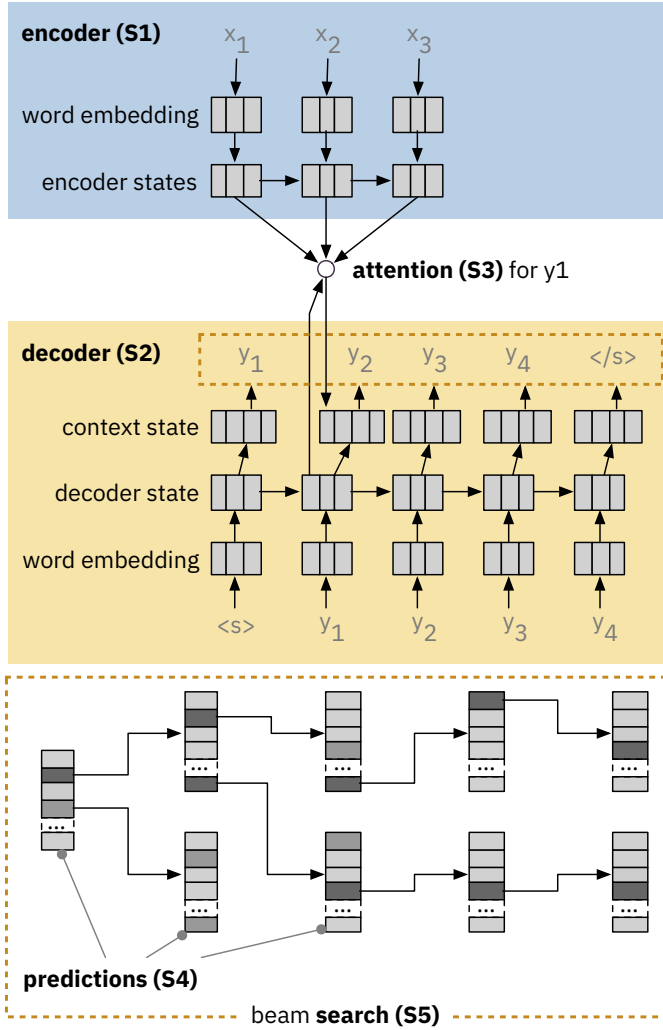


Fig. 2. Seq2seq models go through five stages for translating a source sequence to target sequence: (S1) encoding the source sequence into latent vectors, (S2) decoding to generate a target sentence, (S3) attend between encoder and decoder, (S4) predict word probabilities at each time step, and (S5) search for a best translation (beam search).

## 2 SEQUENCE-TO-SEQUENCE MODELS AND ATTENTION

We begin with a formal taxonomy of seq2seq models that will inform our visual analytics approach. Throughout this work, for brevity and clarity, we will consider the running example of automatic translation

from one language to another. We use the sequence notation  $\mathbf{x}_{1:S}$  to represent an  $S$  word sentence in our source language, and  $\mathbf{y}_{1:T}$  to represent a  $T$  word sentence in our target language. Seq2seq models learn to perform translation in a left-to-right manner, one word at a time, until the a stop token is produced. After observation of a stop token, the model stops, and has generated a full sentence.

We break down the translation process of seq2seq models into five-stages: (S1) *encode* the source sentence, (S2) *decode* the target words, (S3) *attend* to the encoded source, (S4) *predict* the next target word, and (S5) *search* for the best complete translation. Note that some systems use a slightly different order, but most adhere roughly to this setup. Fig. 2 provides a structural overview of these five stages.

**Encoder (S1):** Encoding uses a deep neural network to convert a sequence of source words  $\mathbf{x}_{1:S}$  into a sequence of hidden or latent vectors  $\mathbf{h}_{1:S}^{\text{enc}}$ . Each vector in the sequence  $\mathbf{h}_{1:S}^{\text{enc}}$  roughly represents the one word  $\mathbf{x}_s$  but also takes into account all of the surrounding words that may determine its contextual meaning. This encoding is typically done using a recurrent neural network (RNN) or a long short-term memory network (LSTM), however recently non-RNN-based methods such as CNN [8, 9] and Transformer [29, 50] have also been employed. Our approach supports all types of encoding methods.

**Decoder (S2):** The decoding process is analogous to encoding but instead takes the sequence of previous target words  $\mathbf{y}_{1:t}$  and converts them to a sequence of latent vectors  $\mathbf{h}_{1:t}^{\text{dec}}$ . Each vector represents the state of the sentence up to and including word  $\mathbf{y}_t$ . This provides a similar contextual representation as in the encoder, but is only based on previous words. Upon producing a new word in later steps, the previous prediction is fed back into the decoder.

**Attention (S3):** The attention component matches encoder hidden states and decoder hidden states. For each  $\mathbf{h}_t^{\text{dec}}$  we consider which encoder states  $\mathbf{h}_s^{\text{enc}}$  are relevant to the next prediction. In some similar language pairs, like French and Spanish, the words may align in order, e.g., the fourth French word matches the fourth Spanish word. However for languages such as English and Chinese, the matching might be quite far away. Instead of using absolute position, attention compares the hidden vectors to find which source position to translate. Attention forms a distribution based on the dot product between vectors  $\mathbf{h}_s^{\text{enc}}, \mathbf{h}_t^{\text{dec}}$ . We call this value  $a_{s,t}$ , and it roughly indicates how closely the source and target positions match.

**Prediction (S4):** The prediction step produces a multi-class distribution over the vocabulary of the target language, e.g., words that are more likely to come next have higher probability. This problem takes two factors into account: the current decoder state  $\mathbf{h}_t^{\text{dec}}$  and the encoder states weighted by attention. This last vector,  $\mathbf{c}_s = \sum_t a_{s,t} \mathbf{h}_s^{\text{enc}}$ , is known as the context vector and comes from S3. These two are combined to predict a distribution over the next word  $p(\mathbf{y}_{t+1} | \mathbf{x}_{1:S}, \mathbf{y}_{1:t})$ .

**Search (S5):** To actually produce a translation, these previous steps are combined into a search procedure. Seq2seq models are inherently local, since they predict the next word one step at a time. While one could simply take the highest probability word at each time step, it is possible that a lower probability word prediction may lead to a more probable complete translation. A common heuristic is to use *beam search* to pursue multiple possible *hypothesis* translations each time step. Beam search builds a tree consisting of the top  $K$ - hypothesis translations to extend. At each point, each of these are extended and the most likely  $K$  are preserved.

Each stage of the process is crucial for effective translation, and it is hard to separate them out. However, the model does preserve some separations of concerns. The decoder (S2) and encoder (S1) primarily work with each language individually, and manage the change in hidden representations over time. Attention (S3) provides a link between the two representations and connects them during training. Prediction (S4) combines the current decoder state with the information moving through the attention bottleneck. Finally, search (S5) combines these with a global score table. These five stages provide the foundation for our visual analytics system.

