MOVING

A Case Study of Closed-Domain Response Suggestion with Limited Training Data

Lukas Galke[1], **Gunnar Gerstenkorn[2]** and Ansgar Scherp[3]
ZBW – Leibniz Information Centre for Economics[1 2 3]
Kiel University[1], Potsdam University[2], Sterling University[3]

September 4th, 2018, 15th International Workshop on Technologies for Information Retrieval,
September 3rd - 6th, 29th International Conference on Database and Expert Systems Applications,
Regensburg, Germany.

The Response Suggestion Setting

- **Problem:**
  - The Leibniz Information Center has a chat assistant for searching literature
  - The staff and domain experts receive increasingly more requests
  - Many of the questions are repeating
- **Solution:**
  - Suggest appropriate responses for a given request

Example Requests

- **Patron Request:**
  - "How can i buy an article."
- **Library Response:**
  - "Hello and welcome to the EconDesk chat."
  - "Let me take a look at your question. One moment."
  - "Which article do you mean ?"

Chances and Limitations

- **Chances for response suggestion:**
  - Closed domain
  - Looking for a full answer, not necessarily Natural Language Understanding
- **Limitations for response Suggestion:**
  - Very little data
  - Non-labeled, non-enhanced data

Suggestion Models

1. Retrieval

- Baseline: TF-IDF variants
- KNN
- Word Centroid Distances

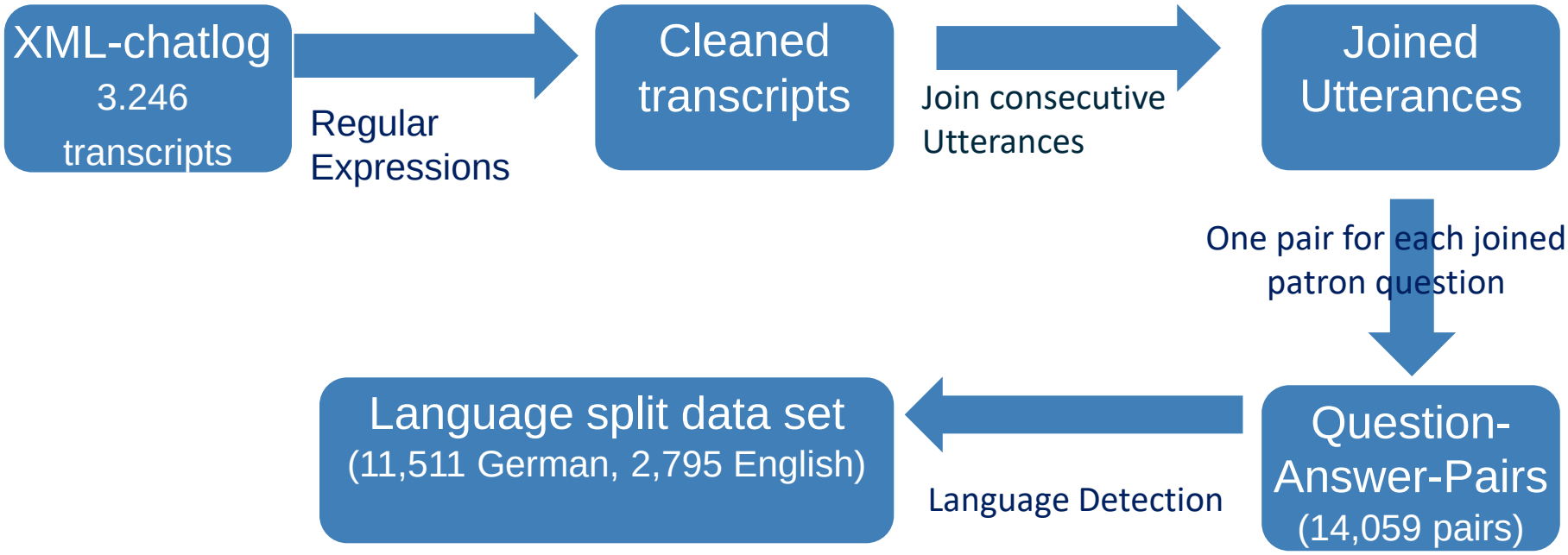
2. Representation Learning

- Feedforward NN to learn scoring function for good responses

3.Conditioned Generation

- Sequence to Sequence: neural word by word generation from input

Preprocessing Pipeline



Joining Consecutive Utterances

<PatronIncident>...

<LibraryIncident> ...

<PatronIncident> Am besten wäre eine Tabellarische Übersicht der Organisationen, ähnlich der John Hopkins Studie von 97.

<LibraryIncident> Wissen Sie welches die Quelle diese Studie war? Oder wurden die Zahlen von den Autoren selbst erhoben?

<LibraryIncident> Ich suche jetzt einmal mit Deutsch* Nonprofit-Organisation*, die Sterne kürzen die Begriffe ab, so dass alle möglichen Endungen gefunden werden können.

<PatronIncident>...

<PatronIncident> Am besten wäre eine Tabellarische Übersicht der Organisationen, ähnlich der John Hopkins Studie von 97.

<LibraryIncident> Wissen Sie welches die Quelle diese Studie war? Oder wurden die Zahlen von den Autoren selbst erhoben? Ich suche jetzt einmal mit Deutsch* Nonprofit-Organisation*, die Sterne kürzen die Begriffe ab, so dass alle möglichen Endungen gefunden werden können.

Forming Question-Answer-Pair

MOVING
www.moving-project.eu

<*PatronIncident*>...

<*LibraryIncident*> ...

<*PatronIncident*> Am besten wäre eine Tabellarische Übersicht der Organisationen, ähnlich der John Hopkins Studie von 97.

<*LibraryIncident*> Wissen Sie welches die Quelle diese Studie war? Oder wurden die Zahlen von den Autoren selbst erhoben?

<*LibraryIncident*> Ich suche jetzt einmal mit Deutsch* Nonprofit-Organisation*, die Sterne kürzen die Begriffe ab, so dass alle möglichen Endungen gefunden werden können.

<*PatronIncident*> Am besten wäre eine Tabellarische Übersicht der Organisationen, ähnlich der John Hopkins Studie von 97.

