University of Technology Sydney

Advanced Data Analytics Algorithms

32513

Master of Information Technology

Submitted By:

[Qingyuan Li]

[13016389]

[9th Oct 2019]

[github: https://github.com/chris0906/UTS_ML2019_ID13016389*]*

[video: https://www.youtube.com/watch?v=AgTV-b7Qpno&t=37s*]*

Contents

Machine translation between French and Chinese

Background

Nowadays, as the machine learning area becomes more and more popular in the industry, it has been applied in so many areas, one of the regions is the NLP (Natural Language Processing). As human translation fee goes higher and higher along the way, and machine learning technology becomes more and more mature, machine translation will be the right substitution in the future. However, the language of the same family translated by machine has done an excellent performance, such as French and Spanish, but the language of the different family has not done well for most of the cases, one example is French and Chinese. So, our project is mainly focused on achieving Chinese and French translation in terms of text. From a technical perspective, the NLP area has experienced a long development history. The first model introduced was classical neural language model (Bengio et al.), we can give the word to predict the next word by using this model, with which a scenario like a spell correction feature can be used. The next stage was multi-task learning in deep neural networks, Rich Caruana was introducing the concept in 1993, and it was used in areas like road tracking and pneumonia prediction(Caruana R 1997), and after that, the multi-task learning was first applied in NLP by Collobert and Weston in 2008. Then word embedding came up into view in 2013. Then NLP had brought in neural networks in 2013 and 2014, and the widely used three neural networks were Convolutional neural network, Recurrent Neural Network, Recursive neural network. Recurrent Neural Network is an optimal technique to deal with the dynamic input sequences which are a common case in NLP, Convolutional neural network was originally a technology widely used in the field of computer vision, but now it is also applied in NLP. Both RNN and CNN treat language as a sequence. From a linguistic perspective, however, language is inherently hierarchical, words are grouped into higher-order phrases and clauses that themselves can be grouped recursively according to a set of production rules. The linguistically inspired idea of sentences as trees rather than sequences gave rise to a recursive neural network. Apart from that, there are also some other models in NLP, but it's out of our projects' range.

Research Aims and Objectives

Our aim is to let machine to be able to translate French and Chinese to each other.
the aim can be divided into 5 objectives:

1. to collect or purchase corpus for our data source.
2. to clear the corpus, and restore them to one to one corresponding relation.
3. to make a mapping table, in which it shows the mapping between original verb and changed verb.
4. to make all the sentences aligned to each other.
5. train our model.

Significance and innovation

Significance:

The machine translation is in massive demand in the modern society, things like traveling, simultaneous interpretation, online translation have been happening all over the place all the time, not to mention that people's communication is more intertwined across the globe as our technology is going to bring us closer. Apart from that, there are some niche areas in the French and Chinese translation that a professional Interpreter has rarely touched, such as patent translation and academic translation. For instance, if we want to invite an interpreter to translate in these areas, not only the interpreter must learn these related words in these areas, but also would pay a considerable amount of fee. in this case, it has enormous benefits for doing such a work.

Innovation:

People prefer to use the Recurrent neural network and convolutional neural network in the past years; however, these two models are processing data in a sequential way instead of a structural way. In some cases, the sequential way would make the sentence ambiguity, but the structural way can get better outcomes.

Tasks Breakdown

1. comparing different neural networks model
2. analyze advantages and disadvantages in our project scenarios
3. data preparation and processing
4. construct training data based on all the pre-processed data
5. calculate weights and errors
6. implement code and test
7. evaluation and analysis
8. improvement and recommendation

Timetable and Plan

Tasks/time	Starting time	Ending time	duration
Recruit talent	2020/1/1	2020/2/1	31 days
Preparation & procedure	2020/2/2	2020/2/29	27 days
Purchase corpus	2020/3/1	2020/3/15	14 days
Data pre-processing	2020/3/16	2020/5/10	55 days
Phrases restoration	2020/4/3	2020/7/8	96 days
Phrases alignment	2020/6/18	2020/9/10	84 days
Construct training data	2020/9/11	2020/9/21	10 days
Training data	2020/9/22	2020/10/22	30 days
Evaluation result	2020/10/23	2020/11/1	9 days
Analyze and improve	2020/11/2	2020/12/31	59 days

Outcomes

1. The product can translate French to Chinese or Chinese to French in text.
2. The investors could have a long term profit by having this product, because we could implant our technology in so many areas.
3. The investors could have potential abilities to advance the technology into not only text translation but also text recognition and speech recognition areas. It provides an excellent base for it.

Budgets

Component	Cost (dollar)

Corpus purchase	10,000
staff salary	630,000
Space rental fee	48,000
Server purchase	12,000
Technology support	10,000
System testing	10,000
Marketing strategy	30,000
Contingency	100,000

Personnel

No	Roles	Responsibility
1	Project manager	Manage the project
2	Language analyst	Pre-process data source
3	Data analyst	Analyze data
4	Machine learning engineer	Building model and training
5	Software engineer (4)	Implement software
6	Marketing analyst	Market research

References

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41-75.

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM.