Paper Abstract Generation Using Neural Network

Kevin Xu[1], Boliang Zhang[2], Qingyun Wang[2], Heng Ji[2]
[1]Bayview Secondary School, Richmond Hill, Ontario, Canada
[2]Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY, USA

ABSTRACT

One of the most powerful aspects of AI (Artificial Intelligence) is machine learning. It uses statistical methods to give computers the ability to progressively improve performance on a specific task, with data, without being explicitly programmed. In this poster, we present an Automatic Paper Abstract Generation framework that generates a paper abstract given an input title. A large dataset of scientific papers has been collected in advance and pre-trained by computer using a sequence to sequence model so that the machine learns about the natural language syntax on the paper abstract. To use it, users input any title and the machine generates a fluent paper abstract by Machine Learning and Natural Language Processing. The Automatic Paper Abstract Generation aims to assist humans to write paper abstracts more efficiently by generating an initial draft. Quantitative and qualitative analysis show that our method is comparable with the state-of-the-art techniques.

RESEARCH OBJECTIVES

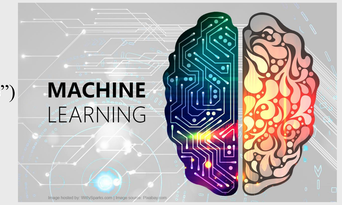
- Explore the idea of using neural network for automatic language generation.
- Implement an end-to-end paper abstract generation system using a supervised sequence-to-sequence neural network[1][2].

BACKGROUND

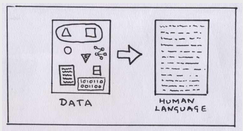
- **Introduction**

AI (Artificial Intelligence) makes the machine become “intelligent” by giving it a large amount of data to process and learn, and computers can be trained to accomplish specific tasks by recognizing patterns in the data. Many applications are powered by AI: autonomous driving, speech recognition, email spam filter, etc.

- **Real life examples**
  Google Maps (faster route)
  Spotify (music suggestions)
  AlphaGo (plays the game “go”)



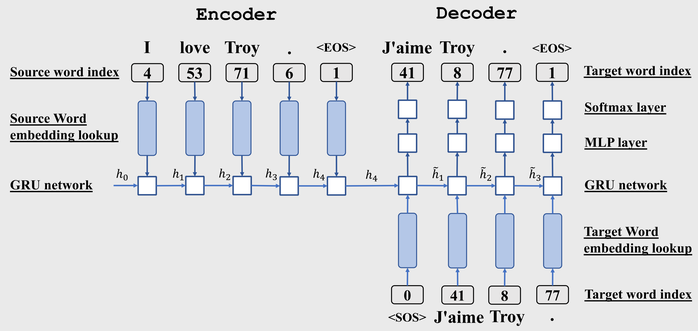
- **Major research fields of AI**
  Machine Learning
  Natural Language Processing
  Robotics
  Computer Vision



- **Paper abstract generation**
  We aim to assist humans to write paper abstracts more efficiently by generating an initial draft.

- **Sequence to Sequence Network**
  We use a sequence-to-sequence neural network in our experiment. It trains models from one sequence to another sequence in a different domain. The architecture is broadly used in many machine learning tasks: machine translation, music composition, image captioning, speech recognition, etc.

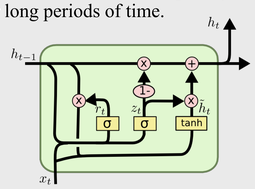
MODEL REPRESENTATION



Sequence-to-Sequence neural network is about training models to convert sequences from one domain (e.g. sentences in English) to sequences in another domain (e.g. the same sentences translated to French).
Encoder network condenses an input sequence into a vector, and a **decoder network** unfolds that vector into a new sequence.

METHODS & FORMULAS

- **Gated Recurrent Unit (GRU) Network:**
  Gated Recurrent Unit network is explicitly designed to remember information of a sequence for long periods of time.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

- **Softmax Layer**
  The output of the softmax function is equivalent to a categorical probability distribution, and it tells you the probability that any of the classes are true.

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}}$$

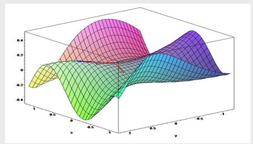
- **Cost Function**
  Cost functions are used to estimate how badly models are performing.

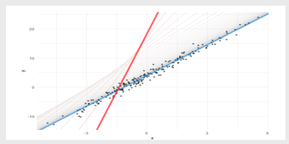
$$\mathcal{L}(\theta) = -\sum_{c=1}^{M} y_c \log(p_c)$$

- **Gradient descent**

Gradient descent is used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). Finding the minimum is to minimize the errors so that machine can produce better results.

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} \mathcal{L}(\theta)$$





RESULTS

- **Data**

We collected a data set of 10,874 paper title and abstract pairs from the ACL Anthology Network (until 2016) for our experiments. We randomly dividing them into training (80%), validation (10%), and testing (10%) sets. On average, each title and abstract pair includes 9 and 116 words, respectively.

- **Qualitative Analysis**

Input title	native language identification using large scale lexical features <EOS>
Desired abstract	this paper describes an effort to perform native language identification (nli) using machine learning on a large amount of lexical features . the features were collected from sequences and collocations of bare word forms , suffixes and character n-grams amounting to a feature set of several hundred thousand features . these features were used to train a linear support vector machine (svm) classifier for predicting the native language category . <EOS>
Output abstract	we present a novel approach to automatically detect the native features of a language . we show that the proposed model is trained on a corpus of wikipedia data and the features of the corpus of features with the features . we show that the proposed model yields the best published and the best published of our best . <EOS>

- **Quantitative Analysis**

Method	METEOR	ROUGE-L
LSTM-LM	8.7	15.1
Seq2Seq	13.5	19.2

METEOR[3] is a metric for the evaluation of machine translation output.
ROUGE-L[4] is a metric for the evaluation of automatic summarization output.

CONCLUSIONS & FUTURE DIRECTIONS

We implemented an effective program for drafting an abstract automatically based on an input title. For future work, we will explore to use this program on other generation tasks, e.g. Automated Journalism where news articles are being generated automatically based on a title input.

ACKNOWLEDGEMENTS

I sincerely thank my PhD mentor Boliang Zhang for the important guidance and significant support he gave. I also express my gratitude to the director of BLENDER Lab, Heng Ji for offering me this great opportunity. Lastly, I would like to thank all the lab members who rendered their help during the period of my research.

REFERENCES

[1] Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji and Kevin Knight, Paper Abstract Writing through Rewriting Mechanism, Proc. ACL 2018
[2] Ilya Sutskever, Oriol Vinyals and Quoc V. Le, Sequence to Sequence Learning with Neural Networks, Proc. NIPS 2014
[3] Satanjeev Banerjee and Alon Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proc. NAACL 2010
[4] Chin-Yew Lin, ROUGE: A Package for Automatic Evaluation of Summaries, Proc. the Workshop on Text Summarization Branches Out 2004

DO NOT USE THIS SPACE