### 3 MOTIVATING CASE STUDY: DEBUGGING TRANSLATION

In order to make our contributions concrete, we begin with a motivating case study. Further case studies are discussed in Sect. 7. This case study involves a model trainer (see [48]) who is building a German-to-English translation model (our model is trained on the small-sized IWSLT '14 dataset [31]).

The user begins by seeing that a specific example was mistranslated in production. She finds the source sentence: *Die längsten Reisen fangen an, wenn es auf den Straßen dunkel wird.*<sup>1</sup> This sentence should have been translated to: *The longest journeys begin, when it gets dark in the streets.* However, she finds the model produces the mistranslation: *The longest journey begins, when it gets to the streets.* Fig. 5(E/D) shows the tokenized input sentence in blue and the corresponding translation of the model in yellow (on the top). The user immediately observes that the model does not translate the word *dunkel* into *dark*.

This mistake exemplifies several goals that motivated the development of Seq2Seq-Vis. The user would like to examine the system's decisions, connect to training examples, and test possible changes. As described in Sect. 2, these goals apply to all five model stages: encoder, decoder, attention, prediction, and search.

**Hypothesis: Encoder (S1) Error?** Seq2Seq-Vis lets the user examine similar encoder states for any example. Throughout, we will use the term *neighborhood* for the twenty closest states in vector space from training data. The tool displays the nearest neighbor sentences for a specific encoder state as red highlights in a list of training set examples. Fig. 3 shows that the nearest neighbors for *dunkel* match similar uses of the word. The majority seem to express variations of *dunkel*. The few exceptions, e.g., *db*, are artifacts that can motivate corrections of the training data or trigger further investigation. Overall, the encoder seems to perform well.

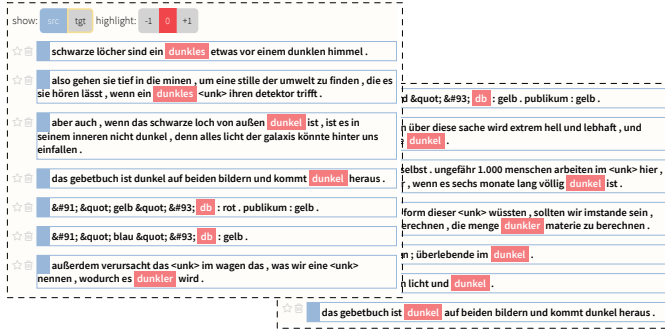


Fig. 3. Hypothesis: Encoder (S1) Error – nearest neighbors of encoder state for *dunkel*.

**Hypothesis: Decoder (S2) Error?** Similarly, the user can apply Seq2Seq-Vis to investigate the neighborhood of decoder states produced at times  $t$  and  $t + 1$  (Fig. 4). In addition to the neighbor list, it gives a projection view that depicts all decoder states for the current translation and all their neighbors in a 2D plane. Our analyst observes that the decoder states produced by *gets* and *streets* are in proximity and share neighbors. Since these states are indicative for the next word we can switch the highlight one text position to the right (+1) and observe that the decoder states at *gets* and *streets* support producing *dark*, *darker*, or *darkness*. Hence, the decoder state does not seem very likely as the cause of the error.

**Hypothesis: Attention (S3) Error?** Since both encoder and decoder are working, another possible issue is that the attention may not focus on the corresponding source token *dunkel*. The previous hypothesis testing revealed that well-supported positions for adding *dark* are after *gets* or *streets*. This matches human intuition, as we can imagine the following sentences being valid translations: *The longest travels begin when it gets dark in the streets.* or *The longest travels begin when it*

S2

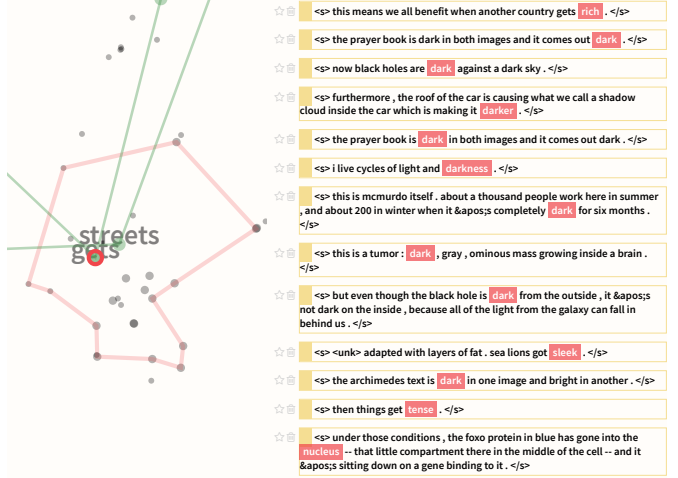


Fig. 4. Hypothesis: Decoder (S2) Error – nearest neighbors of decoder state for *gets* and *streets*, which are close in projection space.

*gets to the streets turning dark.* In Fig. 5(S3) our analyst can observe that the highlighted connection following *get* to the correct next word *dunkel* is very strong. The strong connection width indicates that the attention weight is very high with the right word. Hence, the user can assume that the attention is well set for predicting *dark* in this position. The hypothesis for error in S3 can be rejected with high probability.

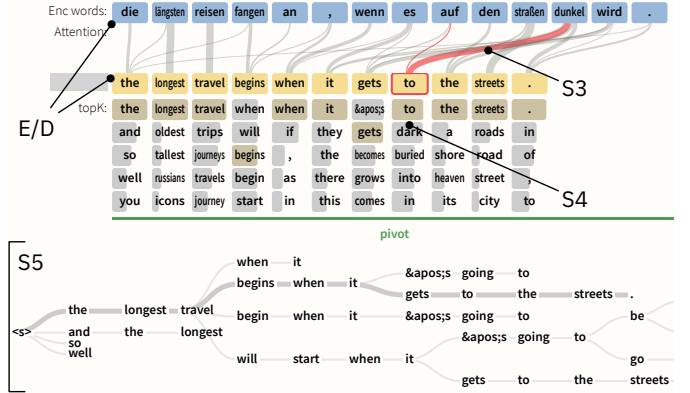


Fig. 5. Hypotheses: Attention (S3), Prediction (S4), or Beam Search (S5) Error – encoder words and decoder words (E/D), Attention (S3), top k predictions for each time step in decoder (S4), and beam search tree (S5)

**Hypothesis: Prediction (S4) Error?** The combination of decoder state and attention is used to calculate the probability of the next word. It may be that an error occurs in this decision, leading to a poor probability of the word *dark*. The tool shows this next step probability in Fig. 5(S4). Here, our analyst can see the model does make a mistake, assigning a higher probability to *to* than *dark*. However, both options are very close in score indicating that the model is quite uncertain and almost equally split between the two choices. These local mistakes should be automatically fixed by the beam search.

