Research Project Proposal:

Neural Statistical Language Model using Logic Tensor Networks

Francisco Javier Medel Medina
MSc Data Science

Supervisor: Dr Tillman Weyde

April 2018

1. Purpose

1.1 Introduction

Natural Language Processing (NLP) is a field in Machine Learning, which has still room for improvement. The idea is to use Statistical Relational Learning (SRL) to improve the results of predictions. We plan to use the Logic Tensor Networks(LTN) framework for a Neural Network architecture in Statistical Language Model, i.e., a model that predicts the next token in a given context. One way to improve Neural Network Language Models is to add knowledge about the grammatical rules that exist in the English Language to the model. We will use grammar rules as Background Knowledge (BK) to build a model using Logic Tensor Networks (LTN), to improve the results of Neural Statistical Language Models.

1.2 Research question

Can the use of grammatical rules as Background Knowledge be applied in the Language Modelling Techniques to increase the accuracy of predicted phases in a model-built using Logic Tensor Network framework?

1.3 Objectives

- Statistical Relational Learning (SLR)
  - Create grammatical rules as input for Background Knowledge (BK)
  - Improve existing Neural Statistical Language Models with Background Knowledge (BK) on Logic Tensor Network (LTN) framework.
- Explore existing Techniques and compare results.
  - Try other approaches like n-grams (unigrams, bigrams, trigrams) and Bag Of Words (BOW) that can predict text and compare the result with the Model generate in Logic Tensor Network.

1.4 Beneficiaries and Project Impact

The purpose of this project is to contribute to the development of Neural tatistical Language Models under the supervision of Dr Weyde using the Logic Tensor Network (LTN) framework developed by the Research Centre of Machine Learning at City, University of London (Artur d'Avila Garcez and Luciano Serafini 2016). I will benefit from this project as I aim to obtain skills and knowledge in one the most exiting areas of Machine Learning: Natural Language Processing. I believe by contributing to this project, I will contribute to this field. The implementation of a model on the Logic Tensor Network with this new approach, it can contribute to Dr Weyde's research for future projects bringing valuable results, that can help to improve model and develop new research areas for Postgraduate and PhD Students. The final milestone is to evaluate the accuracy of text generation model in automatic translation application.

2. Critical Context

2.1 Statistical Language Models

The purpose of a Statistical Language Model or Language Modelling is to determine a probability distribution for the next word in a given context. The probability is learnt by the occurrence of words in examples of text, often very large corpora.

These days, the use of Neural Networks has become widespread in solving machine learning problems. The use of Neural Networks with Statistical Language Modelling is called Neural Statistical Language Modelling. This approach has shown better results than traditional methods like n-grams or bag-of-words, these methods are relatively simple to train but they show sometimes poor result because of the curse of dimensionality. The curse of dimensionality is a phenomenon that is so common in NLP for higher number of word combinations.

Bengio et al (2003), implemented Neural Language Models (NLM) where the words are converted to a vector to be parameterized, this process is known as word embeddings. These word embeddings are then used in the inputs layer of a neural network. This helps overcome the curse of dimensionality because the word embedding helps to reduce the number of inputs in the Neural Network.

2.2 Syntax in the Language

The syntax in a language can be descried in the form of grammar. The syntax describes the order of the words in a sentence, and how the words are used together. In a sense, syntax is a kind of validator that is the sentences are structured correctly. For example, the English language mostly follows a SVN (Subject, Verb, Object) structure.

I (Subject) + love (Verb) + travelling (Object)

However, this basic structure is not enough to express complex sentences. Some grammar studies Huddleston, R. (1988) specify that the English Language has between 8 and 12 parts. For this project we are going to focus on the next 9 syntactic categories as *Verb, Noun, Adjective, Determiner, Adverb, Pronoun, Preposition*, *Conjunction* and *Interjection* Baker, M. (2003).

Consider the following sentence:

interjection	pron.	conj.	det.	adj.	noun	verb	prep.	noun	adverb
Well,	she	and	my	young	sister	walk	to	park	slowly.