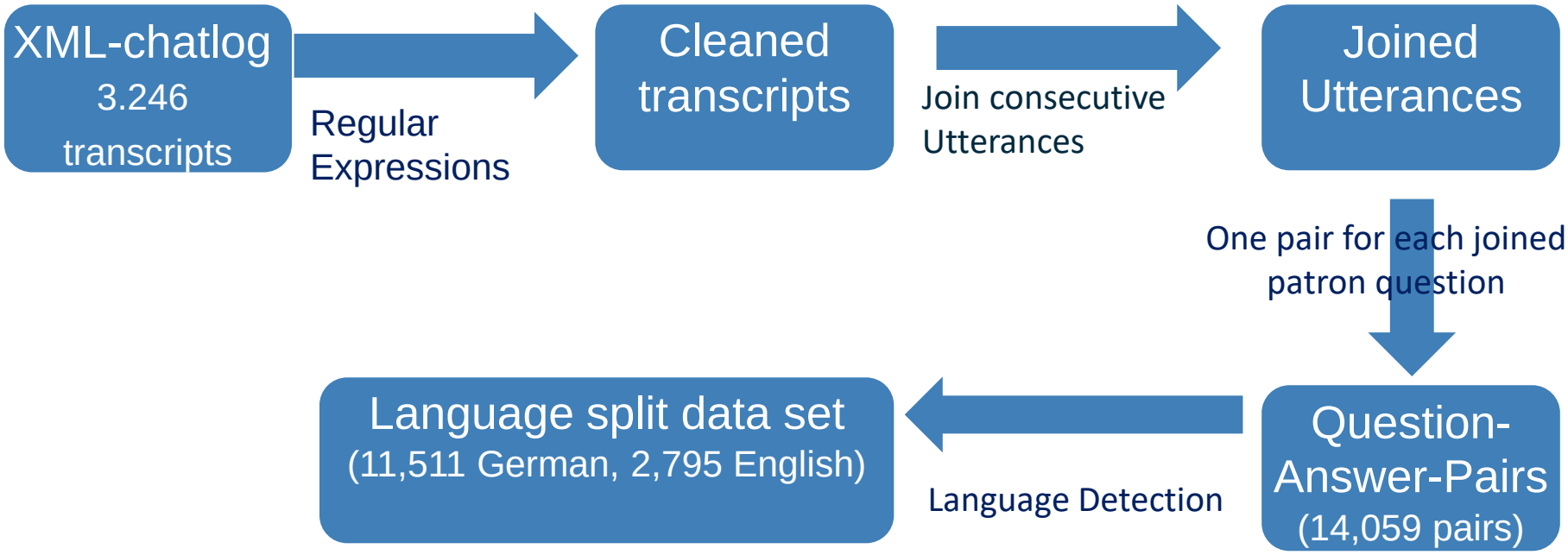
Source

<*LibraryIncident*> Wissen Sie welches die Quelle diese Studie war? Oder wurden die Zahlen von den Autoren selbst erhoben? Ich suche jetzt einmal mit Deutsch* Nonprofit-Organisation*, die Sterne kürzen die Begriffe ab, so dass alle möglichen Endungen gefunden werden können.

Target

<*PatronIncident*>...

Preprocessing Pipeline



Data	Min	Q25	Q50	Q75	Max	Mean	SD
English sources	1	4	9	18	386	14.19	19.20
English targets	1	10	19	34	697	31.34	50.84
German sources	1	4	9	18	386	13.99	16.77
German targets	1	9	17	30	790	24.53	34.14

Number of word tokens per utterance after joining

Retrieval Models

- **TF-IDF (term frequency – inverse document frequency)**
- **WCD (word centroid distance)**
  - German word vectors from *fastText* trained on *Common Crawl* and *Wikipedia*
  - English word vectors from W*ord2Vec* trained on *Google News*
- **Similarity Function: cosine similarity**

TF
TF-IDF
WCD
WCD-IDF

Retrieval Models

- **TF-IDF (term frequency – inverse document frequency)**
- **WCD (word centroid distance)**
  - German word vectors from *fastText* trained on *Common Crawl* and *Wikipedia*
  - English word vectors from W*ord2Vec* trained on *Google News*
- **Apply prefiltering (M-): allowing only suggestions with min 1 word from query improves**
  - performance
  - metric

M-TF
M-TF-IDF
M-WCD
M-WCD-IDF

Retrieval Models - KNN

- Retrieve k nearest requests with cosine similarity
- Let the respective responses cast a vote, weighted by similarity of the requests
- Return voted response

- k = {1,3,5,7}

Representational Learning

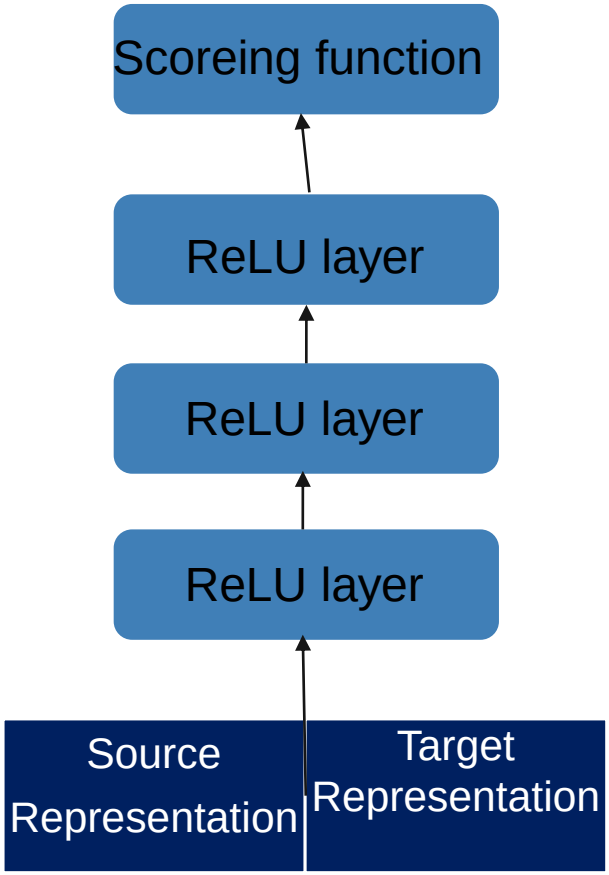
- **Produce a score, given word n-grams of the question and the response**
- **Idea:**
  - Learn from word embeddings of question and either positive or negative examples the score they produce
  - Optimize for a correct ranking