**Hypothesis: Search (S5) Error?** Having eliminated other possible issues, it becomes likely that the problem is a search error. The user can observe the beam search tree in Fig. 5(S5). For the current prediction, our analyst retrieves the beam of highest support and finds out that in none of the branches *dark* gets used. We can conclude that this might be the most likely error case, i.e., the analyst has identified a search error, where the beam search cuts off a better global option in favor of

<sup>1</sup>The closing quote of the book 'Kant' from German author Jörg Fauser, who is attributed as being a forerunner of German underground literature.



a worse local choice.

**Exploring Solutions.** A possible fix for this issue is a more exhaustive search procedure with larger beam size. SEQ2SEQ-VIS allows the user to evaluate a more specific fix to see if the translation would improve. When observing the k-best predictions for the position of *to* the user sees that *dark* and *to* are close together in probability (Fig. 5(S4)). Thus, our analyst can test what would happen if she forced the translation to use *dark* at this critical position. By clicking on *dark* she can produce this probe (shown in Fig. 6) and observe the correct translation. The user can now describe the most likely cause of error (search error) and a local fix to the problem (forced search to include dark). The analyst can now add this case to a list of well described bugs for the model and later consider a global fix.

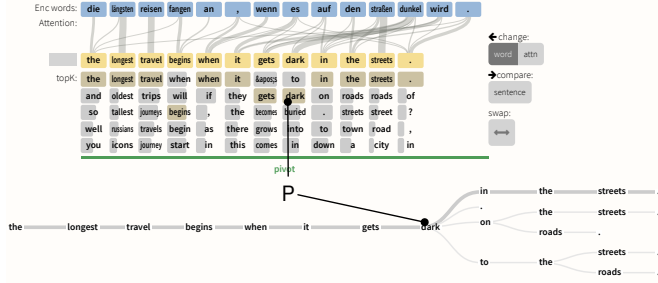


Fig. 6. Testing a fix – by clicking on the correct word *dark* in the predicted top-K, the beam search is forced on a specific path (P) which leads to the correct prediction.

## 4 GOALS AND TASKS

We now step back from a specific instance and consider a common deployment cycle for a deep learning model such as seq2seq. First, a model is trained on a new task with a possibly new data set, and then evaluated with a standard metric. The model performs well in aggregate and so the stakeholders decide to deploy it. However, for a certain subset of examples there exist non-trivial failures. For example, these may be noticed by users, or, in the case of translation, by post-editors who correct the output of the system. While the model itself is still useful, these examples might be significantly problematic as to cause alarm.

While these failures can occur in any system, this issue was much less problematic in previous generations of AI systems. For instance when using rule-based techniques, a user can explore the provenance of a decision through rules fired for a given output. If there is a mistake in the system, an analyst can 1) identify which rule misfired, 2) see which previous examples motivated the inclusion of the rule, and 3) experiment with alternative instances to confirm this behavior.

Ideally, a system could provide both functionalities: the high performance of deep learning with the ability to interactively spot issues and explore alternatives. However, the current architecture of most neural networks makes it more challenging to examine decisions of the model and locate problematic cases. Our work tackles the following challenges and domain goals for seq2seq models analogous to the three steps in rule-based systems:

**Goal G1 – Examine Model Decisions:** It is first important to examine the model’s decision chain in order to track down the error’s root cause. As mentioned in Sect. 2, seq2seq models make decision through several stages. While it has proven difficult to provide robust examination in general-purpose neural networks, there has been success for specific decision components. For example, the attention stage (S3) has proven specifically useful for inspection [3, 52]. Our first goal is to develop interactive visual interfaces that help users understand the model’s components, their relationships, and pinpoint sources of error.

**Goal G2 – Connect Decisions to Samples from Training Data:** Once a model makes a particular decision, a user should be able to trace what factors influenced this decision. While it is difficult to provide specific reasoning about the many factors that led to a decision

in a trained model, we hope to provide other means of analysis. In particular, we consider the approach of mapping example states to those from previous runs of the model. For instance, the training data defines the world view of a model and therefore influences its learned decisions [20]. The goal is to utilize (past) samples from training data as a proxy to better understand the decision made on the example in question.

**Goal G3 – Test Alternative Decisions:** Ultimately, though, the goal of the user is to improve the model’s performance and robustness. While the current state-of-the art for diagnosing and improving deep neural network models is still in an early stage [16, 17, 28, 47], our goal is to allow the user to test out specific interventions. We aim to let the user modify the model at different stages and observe the model’s changed behavior.

Our motivating case study (Sect. 3) followed these goals: First, the user defined five hypotheses for causes of error and tested them by examining the model’s decisions (G1). Some of these decisions (for S1, S2) could have been represented only as latent vectors, which she connected to representative neighbors from the training data (G2). Finally, by probing alternatives in the beam search (G3) she found a temporary alternative that helped her to formulate a better solution.

We use these goals to compile a set of visualization and interaction tasks for Seq2Seq-Vis. The mapping of these tasks to goals is indicated by square brackets:

**Task T1 - Create common visual encodings** of all five model stages to allow a user to examine the learned connections between these modules. [G1]

**Task T2 - Visualize state progression** of latent vector sequences over time to allow for high-level view of the learned representations. [G1]

**Task T3 - Explore generated latent vectors and their nearest neighbors** by querying a large database of training examples to facilitate error identification and training adjustment. [G2]

**Task T4 - Generate sensible alternative decisions for different stages of the model and compare them** to ease model exploration and compare possible corrections. [G1, G3]

**Task T5 - Create a general and coherent interface** to utilize a similar front-end for many sequence-to-sequence problems such as translation, summary, and generation. [G1, G2, G3]

In the following section, we will match these tasks and goals to design decisions for Seq2Seq-Vis.

## 5 DESIGN OF SEQ2SEQ-VIS

Seq2Seq-Vis is the result of an iterative design process and discussions between experts in machine learning and visualization. In regular meetings we evaluated a series of low-fidelity prototypes and tested them for usability. The design presented in this section combines the prevailing ideas into a comprehensive tool.

Seq2Seq-Vis is composed of two main views facilitating different modes of analysis: In the upper part, the *translation view* provides a visual encoding for each of the model stages and fosters understanding and comparison tasks. In the lower part, the *neighborhood view* enables deep analysis based on neighborhoods of training data. Fig. 7 shows the complete tool.

### 5.1 Translation View

In the translation view (Fig. 7a), each functional stage of the seq2seq model is mapped to a visual encoding (T1, T2, G1). We generalize and extend encodings from Olah & Carter [36] and Le et al. [23]. In Attention Vis (Fig. 7c), the encoder words are shown in blue, the decoder words in yellow, and the attention is shown through weighted bipartite connections. To reduce visual clutter the attention graph is pruned. For each decoder step all edges are excluded that fall into the lower quartile of the attention probability distribution.

Right below the yellow decoder words, the top-k predictions (S4 of model) for each time step are shown (Fig. 7d). Each possible prediction encodes information about its probability in the underlying bar chart, as well as an indication if it was chosen for the final output (yellow highlight).