Our main hypothesis is that extracting these rules as Background Knowledge and converting to real logic to feed into the Neural Network using in the Logic Tensor Network framework will bring a more reliable model that is more consistent on dealing with noisy data and good on quality and performance.

To create the model, a set of suitable the grammatical rules needs to be identified, which will be describe the most common sentence structures and will have to be conditional on sentence types. These rules will be implemented in the layers of the Neural Network model to activate the neurons for those phases that are grammatically correct. The purpose is to provide the neural networks the Background Knowledge (BK) rules as a starting point, where there is a prior knowledge of English Language structure and thus less learning from data is required. Our belief is that this implementation would also give better results on noisy data.

One factor to consider is, even where there are grammatical rules for the language, that the natural language often does not commit them. The natural language has a huge variety of terms and combination that introduce ambiguities to the sentences but the language and can be still understood by humans. This will be addressed by conditioning the structures on sentence types, such that the structures would not be applied if the sentence does not have to a grammatical structure.

2.3 Logic Tensor Networks (LTN) and Statistical Relational Learning (SRL)

Statistical Relational Learning (SRL) is a field of Artificial Intelligence (AI) and more particularly a subdiscipline of Machine Learning that deals with relational learning context in complex structures using statistical learning methods. SRL usually uses the first order logic to describe relation properties and tries to describe formalisms in a general manner (Ravkic et al., 2015).

To deal with uncertainty that is prevalent in a real-world problem, logical and relational representation can be expressed by probability theory. This opens the door to the field of statistical relational learning (SRL) (Getoor and Taskar, 2007), that is also known as Probabilistic Logic Learning (De Raedt and Kersting, 2003) that deals with probabilistic approaches to Logical Representation. In Figure 1. We can see how Statistical Relational Learning (SRL) of Natural Language is a research area of Computational Natural Language Learning that emerges at an intersection between Natural Language Processing (NLP) and Machines Learning (ML).

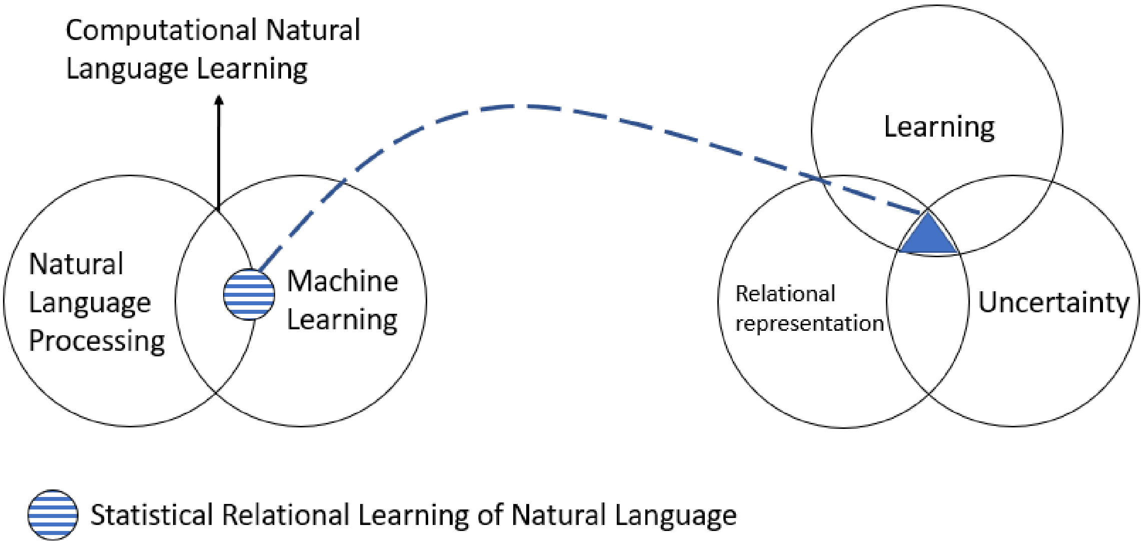
Computational Natural Language Learning
Natural Language Processing
Machine Learning
Learning
Relational representation
Uncertainty
Statistical Relational Learning of Natural Language

Figure. 1 Situating statistical relational learning of natural language (Bianchini, Maggini, 2013)

Logic Tensor Networks (LTN) is a framework that helps build a neural network that must deal with Statistical Relational Learning by applying Real Logic (Serafini and Garcez, 2016). One of the key factors for this project is to find suitable grammatical rules as a Background Knowledge that represent the syntax of the English language to the Neural Network that allow us to implement the Logic Tensor Network to create an model that can predict the words in a phrase based on both the given background knowledge and learning from data.