Representational Learning

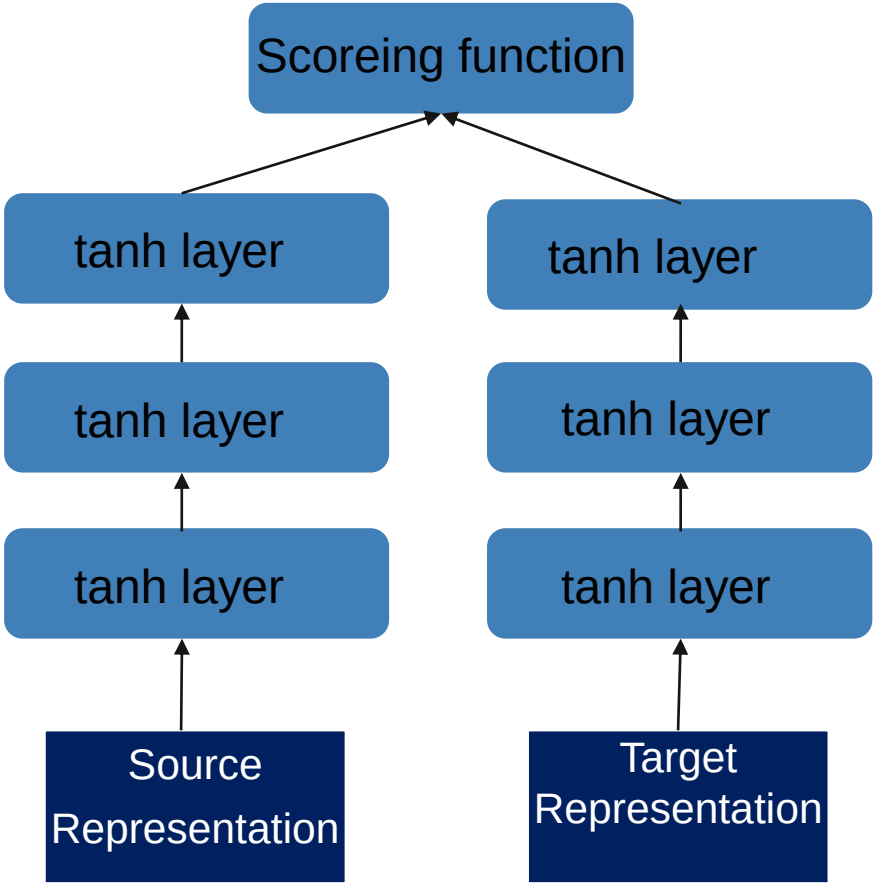
Feature representations

- **Joint**
  - bag-of-ngram representations of the question and the response are concatenated
  - fed to three hidden layers with Rectified Linear Unit (ReLU) activations and a final layer outputs the score
- **Dotproduct**
  - the questions and responses are separately encoded into vector representations
  - using cosine similarity for scoring
  - three hidden layers with Tanh activations

Joined context



Dot-product architecture



Simplified from Henderson etal. 2017

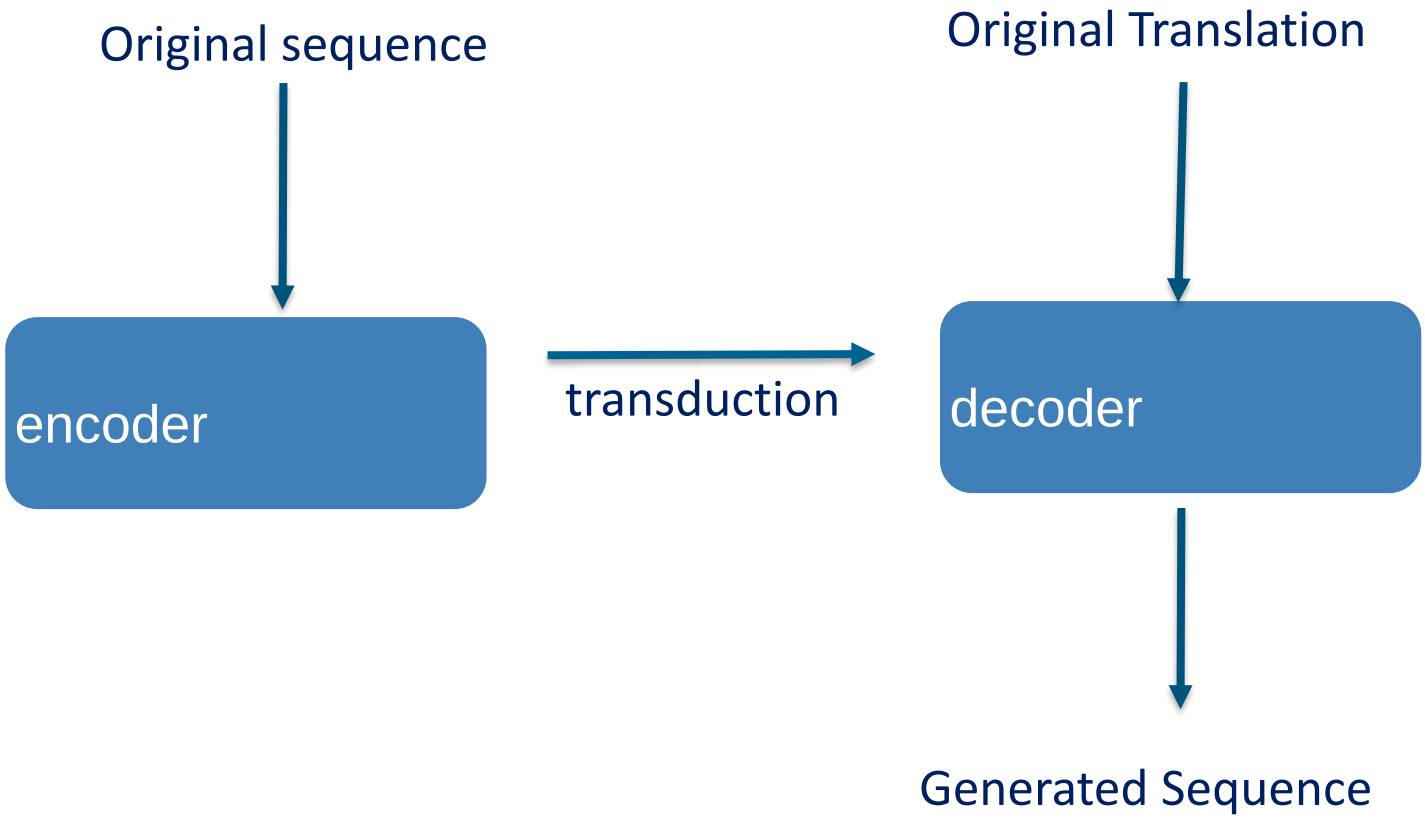
Representational Learning

- **Neural Network**
  - Soft-margin loss as objective function to rank against negative samples
  - Unigrams and bigrams as initial representation
  - Hidden layers size of 100, dropout of 0.2 on each
  - Trained for 50 epochs with Adam optimizer using an initial learning rate of 0.001
  - One to five negative samples