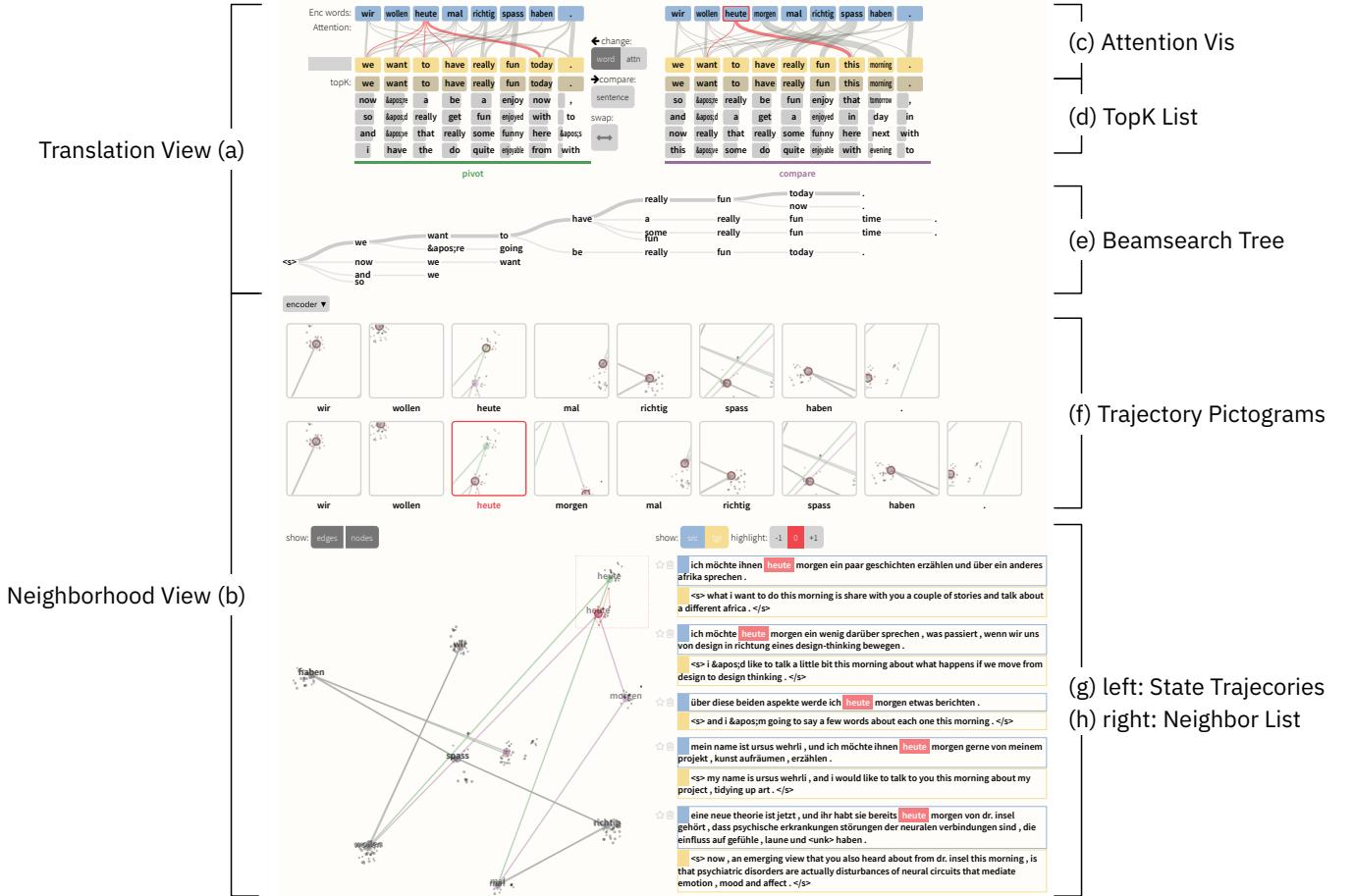


Fig. 7. Overview of Seq2Seq-Vis: The two main views (a) Translation View and (b) Neighborhood View facilitate different modes of analysis. Translation View provides (c) visualizations for attention, (d) the top k word predictions for each time step, and (e) the beam search tree. The Neighborhood View goes deeper into what the model has learned by providing (f,g) a projection of state trajectories and (h) a list of nearest neighbors for a specific model state.

In the bottom part of the translation view, a tree visualization shows the hypotheses from the beam search stage (Fig. 7e). The most probable hypothesis, which results in the final translation sentence, is highlighted. Several interactions can be triggered from the translation view, which will be explained in Sect. 5.4.

## 5.2 Neighborhood View

The neighborhood view (Fig. 7b) takes a novel approach to look at model decisions in the context of finding similar examples (T2, T3, G1, G2). As discussed in Sect. 2, seq2seq models produce high-dimensional vectors at each stage, e.g., encoder states, decoder states, or context states. It is difficult to interpret these vectors directly. However, we can estimate their meaning by looking at similar vectors from examples. To enable this comparison, we precompute the hidden states of a large set of exemplar sentences (we use 50k sentences from training). For each state produced by the model on a given example, Seq2Seq-Vis searches for nearest neighbors from this large subset of pre-recorded states. These nearest neighbors are input to the *state trajectories* (Fig. 7g) and to the *neighbor list* (Fig. 7h).

The state trajectories show the changing internal hidden state of the model with the goal of facilitating task T2. This view encodes the dynamics of a model as a continuous trajectory. First, the set for all states and their closest neighbors are projected using a non-linear algorithm, such as t-SNE [30], non-metric MDS [22], or a custom projection (see Sect. 7). This gives a 2D positioning for each vector. We use these positions to represent each encoder/decoder sequence as a trace connecting its vectors. See Fig. 7g for an example of a trace representing the encoder states for *wir wollen heute mal richtig spass haben*

*haben*.

In the projection, the nearest neighbors to each vector are shown as nearby dots. When hovering over a vector from the input, the related nearest neighbor counterparts are highlighted and a temporary red line connects them. For vectors with many connections (high centrality), we reduce visual clutter by computing a concave hull for all related neighbors and highlight the related dots within the hull. Furthermore, we set the radius of each neighbor dot to be dependent on how many original states refer to it. E.g., if three states from a decoder sequence have one common neighbor, the neighbor's radius is set to  $\sim 2.5$  (we use a  $r(x) = \sqrt{2x}$  mapping with  $x$  being number of common neighbors).

The state trajectories can be quite long. To ease understanding, we render a view showing states in their local neighborhood as a series of trajectory pictograms (Fig. 7f). Each little window is a cut-out from the projection view, derived from applying a regular grid on top of the projection plane. Each pictogram shows only the cut-out region in which the respective vector can be found.

Finally, clicking on any projected vector will trigger the neighbor list on the right side of the view. The neighbor list shows the actual sentences corresponding to the neighbor points, grounding these vectors in particular words and their contexts. Specifically, the neighbor list shows all the nearest neighbors for the selected point with the original sequence pair. The source or target position in the sequence that matches is highlighted in red. The user can facet the list by filtering only to show source (blue) or target (yellow) sequences. She can also offset the text highlight by  $-1$  or  $+1$  to see alignment for preceding or succeeding word positions (see Fig. 4).