3. Approach (40%)

3.1 Software and Frameworks

Python 3.5 will be the primary language to develop this project. The main libraries to use will be NumPy, Pandas and Scikit-Learn. For the framework I will use Google TensorFlow in the version 1.7.0. Data pre-processing techniques will be implemented to prepare the data for implementation.

3.2 Data sources and Pre-processing.

Datasets of English texts will be used for the project. The English language has its own structure, which we aim to make use of by applying grammatical rules. We are going to use a popular data set, the Stanford Natural Language Inference (SNLI) Corpus Samuel R. Bowman (2015). The dataset is a collection of five hundred and seventy thousand English sentences. The sentences are labelled in different categories as contradiction, natural and entailment. The grammatical structure sentences are the source for the background knowledge rules. We are going to do the pre-processing with The Natural Language Toolkit (NLTK). The token is going to be taken from these categories. NLTK has a complete documentation (Bird, Steven, 2009) that focuses on tagging and analysing sentence structures, this pre-processing is going to be fundamental to get data ready to feed into deep neural networks.

3.3 Early Investigation

To get more experience on this filed of NPL that can help to complete this project, I plan to complete a set of tutorials related to the construction of NLP and the previous steps that the process demand. I believe this will help me understand the Language Modelling better.

In Language Modelling there are other interesting approaches like Long Short-Term Memory (LSTM) (Hoch Reiter and Schmid Huber, 1997; Gers et al., 1999). One of the application is how the LSTM are applied to infer the word meaning in large pieces of text and how can fix with other techniques like contextual features as a result you can get text generation, word prediction or topic prediction (Le, P., Dymetman, M. & Renders, J. 2016).

I plan also to explore some other more complex techniques. A key factor for a success project is to develop suitable grammatical rules to feed the LTN. NLP is an mature research field and other useful techniques exist to generate BK. Therefore it will be good to document if I find any another interesting approaches that can this project.

3.4 Framework Implementation

Logistic Tensor Network is a framework that use Google's TensorFlow (Abadi 2016). LTN uses tensor networks combined with first order logic that enable the background knowledge. To understand first order logic and the mathematical formulation of Real Logic is crucial to working with the LTN framework. I plan to complete a tutorial in TensorFlow about Neural Statistical Language Models to get more ability on the implementation of the LTN. The grammatical rules will be determined in the pre-processing state using NLTK. These grammatical rules will then the implemented in the LTN. The idea is to integrate the use of LTN with Recurrent Neural Networks (RNN) to improve the results in Language Modelling area.

3.4.1 Model Evaluation

One good technique to evaluate Statistical Language Models is *perplexity* (Martin and Jurafsky, 2014) that is based on the cross-entropy between the data and the prediction. The calculation is made over the probability distribution of words (or phrases). As a result, a good model is one that gives high probabilities score, and thus a low perplexity, on the test data (J. Hockenmaier 2015).

4. Work plan

In this section I describe the work plan of the proposal as shown in the Figure 2 Gantt chart. The project life runs from July 1st until October 1st 2018. I find four milestones until the project dead line. The first milestone is completing the literature, the second one is to prepare the dataset, the third one is to correctly setup environment. The fourth and final milestone is to complete the project documentation and hand over the project. If there are some external or sudden changes, the plan will be updated to show these new changes.

