Towards Automated Customer Support

Momchil Hardalov[1] Ivan Koychev[1] Preslav Nakov[2]

FMI, Sofia University „St. Kliment Ohridski", Sofia, Bulgaria,
{hardalov,koychev}@fmi.uni-sofia.bg

Qatar Computing Research Institute, HBKU, Doha, Qatar,
pnakov@qf.org.qa

The 18th International Conference on Artificial Intelligence: Methodology, Systems, Applications
Varna, Bulgaria
13 September 2018

Overview

Why is my Internet connection dead?

What time is the next train from Sofia to Varna?

Where is the AIMSA '18 conference dinner going to be?

Introduction

- Need of 24×7 customer support
- Brought communication alternatives:
  e-mail, social networks, forums/message boards, live chat, self-serve knowledge base, etc.
- Employee hiring, training and satisfaction

Can we do better?

Introduction

Can we do better?

1. Can we make the machines answer closed-domain questions?

Introduction

Can we do better?

1. Can we make the machines answer closed-domain questions?
2. Can we automate the process without the need of human interference?

Introduction

Can we do better?

1. Can we make the machines answer closed-domain questions?
2. Can we automate the process without the need of human interference?
3. Can we teach the machine "to talk" like a human?

Overview

Literature Review

- Sequence-to-sequence in context of machine translation [Sutskever et al., 2014, Cho et al., 2014]

Literature Review

- Sequence-to-sequence in context of machine translation [Sutskever et al., 2014, Cho et al., 2014]
- First applications in dialog systems [Vinyals and Le, 2015]

Literature Review

- Sequence-to-sequence in context of machine translation [Sutskever et al., 2014, Cho et al., 2014]
- First applications in dialog systems [Vinyals and Le, 2015]
- In the aspect of QA systems:
  - Ubuntu Dialog [Lowe et al., 2015, Lowe et al., 2017]
  - SemEval-2016 task 3 [Boyanov et al., 2017]

Literature Review

- Sequence-to-sequence in context of machine translation [Sutskever et al., 2014, Cho et al., 2014]
- First applications in dialog systems [Vinyals and Le, 2015]
- In the aspect of QA systems:
  - Ubuntu Dialog [Lowe et al., 2015, Lowe et al., 2017]
  - SemEval-2016 task 3 [Boyanov et al., 2017]
- Twitter related bots:
  - Hierarchical structure [Serban et al., 2015]
  - Context [Sordoni et al., 2015]
  - One-shot conversations [Shang et al., 2015]

Literature Review

- Sequence-to-sequence in context of machine translation [Sutskever et al., 2014, Cho et al., 2014]
- First applications in dialog systems [Vinyals and Le, 2015]
- In the aspect of QA systems:
  - Ubuntu Dialog [Lowe et al., 2015, Lowe et al., 2017]
  - SemEval-2016 task 3 [Boyanov et al., 2017]
- Twitter related bots:
  - Hierarchical structure [Serban et al., 2015]
  - Context [Sordoni et al., 2015]
  - One-shot conversations [Shang et al., 2015]
- Hybrid models IR + NN [Qiu et al., 2017, Cui et al., 2017]

Overview

Data & Preprocessing

- Customer support on Twitter dataset [1] (Over 2M posts)

[1] `https://www.kaggle.com/thoughtvector/customer-support-on-twitter`

Data & Preprocessing

- Customer support on Twitter dataset [1] (Over 2M posts)
- Used only tweets from Apple support **49 626** dialog tuples

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

Data & Preprocessing

- Customer support on Twitter dataset [1] (Over 2M posts)
- Used only tweets from Apple support **49 626** dialog tuples
- Filtered all utterances that redirect the user to another communication channel

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

Data & Preprocessing

- Customer support on Twitter dataset [1] (Over 2M posts)
- Used only tweets from Apple support **49 626** dialog tuples
- Filtered all utterances that redirect the user to another communication channel
- Filtered all posts older then 60 days

[1] `https://www.kaggle.com/thoughtvector/customer-support-on-twitter`

Data & Preprocessing

- Customer support on Twitter dataset [1] (Over 2M posts)
- Used only tweets from Apple support **49 626** dialog tuples
- Filtered all utterances that redirect the user to another communication channel
- Filtered all posts older then 60 days
- Divided the dialogs in train and test by post time:
  45 582 for training and 4044 for testing (last 5 days)