### 5.3 Global Encodings and Comparison Mode

Seq2Seq-Vis uses shared visual encodings that apply across all views to create a coherent experience and simplify the tool’s learning curve for model architects and trainers (T5). First, a consistent color scheme allows the user to identify the stage and origin of data (encoder - blue, decoder - yellow, pivot - green, compare - violet). Furthermore, every visual element with round corners is clickable and leads to a specific action. Across all views, hovering highlights related entities in red.

Additionally, the tool has a global comparison mode (T4, G1, G3). As soon as we generate a comparison sample from one of several triggers (Sect. 5.4), all views switch to a mode that allows comparison between examples. Attention Vis, Trajectory Pictograms, and State Projector display a superpositioned layer of visual encodings labeled with a color (violet) different from the pivot color (green). To ease understanding, we disable further triggers in the comparison view. However, by providing an option to swap pivot and compare roles (arrow button), we allow a consistent exploration flow from one finding to another. The one exception is the Beam Search Tree, which is only shown for the pivot to save space.

### 5.4 Interacting With Examples

A major focus of Seq2Seq-Vis is interactive comparison between different sources and targets (T4, G1, G3). We consider two different modes of interactions to produce comparison samples or to modify the pivot: **model-focused** and **language-focused** changes. Model-focused interactions let the user (model architect) produce examples that the model believes are similar to the current pivot to test small, reasonable variations for the different model stages. Language-focused interactions enable the user (model trainer) to produce examples focussed on the language task and observe model behavior.

For the model-focused interactions, we utilize a different variant of neighborhoods. To replace a word with a slightly different, but interpretable substitute, we search for neighbors from the model’s word vectors. The user can trigger the substitution process by clicking on the word to be replaced. As a result, a word cloud projecting the closest words w.r.t. their vector embedding in a 2D plane is shown. A click on one of the words in the cloud replaces the original and triggers a new translation in comparison mode.

Another model-focussed interaction is to modify the model directly, for instance, by altering attention weights (S3 in model). For this step, the user can switch to attention modification and select a target word for which attention should be modified. By repeatedly clicking on encoder words, she gives more weights to these encoder words. Fig. 8 shows how the attention can be modified for an example. After hitting *apply attn*, the attention is applied for this position, overwriting the original attention distribution.



Fig. 8. To re-direct attention in Seq2Seq-Vis, the user first observes a split of attention between the input 8 and 9 for converting the last digits of a year in a date conversion model. She can (a) select attention mode, (b) select the decoder word, (c) click on the preferred encoder word, (d) apply the attention change, and (e) see the models reaction.

For language-focused interactions, the user can specify direct

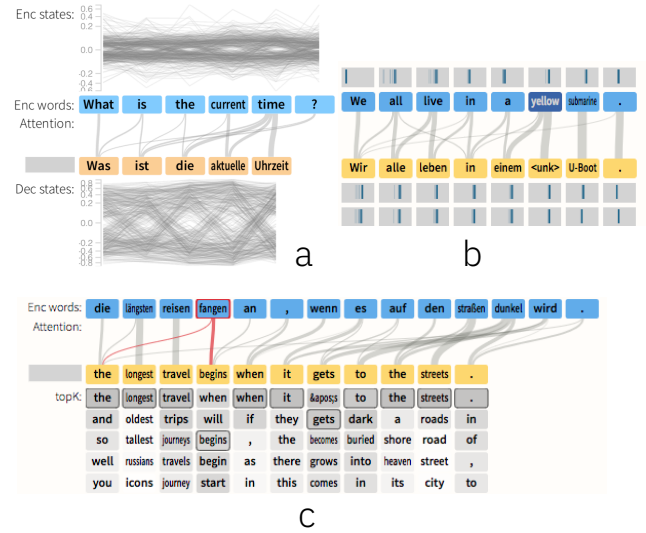


Fig. 9. Design variants for additional token information: (a) progression of hidden states, (b) density of neighborhoods, or (c) top-k predictions as heatmap.

changes to either the source or the target. The user can trigger the changes by using the *manual compare* button and enter a new source or a new target sentence. When the source is changed, a new full translation is triggered. If the target is changed, a prefix decode is triggered that constrains the search on a predefined path along the words entered and continues regular beam search beyond.

Alternatively, the user can select the word from the top-k predictions (Fig. 7d) that seems to be the best next word. By clicking on one of these words, a prefix decode is triggered as described above and shown in Fig. 6.

Initiating either of these interactions switches Seq2Seq-Vis into comparison mode. As a core analysis method, comparison allows to derive insights about model mechanics (model-focused probing) or how well the model solves the task (language-focused testing).

### 5.5 Design Iterations

We considered several different variants for both main views of the system. For the translation view, we considered incorporating more state information directly into the encoding. Fig. 9 shows iterations for adding per-word information around encoder and decoder. Similar to LSTMVis [48], the hidden state line charts show progression along encoder and decoder hidden state (Fig. 9a). Domain scientists rejected this as too noisy for the given domain goals. In a later iteration the visualization experts proposed to indicate the closeness of the nearest neighbors with a simple histogram-like encoding (Fig. 9b). This information did not help to formulate hypotheses. In addition, it did not reveal a lot of variance (see abundance of similar small gray boxes). The next design focused on incorporating language features rather than latent vectors. It showed for each time step of the decoder the top-k predicted words being produced as if there was only the top beam evaluated until then. Finally, we decided to use the stronger visual variable *length* to encode the probability values (Fig. 7d).

In the neighborhood view, the trajectory pictograms are a result of a series of visual iterations around combining the linear nature of sequences with keeping some spatial information describing vector proximity. We divide the state trajectory view into cutout regions forming a regular grid. Using a regular grid limits the variants of basic pictograms to a small and recognizable number. Alternative ideas were to center the cutout area around each state or to use the bounding box of the sequence state and all its neighbors as area. Both alternatives created highly variant multiples that introduced visual noise.



## 6 IMPLEMENTATION

Seq2Seq-Vis allows for querying and interaction with a live system. To facilitate this, it uses tight integration of a seq2seq model with the visual client. We based the interface between both parts on a REST API, and we used OpenNMT [18] for the underlying model framework. We extended the core OpenNMT-py distribution to allow easy access to latent vectors, the search beams, and the attention values. Furthermore, we added non-trivial model-diagnostic modifications for translation requests to allow prefix decoding and to apply user-specific attention. We plan to distribute Seq2Seq-Vis as the default visualization mode for OpenNMT.

To allow fast nearest neighbor searches, custom Python scripts extract the hidden state and context values from the model for points in a large subset of the training data. These states are saved in HDF5 files and indexed utilizing the Faiss [14] library to allow fast lookups for closest dot products between vectors. For TSNE and MDS projections we use the SciKit Learn package [40] for Python.