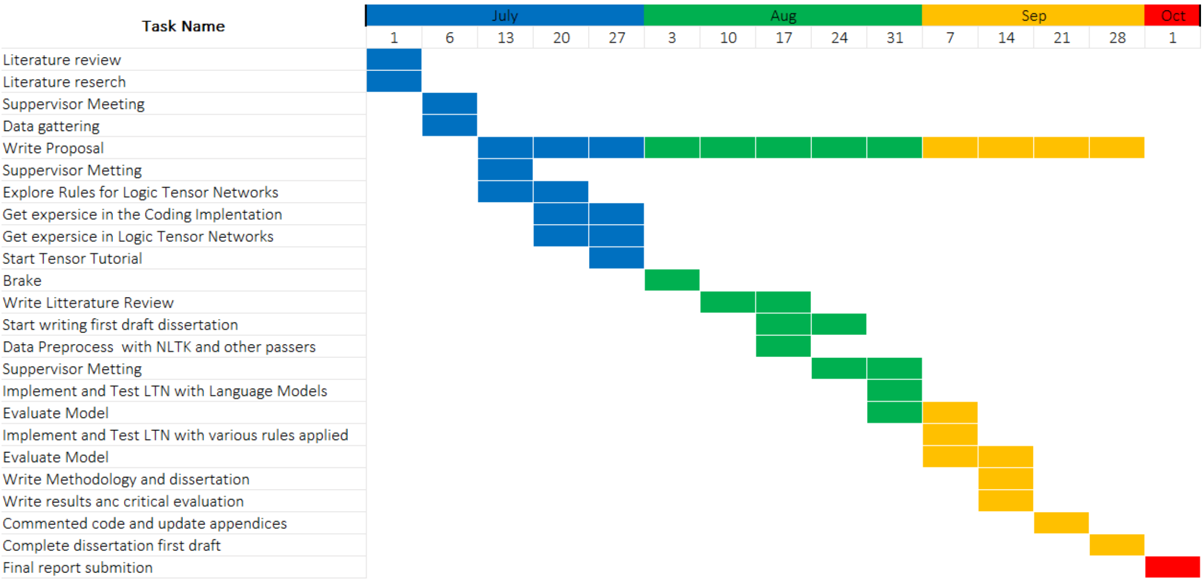

Figure 2 Work Plan Schedule

5. Risks

How to manage and mitigating possible risks that can affect the project is a key component. I found Dawson's (2006) framework for risk management very useful in this context. A list of possible risk that can affect the project success has been observed.

I have a list of possible risks, outlined by the nature. Each risk has assigned a likelihood from 1 to 5, Each risk consequence scale from 1 to 5, I calculate the impact by (likelihood * consequence). For each risk I have mitigating steps to reduce the likelihood that the risk happened, and I have too a contingency plan if the risk eventuates.

#	Description	Likelihood	Consequence	Impact	Mitigation	Contingency
1	Difficulties with coding in TENSORFLOWTM	3	3	6	Spend least 4 weeks on tutorial and complete some mini projects	Ask the supervisor to help overcome this gap.
2	Difficulty in understanding the language model techniques	3	4	12	Select carefully literature that help to understand the models.	Take an extra course that can help to get some practical knowledge
3	LTN does not produce the expected results	3	4	12	In a research project it is possibility. Keep posted the	Document well all the effort, emphasize the

					supervisor of realistic results.	future work for future projects.
4	Tasks taking longer that the first schedule	3	5	15	Carefully check the project plan and the time assigned to each task.	Move task with high uncertainty to fist period of the project.
5	Laptop fails/stolen.	1	5	5	Take extra care of always have backup on the cloud of the last version	Purchase a new Laptop
6	Personal computer cannot process computational task and it takes longer	2	2	4	Set up the VPN to access the servers of the University. And setup an environment for the project.	Use the server that the university give to train the models
7	Loss of interest	1	5	5	The topic was carefully chosen	Talk with the supervisor and check the state of the art of new application to get a extra motivation.

6. Ethics Review Form

This project only involves technical and coding work, most of the task are apply statistical models and machine learning models on the computer of the author. This project does not work on data that can help to identify persons individually.

Ethics Review Form: BSc, MSc and MA Projects
Computer Science Research Ethics Committee (CSREC)

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

Part A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

Part B: Ethics Proportionate Review Form. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in this case. The approval may be provisional: the student may need to seek additional approval from the supervisor as the project progresses.