[1] `https://www.kaggle.com/thoughtvector/customer-support-on-twitter`

Data & Preprocessing

- Customer support on Twitter dataset [1] (Over 2M posts)
- Used only tweets from Apple support **49 626** dialog tuples
- Filtered all utterances that redirect the user to another communication channel
- Filtered all posts older then 60 days
- Divided the dialogs in train and test by post time:
  45 582 for training and 4044 for testing (last 5 days)
- Replaced slang, special words (users, urls, etc.), short forms

[1] `https://www.kaggle.com/thoughtvector/customer-support-on-twitter`

Information Retrieval

- Based on ElasticSearch [2]

[2] `https://www.elastic.co/products/elasticsearch`

Information Retrieval

- Based on ElasticSearch [2]
- Using BM25 (Improved TF.IDF) [Robertson and Zaragoza, 2009]

[2] `https://www.elastic.co/products/elasticsearch`

Information Retrieval

- Based on ElasticSearch [2]
- Using BM25 (Improved TF.IDF) [Robertson and Zaragoza, 2009]
- English analyzer

[2] https://www.elastic.co/products/elasticsearch

Information Retrieval

- Based on ElasticSearch [2]
- Using BM25 (Improved TF.IDF) [Robertson and Zaragoza, 2009]
- English analyzer
- whitespace- and punctuation-based tokenization

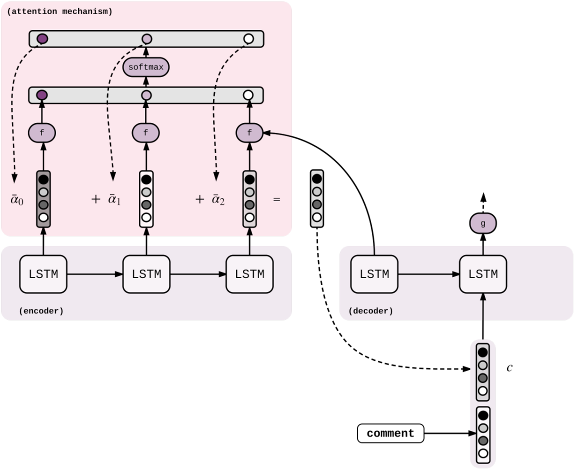
[2] https://www.elastic.co/products/elasticsearch

Information Retrieval

- Based on ElasticSearch [2]
- Using BM25 (Improved TF.IDF) [Robertson and Zaragoza, 2009]
- English analyzer
- whitespace- and punctuation-based tokenization
- Word 3-grams

[2] https://www.elastic.co/products/elasticsearch

Seq2seq model with Bahdanau attention [Bahdanau et al., 2014]



[3]

---

[3] https://guillaumegenthial.github.io/sequence-to-sequence.html

Sequence-to-Sequence II

- Embedding layer (8192 words)
  (GloVe + positional)

Sequence-to-Sequence II

- Embedding layer (8192 words)
  (GloVe + positional)
- Encoder: Bi-directional LSTM layer (512)

Sequence-to-Sequence II

- Embedding layer (8192 words)
  (GloVe + positional)
- Encoder: Bi-directional LSTM layer (512)
- Decoder: Two stacked LSTM layers (2×512)

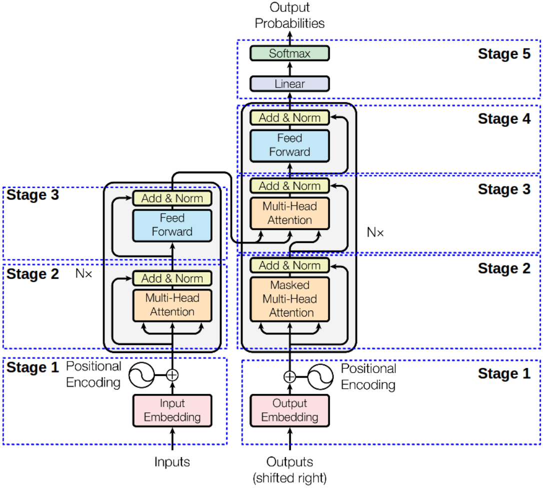
Sequence-to-Sequence II

- Embedding layer (8192 words)
  (GloVe + positional)
- Encoder: Bi-directional LSTM layer (512)
- Decoder: Two stacked LSTM layers (2×512)
- Dropout (0.2) after LSTMs

Sequence-to-Sequence II

- Embedding layer (8192 words)
  (GloVe + positional)
- Encoder: Bi-directional LSTM layer (512)
- Decoder: Two stacked LSTM layers (2×512)
- Dropout (0.2) after LSTMs
- Adam with exp. decay [Kingma and Ba, 2015]