The model framework and the index work within a Python Flask server to deliver content via a REST interface to the client. The client is written in Typescript. Most visualization components are using the d3js library. Source code, a demo instance, and a descriptive webpage will be available at <http://seq2seq-vis.io>.

## 7 USE CASES

We demonstrate the application of Seq2Seq-Vis and how it helps to generate insights using examples from a toy date conversion problem, abstractive summarization, and machine translation (Sect. 3).

**Date Conversion.** Seq2seq models can be difficult to build and debug even for simple problems. A common test case used to check whether a model is implemented correctly is to learn a complex but well-specified deterministic task. Here we consider the use case of converting various date formats to the unified format YEAR-MONTH-DAY. For example, the source *March 25, 2000* should be converted to the target *2000-03-25*. While this problem is much simpler than language translation, it tests the different components of the system. Specifically, the encoder (S1) must learn to identify different months, the attention (S3) must learn to reorder between the source and the target, and the decoder (S2) must express the source word in a numeric format.

SEQ2SEQ-VIS provides tools for examining these different stages of the model. Figure 10 shows an example use case, where the user, following Goal 3, employs a comparison between two different translations, one starting with *March* and the other with *May*. These two translations are nearly identical, except one yields the month 3 and the other 5. Following Goal 1, the user might want to examine the models decisions. The upper translation view provides a way to compare between the attention on the two inputs. The red highlighted connections indicate that the first sentence attention focuses on *rc* whereas the second focuses on *y*. These characters are used by the model to distinguish the two months since it cannot use *Ma*. The user can also observe how the encoder learns to use these letters. The trajectory view compares the encoder states of sentence 1 and sentence 2. Here we use a custom projection, where the y-axis is the relative position of a word in a sentence and the x-axis is a 1-d projection of the vector. This reveals that the two trajectories are similar before and after these characters, but diverge significantly around *r* and *c*. Finally, following Goal 2, the user can connect these decisions back to the training data. On the right, she can see the nearest neighbors around the letter *a* in *May* (highlighted in red). Interestingly, the set of nearest neighbors is almost equally split between examples of *May* and *March*, indicating that at this stage of decoding the model is preserving its uncertainty between the two months.

**Abstractive Summarization.** For our second use case we apply the tool to a summarization problem. Recently, researchers have developed methods for *abstractive* text summarization that learn how to produce a shorter summarized version of a text passage. Seq2seq models are commonly used in this framework [33, 39, 44, 45]. In abstractive summarization, the target passage may not contain the same wording as the original. Instead, the model learns to paraphrase and alter the wording in the process of summarization.

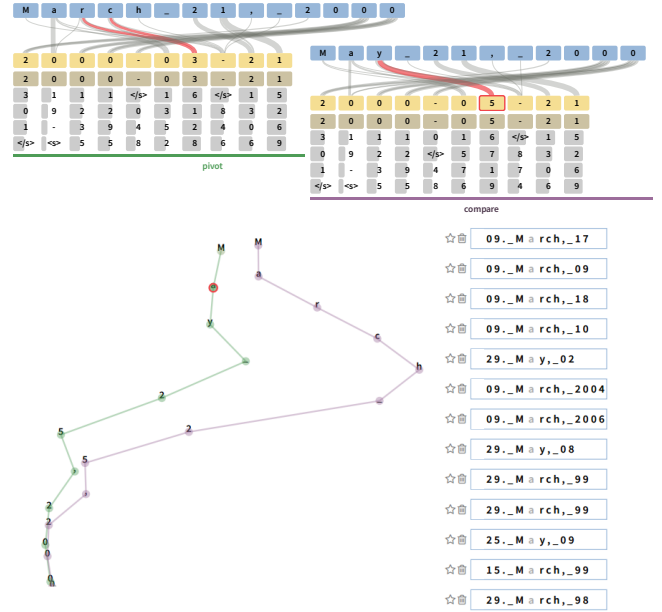


Fig. 10. Comparing translations for a date conversion model. The input sequences *March 21, 2000* and *May 21, 2000* are only different by some letters. The attention (top) for predicting the correct months 3 and 5 is focussed on this difference (*y* vs. *rc*). The trajectory view (bottom left) shows this difference along the progression of encoder states. The neighborhood list (bottom right) indicates that after input of *Ma* the model is still undecided.

Studying how paraphrasing happens in seq2seq systems is a core research question in this area. Rush et al. [44] describe a system using the Gigaword data set (3.8M sentences). They study the example source sentence *russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism* to produce a summary *russia calls for joint front against terrorism*. Here *russia* summarizes the phrase *russian defense minister ivanov* and *against* paraphrase for *combating*.

To replicate this use case we consider a user analyzing this sentence. In particular, he is interested in understanding how the model selects the length and the level of abstraction. He can analyze this in the context of Goal 3, testing alternatives predictions of the model, in particular targeting Stage 4. As discussed in Sect 5, SEQ2SEQ-VIS shows the top-K predictions at each time step. When the user clicks on a prediction, the system will produce a sentence that incorporates this prediction. Each choice is “locked” so that further alterations can be made.

Fig. 11 shows the source input to this model. We can see four different summarizations that the model produces based on different word choices. Interestingly, specific local choices do have a significant impact on length, ranging from five to thirteen words. Furthermore, the model reacts to word choices by altering its path. Switching from *for* to *on* leads the decoder to insert an additional phrase *on world leaders* to maintain grammaticality. While the model outputs the top choice, all other choices have relatively high probability. This observation has motivated research into adding constraints to the prediction at each time step. Consequently, we have added methods for constraining length and prediction into the underlying seq2seq system to produce different outputs.

**Machine Translation.** Finally, we consider a more in-depth use case of a real-world machine translation system using a complete model trained on WMT ’14 (3.96M examples) to translate from German to English. This use case considers a holistic view of how an expert might go about understanding the decisions of the system.