A.1 If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval.		*Delete as appropriate*
1.	Does your project require approval from the National Research Ethics Service (NRES)? For example, because you are recruiting current NHS patients or staff? If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/.	**No**
2.	Does your project involve participants who are covered by the Mental Capacity Act? If so, you will need approval from an external ethics committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/.	**No**
3.	Does your project involve participants who are currently under the auspices of the Criminal Justice System? For example, but not limited to, people on remand, prisoners and those on probation? If so, you will need approval from the ethics approval system of the National Offender Management Service.	**No**

A.2 If your answer to any of the following questions (4 – 11) is YES, you must apply to the City University Senate Research Ethics Committee (SREC) for approval (unless you are applying to an external ethics committee).		*Delete as appropriate*
4.	Does your project involve participants who are unable to give informed consent? For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf?	**No**
5.	Is there a risk that your project might lead to disclosures from participants concerning their involvement in illegal activities?	**No**
6.	Is there a risk that obscene and or illegal material may need to be accessed for your project (including online content and other material)?	**No**
7.	Does your project involve participants disclosing information about sensitive subjects? For example, but not limited to, health status, sexual behaviour, political behaviour, domestic violence.	**No**

8.	Does your project involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (See http://www.fco.gov.uk/en/)	**No**
9.	Does your project involve physically invasive or intrusive procedures? For example, these may include, but are not limited to, electrical stimulation, heat, cold or bruising.	**No**
10.	Does your project involve animals?	**No**
11.	Does your project involve the administration of drugs, placebos or other substances to study participants?	**No**

A.3 If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
12.	Does your project involve participants who are under the age of 18?	**No**
13.	Does your project involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.	**No**
14.	Does your project involve participants who are recruited because they are staff or students of City University London? For example, students studying on a specific course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your project involve intentional deception of participants?	**No**
16.	Does your project involve participants taking part without their informed consent?	**No**
17.	Does your project pose a risk to participants or other individuals greater than that in normal working life?	**No**
18.	Does your project pose a risk to you, the researcher, greater than that in normal working life?	**No**

A.4 If your answer to the following question (19) is YES and your answer to all questions 1 – 18 is NO, you must complete part B of this form.		
19.	Does your project involve human participants or their identifiable personal data? For example, as interviewees, respondents to a survey or participants in testing.	**No**

7. References:

The Stanford Natural Language Inference (SNLI) Corpus https://nlp.stanford.edu/projects/snli/

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. and Ghemawat, S., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint
arXiv:1603.04467.

Artur d'Avila Garcez and and Luciano Serafini, 2016, Logic Tensor Networks

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

Baker, M. (2003). Lexical Categories: Verbs, Nouns and Adjectives (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511615047

De Raedt and Kristian Kersting. Probabilistic inductive logic programming. In Shai Ben-David, John Case, and Akira Maruoka, editors, ALT, volume 3244 of Lecture Notes in Computer Science, pages 19–36. Springer, 2004. ISBN 3-540-23356-3.

Daniel Jurafsky and James H. Martin, 2000,"Speech and Language Processing"

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation

Getoor and Taskar, 2007, Introduction to Statistical Relational Learning. MIT Press

Kim, Y., Jernite, Y., Sontag, D. and Rush, A.M., 2016, February. Character-Aware Neural Language Models. In AAAI

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Jurafsky, D. and Martin, J.H., 2014. Speech and language processing (Vol. 3). London: Pearson.

Huddleston, R. (1988). English Grammar: An Outline. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139166003
Bird, Steven, Edward Loper and Ewan Klein ,2009, *Natural Language Processing with Python*. O'Reilly Media Inc.

J. Hockenmaier, 2015, CS447: Natural Language Processing

Bianchini, M., Maggini, M., Jain, L.C. & SpringerLink eBook Collection, 2013, Handbook on Neural Information Processing, Springer Berlin Heidelberg, Berlin, Heidelberg.

Le, P., Dymetman, M. & Renders, J. 2016, "LSTM-based Mixture-of-Experts for Knowledge-Aware Dialogues".