Attention is all you need [Vaswani et al., 2017]



[4]

[4] [Vaswani et al., 2017]

Transformer II

- Sinusoidal positional embedding (8192 words)

Transformer II

- Sinusoidal positional embedding (8192 words)
- $N = 4$, $d_{model} = 256$, $d_{inner} = 512$, $d_k = 64$, $d_v = 64$

Transformer II

- Sinusoidal positional embedding (8192 words)
- $N = 4$, $d_{model} = 256$, $d_{inner} = 512$, $d_k = 64$, $d_v = 64$
- Dropout (0.1) after sub-layers

Transformer II

- Sinusoidal positional embedding (8192 words)
- $N = 4$, $d_{model} = 256$, $d_{inner} = 512$, $d_k = 64$, $d_v = 64$
- Dropout (0.1) after sub-layers
- Adam as suggested in [Vaswani et al., 2017]

Overview

Metrics

Our metrics are divided in two categories:

Metrics

Our metrics are divided in two categories:

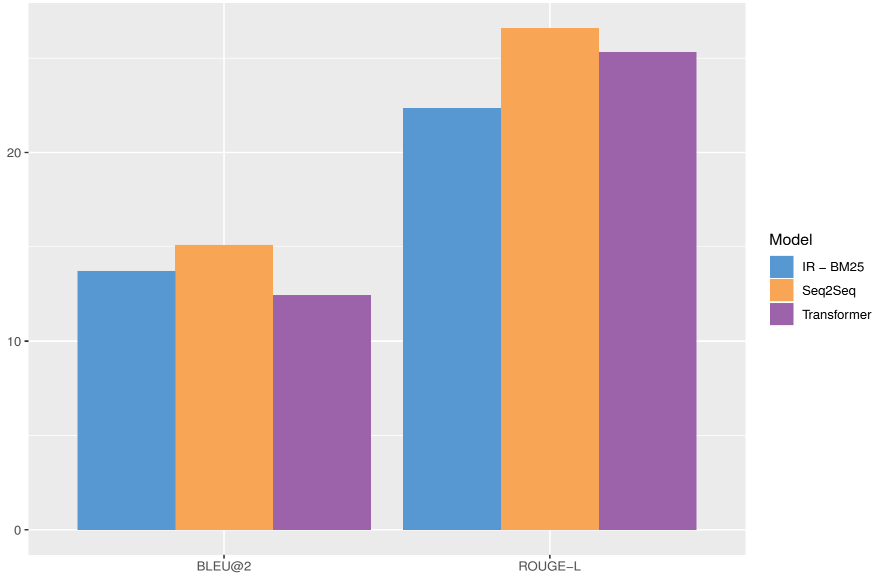
- Word Overlap Measures
  - BLEU@2 [Papineni et al., 2002]
  - ROUGLE-L [Lin, 2004]

Metrics

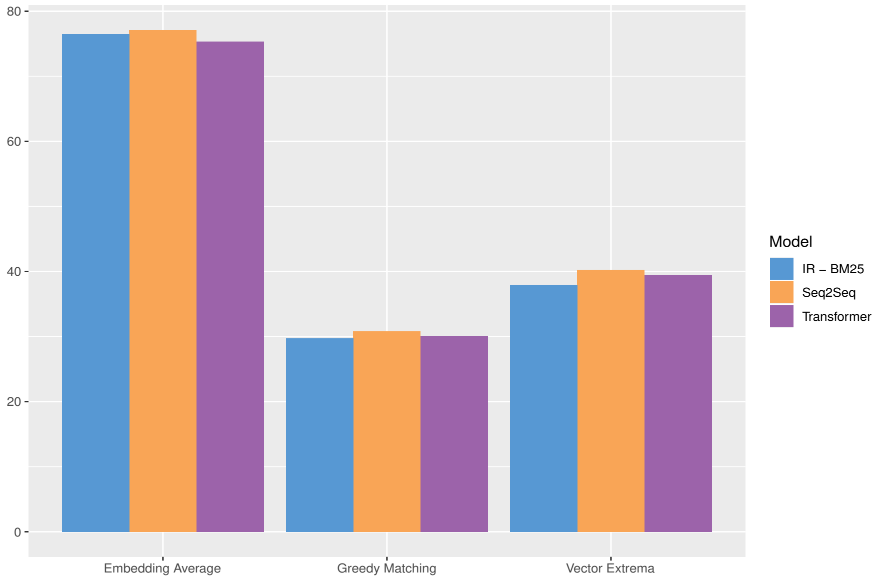
Our metrics are divided in two categories:

- Word Overlap Measures
  - BLEU@2 [Papineni et al., 2002]
  - ROUGLE-L [Lin, 2004]
- Semantic Evaluation Measures
  - Embedding Average
  - Greedy Matching [Rus and Lintean, 2012]
  - Vector Extrema [Forgues et al., 2014]

Word Overlap Measures



Semantic Evaluation Measures



Examples I

Question & Answer	Response
(1) my phones speaker just broke down on me thanks <user> do better	**IR:** <user> we can help with your device . do you notice the slowness when performing a particular activity ?
<user> we are here to help. which iphone and ios version are you using? are you having a sound quality issue or is there no sound at all?	**Seq2Seq:** <user> that is not the experience we want you to have . which iphone model do you own ? **Transformer:** <user> thanks for reaching out to us . we are always happy to help . can you tell us which version of ios 11 you are using ?

Examples II

Question & Answer	Response
(2) i am trying to type it not fuckin i.t, your product is absolute garbage. figure it out. <user> <user> <user> lets take a closer look at this to get your device working right. which model do you have and that is the exact os version number installed? have you tried any steps so far and when did this start happening?	**IR:** <user> thank you for reaching out . we want to help . to start , lets go to settings > general > about and see which version of ios your iphone is on please . **Seq2Seq:** <user> we would be happy to look into that is going on . can you tell us more? **Transformer:** <user> heres what you can do to work around the issue until it is fixed in a future software update : <url>

Examples III

Question & Answer	Response
(3) capaz que si actualizo ios me teléfono se convierta en ladrillo :(....ya no confío en sus productos <user> <user> we offer support via twitter in english. get help in spanish here: <url> or join <url>	**All models:** <user> we offer support via twitter in english get help in spanish here <url> or join <url>

Examples IV

Question & Answer	Response
(4) <user> excuse me , i am got a refund confirmation email but nothing happened . do i have to do anythings ? → <user> thanks for reaching out to us about your refund ! to clarify , was this a refund for a purchase in the itunes store ? or was this a product purchase ? <user> thanks for reply . i am purchased music in itunes store . <user> thanks. our itunes store support team would have the best information for you on that. you can contact them here: <url>	**IR:** <user> alright . our itunes support team will be able to look at this with you . please reach out to them here : <url> **Seq2Seq:** <user> we are investigating and will update our system status page as more info becomes available <url> **Transformer:** <user> thanks for reaching out . we would recommend leaving that request on our feedback page : <url>

Overview

Conclusions & Future Work

Conclusions

Conclusions & Future Work

Conclusions

- Information Retrieval

Conclusions & Future Work

Conclusions

- Information Retrieval
  + Very fast with smooth utterances

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses

Conclusions & Future Work

Conclusions

- Information Retrieval
  + Very fast with smooth utterances
  - Cannot generate new responses
- Attentive Seq2Seq

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data
  - \- Slower training

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data
  - \- Slower training
- Transformer

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data
  - \- Slower training
- Transformer
  - \+ Fast training, no recurrence

Conclusions & Future Work

Conclusions

- Information Retrieval
  - + Very fast with smooth utterances
  - - Cannot generate new responses
- Attentive Seq2Seq
  - + Can work with less data
  - - Slower training
- Transformer
  - + Fast training, no recurrence
  - - Very large parameter space

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data
  - \- Slower training
- Transformer
  - \+ Fast training, no recurrence
  - \- Very large parameter space

Future Work

- Combine the three approaches into an ensemble

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data
  - \- Slower training
- Transformer
  - \+ Fast training, no recurrence
  - \- Very large parameter space

Future Work

- Combine the three approaches into an ensemble
- Experiment with byte-pair encoding [Sennrich et al., 2016]

Conclusions & Future Work

Conclusions

- Information Retrieval
  - \+ Very fast with smooth utterances
  - \- Cannot generate new responses
- Attentive Seq2Seq
  - \+ Can work with less data
  - \- Slower training
- Transformer
  - \+ Fast training, no recurrence
  - \- Very large parameter space

Future Work

- Combine the three approaches into an ensemble
- Experiment with byte-pair encoding [Sennrich et al., 2016]
- Handle questions whose correct answers evolve over time

References I

Bahdanau, D., Cho, K., and Bengio, Y. (2014).
Neural machine translation by jointly learning to align and translate.
arXiv preprint arXiv:1409.0473.