Figure 12 shows an example source input and its translation. Here the user has input a source sentence, translated it, and activated the neighbor view to consider the decoder states. She is interested in better

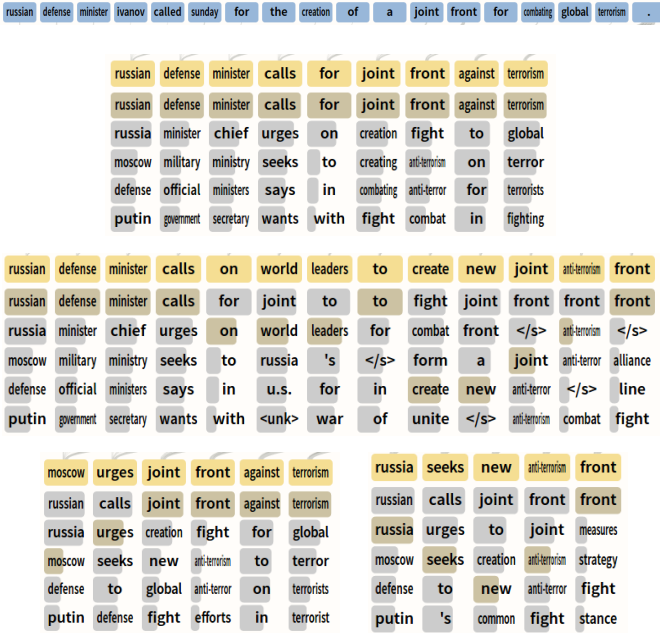


Fig. 11. Use case of abstractive summarization. The input sentence *russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism* can be summarized in different ways. The yellow boxes indicate alternative translations for different prefix decode settings. Top: the unconstrained abstraction; middle: changing prediction from *for* to *on* leads to automatic insertion of *on world leaders* to stay grammatically correct; bottom left: changing the first word from *russian* to *moscow* or *russia* compresses the sentence even more while retaining its meaning.

understanding each stage of the model at this point. This sentence is interesting as there is significant reordering that must occur to translate from the original German to English. For instance, the subject *he* is at the beginning of the clause, but must interact with the verb *gesprochen* at the end of the German sentence.

We consider Goals 1 and 2 applied to this example, with the intent of analyzing the encoder, decoder, attention, and prediction (S1-S4). First we look at the attention. Normally, this stage focuses on the word it is translating (*er*), but researchers have noted that neural models often look ahead to the next word in this process [19]. We can see branches going from *he* to potential next steps (e.g., *von* or *gesprochen*). We can further view this process in the decoder trajectory shown below, where *he* and *spoke* are placed near each other in the path. Hovering over the vector *he* highlights it globally in the tool. Furthermore, if we click on *he*, we can link this state to other examples in our data (Goal 2). On the right we can see these related examples, with the next word (+1) highlighted. We find that the decoder is representing not just the information for the current word, but also anticipating the translation of the verb *sprechen* in various forms.

In this case we are seeing the model behaving correctly to produce a good translation. However, the tool can also be useful when there are issues with the system. One common issue in under-trained or under-parameterized seq2seq models is to repeatedly generate the same phrase. Figure 13 shows an example of this happening. The model repeats the phrase *in Stuttgart in Stuttgart*. We can easily see in the pictogram view that the decoder model has produced a loop, ending up in nearly the same position even after seeing the next word. As a short-term fix, the tool's prefix decoding can get around this issue. It remains an interesting research question to prevent this type of cycle from occurring in general.

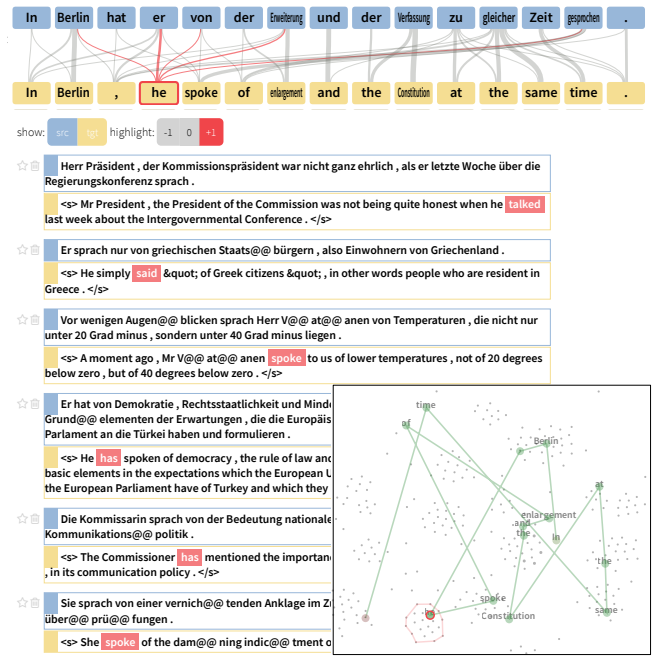


Fig. 12. Use case language translation using WMT'14 data. The attention graph (top) shows how attention for the target word *he* is not only focused on the decoder counterpart *er* but also on the following words, even to the far away verb *gesprochen* (spoke). The state trajectory (bottom left) for the decoder states reveals how close *he* and *spoke* are. The neighborhood list indicates that the model sets the stage for predicting *spoke* as next word.

## 8 RELATED WORK

Various methods [4, 34] have been proposed to generate explanations for deep learning model predictions. Understanding them still remains a difficult task. To better address the specific issues of our users, we narrow the target audience for our proposed tool. Following the classifications by Strobelt et al. [48] and Hohman et al. [13], our tool aims at model developers who have at least a conceptual understanding of how the model works. This is opposed to end users, who are agnostic to the technique used to arrive at a specific result. Following Hohman et al., analysis itself can broadly be divided into *global model analysis* and *instance-based analysis*. In global model analysis, the most commonly seen methods are visualizations of the internal structure of trained deep learning models. Instance-based analysis may be coupled with interactive experimentation with the goal of understanding a particular prediction using the local information around only one input [38].

**Global Model Analysis** Most recent work focuses on visualizing hidden representations of convolutional neural networks (CNNs) [24] for computer vision applications. Techniques for visualizing CNNs include showing neural activity in the convolutional layers as overlay over the image [7, 23] and directly showing the images that maximize the activity [46]. Zeiler and Fergus [54] use deconvolutional networks to explore the layers of a CNN. This approach is widely used to generate explanations of models, for example, by Yosinski et al. [53].

A similar line of work has focused on visualizing recurrent neural networks (RNNs) and other sequence models. Preliminary work by Karpathy et al. [17] uses static visualizations to understand hidden states in language models. They demonstrate that selected cells can model clear events such as open parentheses and the start of URLs. Strobelt et al. [48] introduce LSTMVis, an interactive tool that allows users to understand activation patterns of combinations of hidden states. LSTMVis shows the neighborhood of activations within the training data as an approach of making sense of the complex interactions in a context.

Similar to our approach, Kahng et al. [16] propose using the model



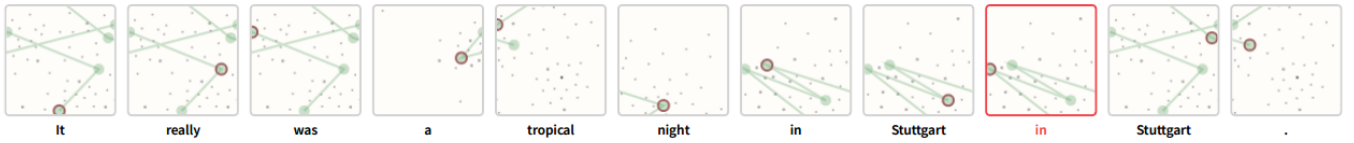


Fig. 13. An under-trained English-German model. Repetition repetition is a commonly observed phenomenon in under-trained or under-parametrized models. Here the trajectory pictograms show that for the repetition *in Stuttgart in Stuttgart* the decoder states alternate in the same region before being able to break apart.

structure as the entry point into the analysis. In their approach, they try to understand connections between misclassified examples and hidden states of parts of the network by showing activation pattern differences between correct and false examples from the training data. Ming et al. [32] propose RNNVis, a tool that uses word clouds instead of full contexts or sentences to show typical words that appear for activation patterns. Our approach to show embeddings of a whole phrase is similar to that of Johnson et al. [15]. They use three-dimensional tSNE in order to visualize progressions of context vectors. Novel in our approach are different types of progressions as well as the connection and embedding with neighborhoods.