  - Negative sampling

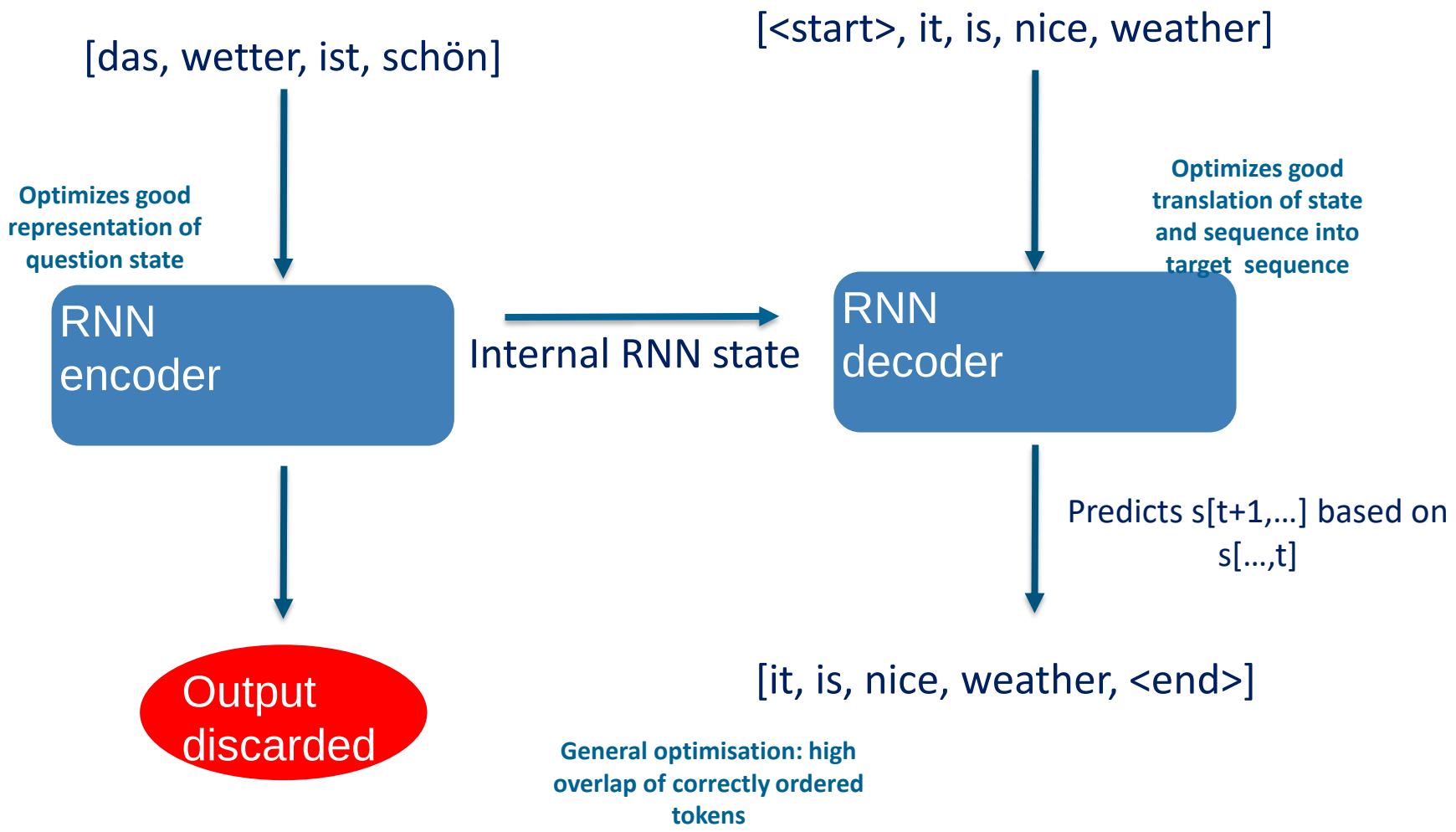
Conditioned Generation

- Idea: for a specific sequence of word generate a respective sequence
- Also called: encoder-decoder sequence to sequence models