Boyanov, M., Nakov, P., Moschitti, A., Da San Martino, G., and Koychev, I. (2017).
Building chatbots from forum data: Model selection using question answering metrics.
In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 121–129, Varna, Bulgaria.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014).
Learning phrase representations using RNN Encoder–Decoder for statistical machine translation.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP '14, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., and Zhou, M. (2017).
SuperAgent: A customer service chatbot for e-commerce websites.
In *Proceedings of the Association for Computational Linguistics 2017, System Demonstrations*, ACL '17, pages 97–102, Vancouver, Canada.

Forgues, G., Pineau, J., Larchevêque, J.-M., and Tremblay, R. (2014).
Bootstrapping dialog systems with word embeddings.
In *Proceedings of the NIPS Workshop on Modern Machine Learning and Natural Language Processing*, Montreal, Canada.

Kingma, D. P. and Ba, J. (2015).
Adam: A method for stochastic optimization.
In *Proceedings of the 2015 International Conference on Learning Representations*, ICLR '15, San Diego, California.

References II

Lin, C.-Y. (2004).
ROUGE: A package for automatic evaluation of summaries.
In *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015).
The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems.
In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '15, pages 285–294, Prague, Czech Republic.

Lowe, R. T., Pow, N., Serban, I. V., Charlin, L., Liu, C.-W., and Pineau, J. (2017).
Training end-to-end dialogue systems with the Ubuntu dialogue corpus.
Dialogue & Discourse, 8(1):31–65.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).
BLEU: A method for automatic evaluation of machine translation.
In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J., and Chu, W. (2017).
AliMe chat: A sequence to sequence and rerank based chatbot engine.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 498–503, Vancouver, Canada.

Robertson, S. and Zaragoza, H. (2009).
The probabilistic relevance framework: BM25 and beyond.
Found. Trends Inf. Retr., 3(4):333–389.

References III

Rus, V. and Lintean, M. (2012).
A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics.
In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montreal, Canada.

Sennrich, R., Haddow, B., and Birch, A. (2016).
Neural machine translation of rare words with subword units.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 1715–1725.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. (2015).
Hierarchical neural network generative models for movie dialogues.
CoRR, abs/1507.04808.

Shang, L., Lu, Z., and Li, H. (2015).
Neural responding machine for short-text conversation.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP'15, pages 1577–1586, Beijing, China.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015).
A neural network approach to context-sensitive generation of conversational responses.
In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '15, pages 196–205, Denver, Colorado.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to sequence learning with neural networks.
In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, NIPS '14, pages 3104–3112, Montreal, Canada.

References IV

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).
Attention is all you need.
In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, NIPS '17, pages 5998–6008, Long Beach, California.

Vinyals, O. and Le, Q. V. (2015).
A neural conversational model.
CoRR, abs/1506.05869.

Results

	BLEU@2	ROUGE-L
IR - BM25	13.73	22.35
Seq2Seq	**15.10**	**26.60**
Transformer	12.43	25.33

Table: Results based on word-overlap measures.

	Embedding Average	Greedy Matching	Vector Extrema
IR - BM25	76.53	29.72	37.99
Seq2Seq	**77.11**	**30.81**	**40.23**
Transformer	75.35	30.08	39.40

Table: Results based on semantic measures.

Metrics

Greedy Matching

$$greedy(v, w) = \frac{\sum_{v \in u_1} weight(v) * \max_{w \in u_2} cos(v, w)}{\sum_{v \in u_1} weight(v)} \quad (1)$$

$$simGreedy(v, w) = \frac{greedy(v, w) + greedy(w, v)}{2} \quad (2)$$

Vector Extrema

$$extrema(u_i) = \begin{cases} \max u_i, & if \max u_i \geq |\min u_i| \\ \min u_i, & \text{otherwise} \end{cases} \quad (3)$$

Adam for Transformer

Learning rate in Transformer

$$lrate = d_{model}^{-0.5} \cdot \min\left(step\_num^{-0.5}, step\_num \cdot warmup\_steps^{-1.5}\right) \quad (4)$$

Discussions

- Seq2Seq performed best:
  - Embedding Average by +0.58
  - Greedy Matching by +0.73
  - Vector Extrema by +0.83
  - 15.10 on BLEU@2 (+1.37)
  - 26.60 on ROUGE-L (+1.27)
- Transformer is ranked 2nd in 3/5