An alternative to visualizing what a model has learned is visualizing how it is learning. RNNbow by Cashman et al. [5] shows the gradient flow during backpropagation training in RNNs to visualize how the network is learning.

**Instance-Based Analysis** Instance-based analysis is commonly used to understand local decision boundaries and relevant features for a particular input. For example, Olah et al. [37] extend methods that compute activations for image classification to build an interactive system that assesses specific images. They show that not only the learned filters of a CNN matter, but also their magnitudes. The same type of analysis can be used to answer counter-factual “what if” questions to understand the robustness of a model to perturbations. Nguyen et al. [35] show that small perturbations to inputs of an image classifier can drastically change the output. Interactive visualization tools such as Picasso [11] can manipulate and occlude parts of an image as input to an image classifier. Krause et al. [21] use partial dependence diagnostics to explain how features affect the global predictions, while users can interactively tweak feature values and see how the prediction responds to instances of interest.

There is an intrinsic difficulty in perturbing inputs of models that operate on text. While adding noise to an image can be achieved by manipulating the continuous pixel values, noise for categorical text is less well defined. However, there is a rich literature for methods that compute relevant inputs for specific predictions, for example by computing local decision boundaries or using gradient-based saliency [1, 27, 41, 42, 55]. Most of these methods focus on classification problems in which only one output exists. Ruckle et al. [43] address this issue and extend saliency methods to work with multiple outputs in a question-answering system. As an alternative to saliency-methods, Ding et al. [6] use a layer-wise relevance propagation technique [2] to understand relevance of input with regard to an output in sequence-to-sentence models. Yet another approach to understand predictions within text-based models is to find the minimum input that still yields the same prediction [26, 28]. None of the previous methods use our approach of using nearest neighbors of word embeddings to compare small perturbations of RNNs.

One commonality among all these approaches is that they treat the model as a black box that generates a prediction. In contrast, we are assuming that our users have an understanding of the different parts of a sequence-to-sequence model. Therefore, we can use more in-depth analysis, such as interactive manipulations of input, output, and attention. Our beam search and attention manipulations follow the approach by Lee et al. [25] who show a basic prototype to manipulate these parts of a model.

## 9 CONCLUSIONS AND FUTURE WORK

Seq2Seq-Vis is a tool to facilitate deep exploration of all stages of a seq2seq model. We apply our set of goals, inspired by rule-based systems, to deep learning models that are traditionally difficult to interpret. To our knowledge, our tool is the first of its kind to combine insights about model mechanics (translation view) with insights about model semantics (neighborhood view).

There are many avenues for future work on the algorithmic and visualization side. Improving the projection techniques to better respect the linear order of sequences would be helpful. The tool could be extended to different sequence types, including audio, images, and video. These different data types require an expansion of visual encoding for input and output as well as model modification methods. A larger deployment as an open source tool will generate more insights about model dynamics from users. We hope this will shed light on novel methods to fix models without retraining them entirely.

## REFERENCES

- [1] D. Alvarez-Melis and T. S. Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 412–421, 2017.
- [2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7):e0130140, 2015.
- [3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [4] O. Biran and C. Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, p. 8, 2017.
- [5] D. Cashman, G. Patterson, A. Mosca, and R. Chang. Rnnbow: Visualizing learning via backpropagation gradients in recurrent neural networks. In *Workshop on Visual Analytics for Deep Learning (VADL)*, 2017.
- [6] Y. Ding, Y. Liu, H. Luan, and M. Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1150–1159, 2017.
- [7] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical report, University of Montreal, 2009.
- [8] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin. A Convolutional Encoder Model for Neural Machine Translation. *ArXiv e-prints*, Nov. 2016.
- [9] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional Sequence to Sequence Learning. *ArXiv e-prints*, May 2017.
- [10] H. Hassan Awadalla, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou. Achieving human parity on automatic chinese to english news translation. March 2018.
- [11] R. Henderson and R. Rothe. Picasso: A modular framework for visualizing the learning process of neural network image classifiers. *Journal of Open Research Software*, 5(1), 2017.
- [12] A. Hern. Facebook translates ‘good morning’ into ‘attack them’, leading to arrest. *The Guardian*, Oct 2017.

- [13] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *arXiv preprint arXiv:1801.06889*, 2018.
- [14] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.
- [15] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google’s multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.
- [16] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau. Activis: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, 24(1):88–97, 2018.
- [17] A. Karpathy, J. Johnson, and F.-F. Li. Visualizing and understanding recurrent networks. *ICLR Workshops*, 2015.
- [18] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.
- [19] P. Koehn and R. Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.
- [20] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.
- [21] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5686–5697. ACM, 2016.
- [22] J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.
- [23] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. C. 0010, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. *ICML*, 2012.
- [24] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [25] J. Lee, J.-H. Shin, and J.-S. Kim. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 121–126, 2017.
- [26] T. Lei, R. Barzilay, and T. S. Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 107–117, 2016.
- [27] J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and Understanding Neural Models in NLP. In *NAACL*, pp. 1–10. Association for Computational Linguistics, San Diego, California, jun 2016.
- [28] J. Li, W. Monroe, and D. Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.
- [29] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.
- [30] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [31] C. Mauro, G. Christian, and F. Marcello. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pp. 261–268, 2012.
- [32] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu. Understanding hidden memories of recurrent neural networks. *arXiv preprint arXiv:1710.10777*, 2017.
- [33] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.
- [34] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.
- [35] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.
- [36] C. Olah and S. Carter. Attention and augmented recurrent neural networks. *Distill*, 2016. doi: 10.23915/distill.00001
- [37] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.
- [38] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist. Conceptvector: text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1):361–370, 2018.
- [39] R. Paulus, C. Xiong, and R. Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.
- [40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [41] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.
- [42] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.
- [43] A. Rücklé and I. Gurevych. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. *Proceedings of ACL 2017, System Demonstrations*, pp. 19–24, 2017.
- [44] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.
- [45] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.
- [46] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [47] D. Smilkov, S. Carter, D. Sculley, F. B. Viégas, and M. Wattenberg. Direct-manipulation visualization of deep networks. *arXiv preprint arXiv:1708.03788*, 2017.
- [48] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2018.
- [49] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.
- [50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.
- [51] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.
- [53] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.
- [54] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision—ECCV*, vol. 8689, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1\_53
- [55] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. *ICML*, 2017.