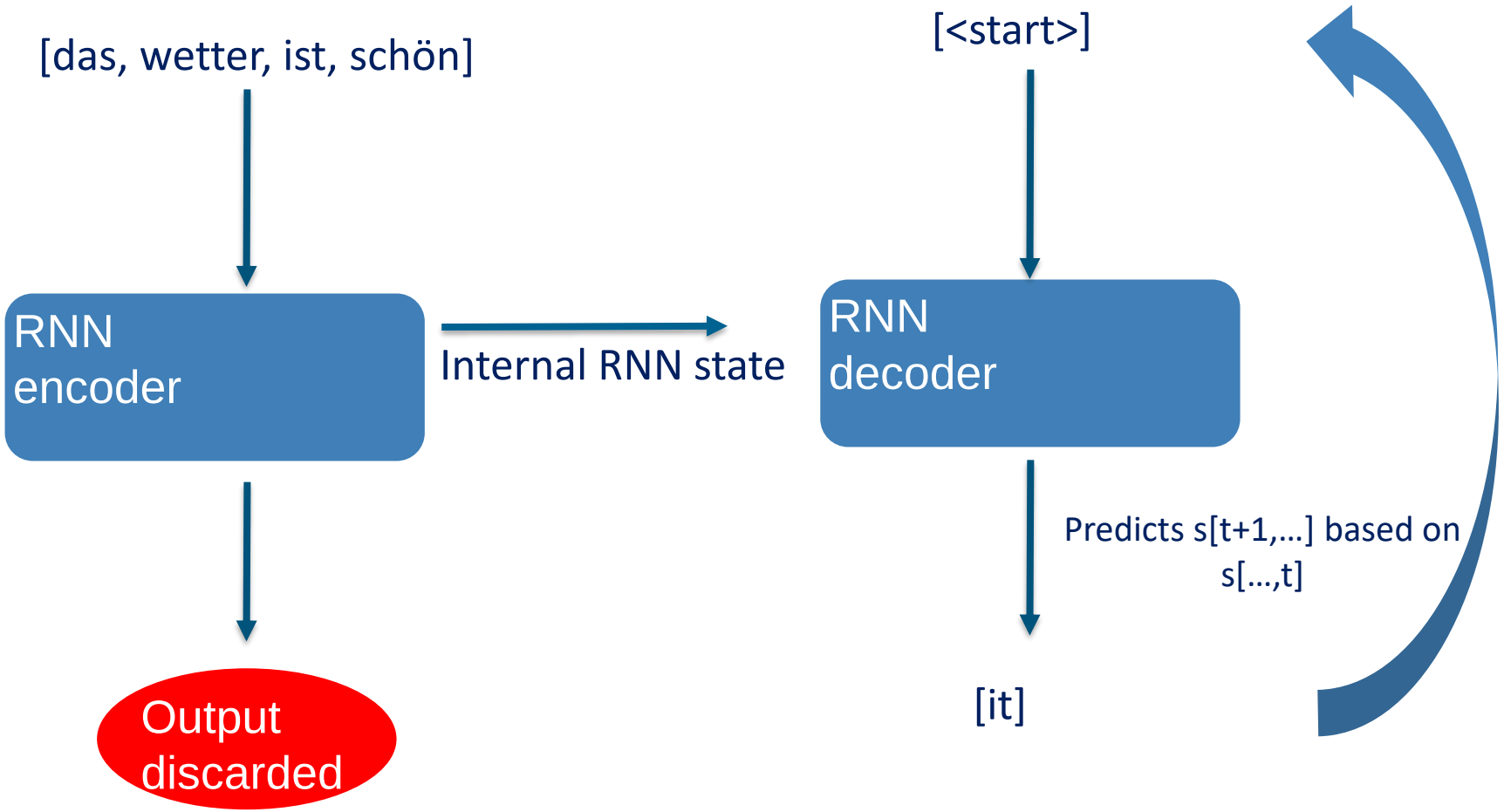
Conditioned Generation

- Training:



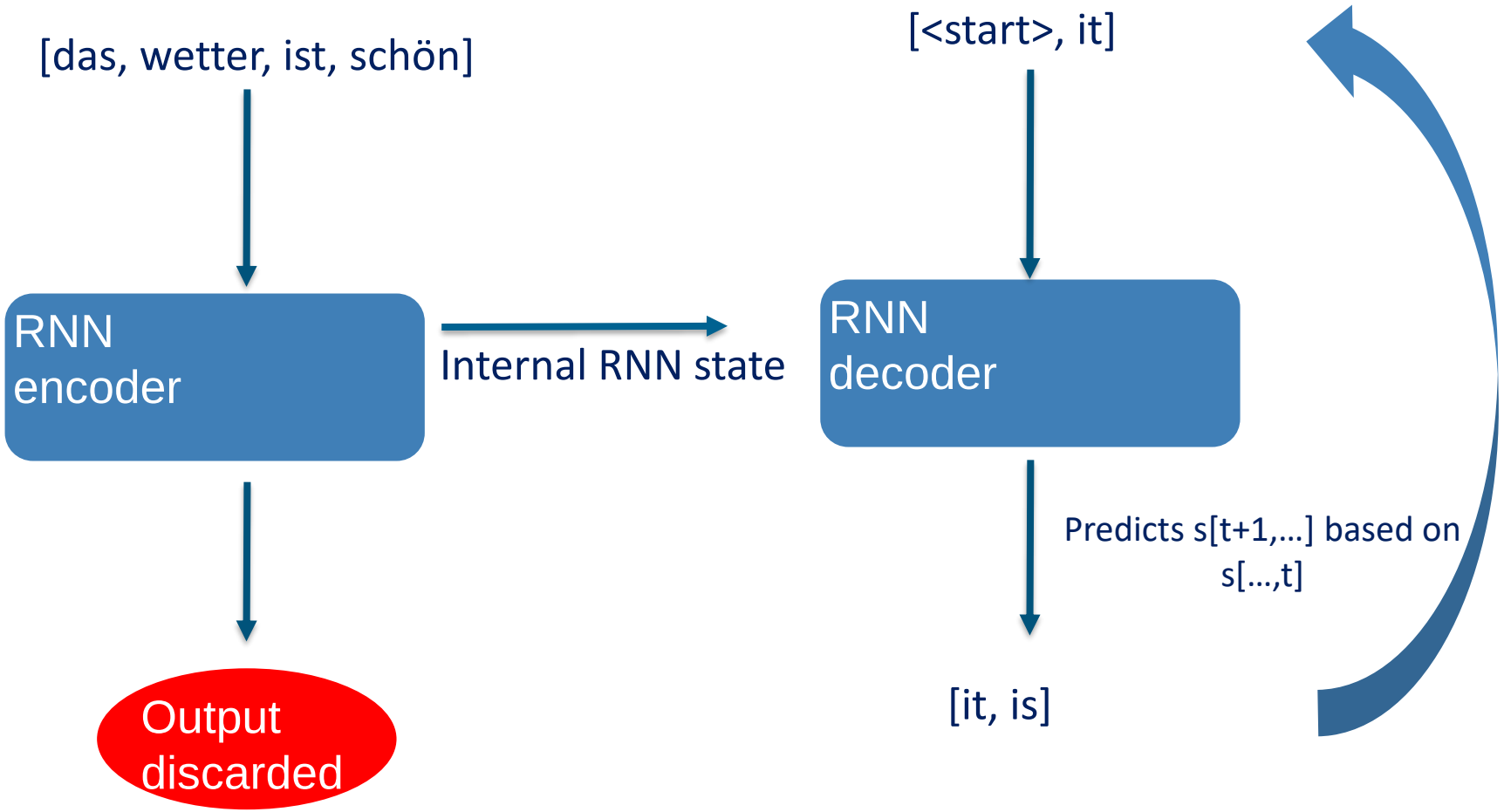
Conditioned Generation

- Prediction:



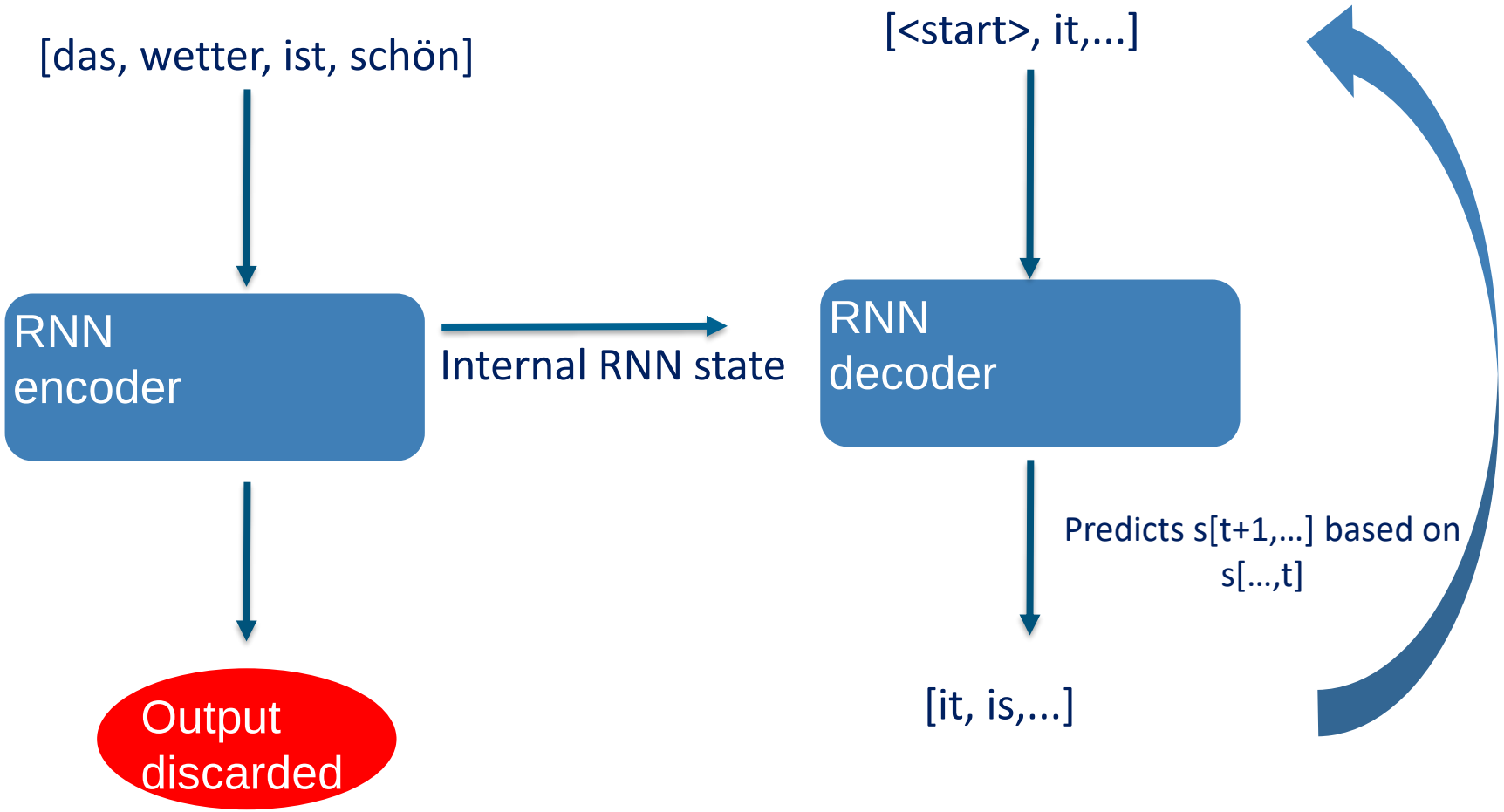
Conditioned Generation

- Prediction:



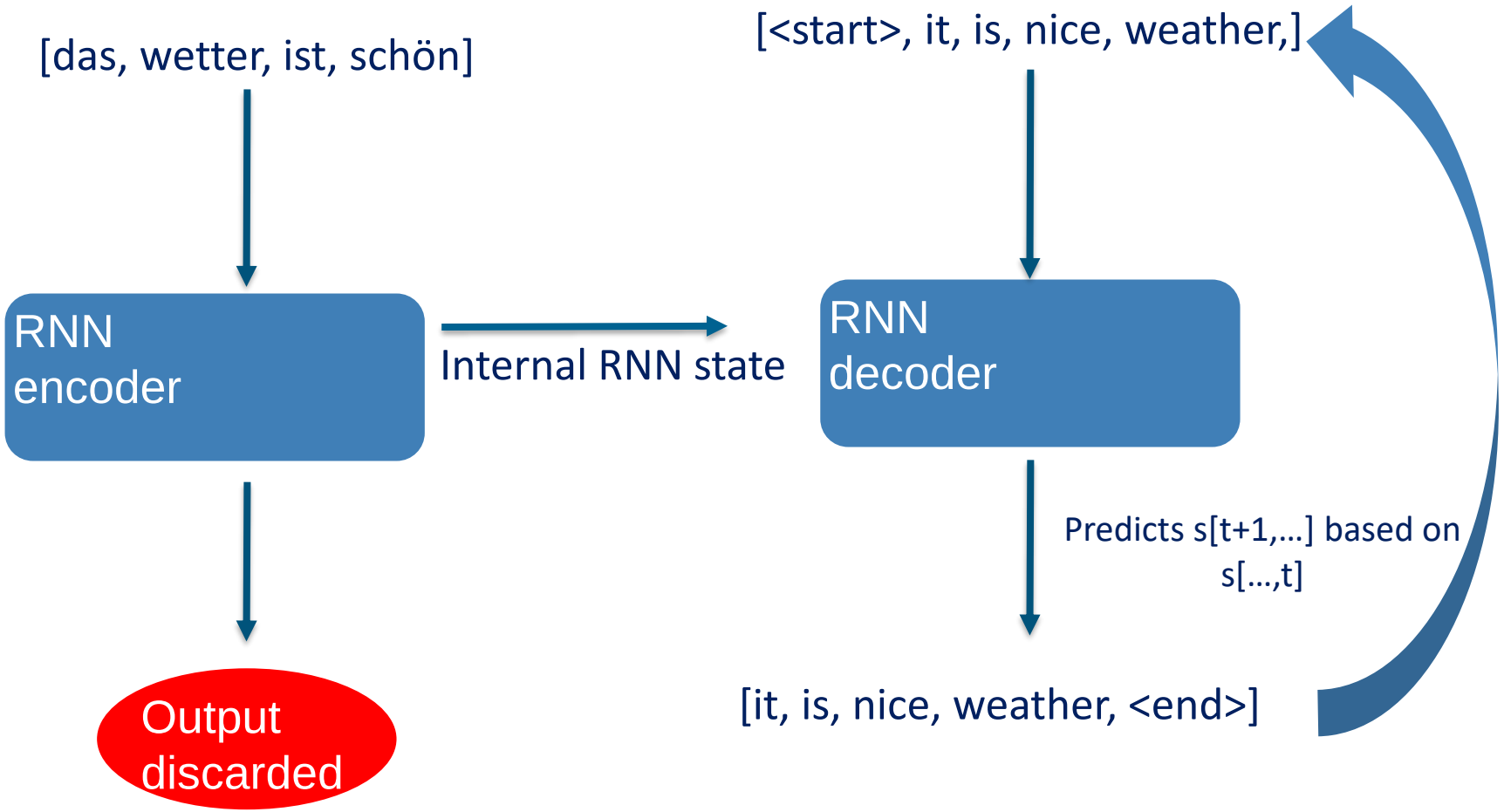
Conditioned Generation

- Prediction:



Conditioned Generation

- Prediction:

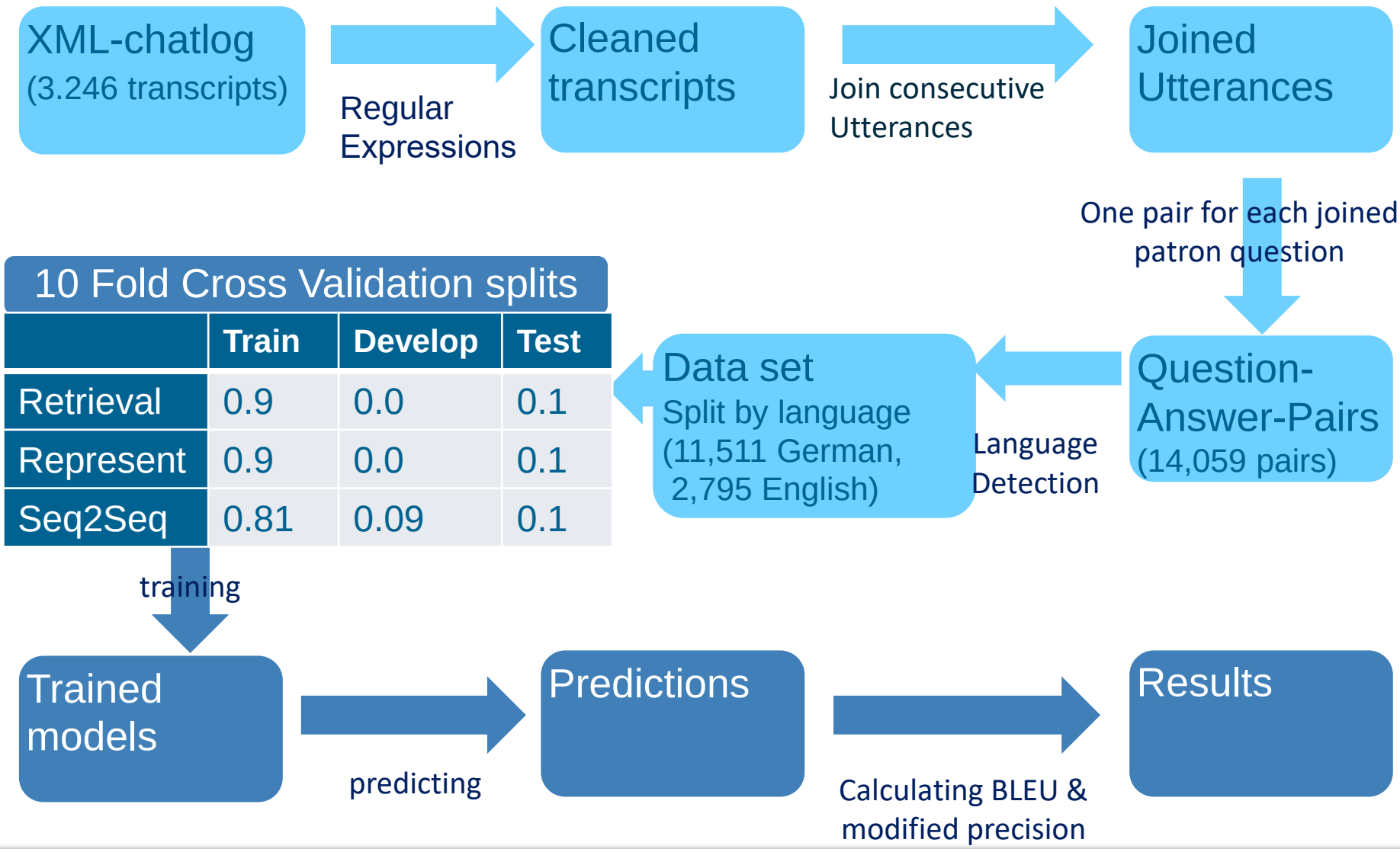


Conditioned Generation

Hyperparameters from tensorflow tutorial

- 2 hidden layers, 128 hidden units each
- All had a dropout probability of 0.2
- Learning rate: 1
- Vocabulary generated from training data

Training Process



	Train	Develop	Test
Retrieval	0.9	0.0	0.1
Represent	0.9	0.0	0.1
Seq2Seq	0.81	0.09	0.1

Evaluation Metrics: BLEU

- Modified n-gram precision:
  - Counting the shared n-grams between predicted and actual response with the upper bound of counts in the actual response
  - Then dividing by ne number of n-grams
- Bilingual evaluation understudy (BLEU)
  - Uses modified n-gram precision against several references
  - Weights modified precision
  - Applies length penalties
- We use a NLTK implementation with neighbour averageing smoothing

Example Suggestions

- **Patron Request:**
  - "How can i buy an article."
- **Library Response:**
  - "Hello and welcome to the EconDesk chat . Let me take a look at your question . One moment . Which article do you mean ?"
- **Joint Representation Suggestion:**
  - "Hello and welcome to the chat. Let me take a look at your question. One moment . Which article to you like to reed?"
- **Seq2Seq Suggestion:**
  - Hello and welcome to the chat . Let me take a look at your question. One moment.
- **TF-IDF Suggestion:**
  - "Yes"

Example Suggestions

- **Patron Request:**
  - "what exactly do I have to type in ? I there a filter just for articles or a filter only for J`s ?"
- **Library Response:**
  - "I think it is easier , if i forward both questions to a subject librarian ."
- **Joint Representation Suggestion:**
  - "Thank you for your translation . When you can search in our database EconBiz for your topic please click `` Open Access material only '' When you get the results please click at the left side under `` Type of Publication '' the button `` article '' Is that helping you ?"
- **Seq2Seq Suggestion:**
  - "Yes , your library can do that . Payment is by IFLA vouchers ."
- **TF-IDF Suggestion:**
  - "Do you mean the book with the bookshelf B 391758 , published 2009 ? Is the `` ZBW view '' on at EconBiz ? You find it under the blue line at the top of the EconBiz side . There have to stand `` Leave ZBW view '' . Than you can see the `` reserve '' - Button ."

Results

German Data (1.151 test set pairs)

Model	p1	p2	p3	BLEU
Traditional retrieval				
M-TF-IDF	26.73	17.45	15.82	23.02
M-TF	**27.19**	**17.79**	**16.09**	23.24
TF-IDF	26.77	17.51	15.89	23.04
TF	27.05	17.64	15.98	23.12
Retrieval with KNN				
M-WCD-IDF	26.95	17.41	15.78	23.44
M-WCD	27.05	17.53	15.89	23.42
WCD-IDF	27.15	17.49	15.79	23.34
WCD	27.04	17.36	15.73	23.25
Retrieval with word vectors				
1NN	26.84	17.38	15.78	**23.46**
3NN	26.86	17.56	15.93	23.19
5NN	26.72	17.54	15.93	23.21
7NN	26.83	17.58	15.96	23.09
Representation learning				
dotproduct	12.35	01.32	00.44	7.59
joint	25.84	14.46	12.66	18.26
Conditioned-generation				
seq2seq	14.80	06.23	03.93	4.10

English Data (279 test set pairs)

model	p1	p2	p3	BLEU
Traditional retrieval				
TF	26.61	14.62	12.68	**18.01**
TF-IDF	26.02	14.35	12.60	17.59
M-TF	26.84	**14.79**	**12.77**	17.90
M-TF-IDF	26.06	14.29	12.48	17.53
Retrieval with KNN				
1-NN	26.00	14.30	12.52	17.63
3-NN	25.97	14.37	12.63	17.52
5-NN	25.76	14.20	12.54	17.54
7-NN	25.51	14.21	12.56	17.78
Retrieval with word vectors				
WCD	26.01	13.96	12.12	17.35
WCD-IDF	26.35	14.31	12.44	17.54
M-WCD	25.68	13.94	12.15	17.32
M-WCD-IDF	26.16	14.24	12.38	17.62
Representation learning				
Dotproduct-n5	12.95	00.76	00.17	6.61
Joint-n5	**29.86**	11.83	09.78	10.71
Conditioned-generation				
seq2seq	17.78	08.69	06.92	4.67

Averaged over a 10 fold cross validation

Discussion

- Retrieval >= tuned joint representational
- tuned joint representational >> conditioned generation
- conditioned generation >= dotproduct representational model

- Limited data set size
  - Parameter-learning approaches likely lack training data
  - Our RNN sequence-to-sequence architecture is underexplored
- Missing context information
  - Preexperiments showed that complete contexts resulted in loss of the immediate context and worse results
- BLEU Metric
  - targeted at translations with *several* human translations or responses

Project consortium and funding agency

MOVING
www.moving-project.eu

MOVING is funded by the EU Horizon 2020 Programme under the project number INSO-4-2015: 693092

Thank you for your attention!

Any questions?

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., et al.: Tensorflow: Large-scalemachine learning on heterogeneous distributed systems. CoRR abs/1603.04467 (2016)

2. Al-Rfou, R., Pickett, M., Snaider, J., Sung, Y., Strope, B., Kurzweil, R.: Conversational contextual cues: The case of personalization and history for response ranking. CoRR abs/1606.00372 (2016)

3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473 (2014)

4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. TACL 5, 135–146 (2017)

5. Chen, B., Cherry, C.: A systematic comparison of smoothing techniques for sentence-level BLEU. In: WMT@ACL. The Association for Computer Linguistics (2014)

6. Galke, L., Saleh, A., Scherp, A.: Word embeddings for practical information retrieval In: GI-Jahrestagung. LNI, vol. P-275. GI (2017)

7. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)

8. Henderson, M., Al-Rfou, R., Strope, B., Sung, Y., Lukács, L., Guo, R., Kumar, S., Miklos, B., Kurzweil, R.: Efficient natural language response suggestion for smart reply. CoRR abs/1705.00652 (2017)

9. Huang, P., He, X., Gao, J., Deng, L., Acero, A., Heck, L.P.: Learning deep structured semantic models for web search using clickthrough data. In: CIKM. ACM (2013)

10. Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukács, L., Ganea, M., Young, P., Ramavajjala, V.: Smart reply: Automated response suggestion for email. In: KDD. ACM (2016)

11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR abs/1412.6980 (2014)

12. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: ICML. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org (2015)

References

13. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)

14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)

15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. ACL-200: 40th Annual meeting of the Association for Computational Linguistics (2002)

16. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: EMNLP. ACL (2011)

17. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5) (1988)
18. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1) (2014)
19. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
21. Vinyals, O., Le, Q.V.: A neural conversational model. CoRR abs/1506.05869 (2015)
22. Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., Weston, J.: Starspace: Embed all the things! CoRR abs/1709.03856 (2017)
23. Wu, Y., Wu, W., Yang, D., Xu, C., Li, Z., Zhou, M.: Neural response generation with dynamic vocabularies. CoRR abs/1711.11191 (2017)
24. Xu, Z., Liu, B., Wang, B., Sun, C., Wang, X., Wang, Z., Qi, C.: Neural response generation via GAN with an approximate embedding layer. In: EMNLP. Association for Computational Linguistics (2017)

Example Suggestions: Mid Conversation

- Patron Request:
  - "Yes it is published in the journal of finance . I will be very thankful if u find a free version ."
- Librarian Response:
  - "I am sorry . It is available through JSTOR - a relatively cheap database that is available in many libraries . Is there really no single library in your area ? There is no free version . However : This paper is from 1975 . Maybe you want to search for other relevant literature that is more up to date ? Or do you only need this one ?"

Example Suggestions: Mid Conversation

- Patron Request:
- "Yes it is published in the journal of finance . I will be very thankful if u find a free version ."

- K(=7)NN Suggestion:
  - "Do you have access to a library ?"

- Joint Representation Suggestion:
  - "Ok . The only way I see is a) international library loan , b) document delivery service subito c) pubilshers websites d) Visit a big library in your area that has the articles . I 'm afraid copyright will allow us to send printed copies only in most of the cases . That depends on the single journal an their allowances ."

- TF-IDF Suggestion:
  - "Please be patient , I quickly check"

- Seq2Seq Suggestion:
  - "Ok . I think it helps . Happy to assist you . Do you have other questions ?"

Example Seq2Seq generated Suggestions

German:

- "Es ist ein Hinweis . Danke für den Hinweis . Dann können Sie den Chat erst einmal schließen . Vielen Dank für den Chat und auf Wiedersehen ."
- "Soll ich Ihnen sonst noch etwas für die Recherche helfen ?"
- "Recherchieren Sie , in welchem Zusammenhang Sie diese Zeitschrift haben ?"
- "Ich bin mir nicht sicher , ob ich das richtige Dokument nicht gefunden habe . Ich kann Ihnen aber nicht sagen , wie ich herausfinde ."

English:

- "Okay , that is an article in a book . Please ask in your local library for the journal youo need the journal in its stock . In which country are you interested in ?"
- "Here is a library near you has access to a library which has access to the journal , you could order the article for you from an other library."