# Understanding the Importance of Entities and Roles in Natural Language Inference :

# A Model and Datasets

by

Ishan Shrivastava

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2019 by the
Graduate Supervisory Committee:

Chitta Baral, Chair
Saadat Anwar
Yezhou Yang

ARIZONA STATE UNIVERSITY

August 2019

ABSTRACT

The task of Natural Language Inference (NLI) is to determine the possibility of a sentence referred to as "Hypothesis" being true given that another sentence referred to as "Premise" is true. NLI is a precursor to solving many Natural Language Processing(NLP) tasks such as Question Answering and Semantic Search. Considering the applications of NLI, the importance of having a strong NLI system can't be stressed enough.

While the existing state of the art models do get good accuracy on the test sets of various large-scale datasets they were trained on, they fail to capture the basic understanding of "Entities" and "Roles". They often make the mistake of inferring "John went to the market." from "Peter went to the market." failing to capture the notion of "Entities". These models also fail to capture the notion of"Roles" as they end up wrongly inferring "Peter drove John to the stadium." from "John drove Peter to the stadium." The lack of understanding of "Roles" can be attributed to the lack of such examples in the various existing datasets. The failure in capturing the notion of "Entities" is not just due to the lack of such examples in the existing NLI datasets but also due to the strict use of vector similarity in the "word-to-word" attention mechanism being used in the existing architectures.

To overcome these issues, this work presents two new datasets and a modification to the existing models in the form of a novel attention mechanism. The two new datasets: "NER Changed"(NC) and "Role-Switched"(RS) datasets contain examples that require the understanding of "Entities" and "Roles" respectively to make correct inferences. This work shows how the existing architectures perform poorly on the NC dataset even after being trained on the new datasets. In order to help the existing architectures understand the notion of "Entities", this work proposes a modification to

the "word-to-word" attention mechanism that incorporates the "Symbolic Similarity" by using the Named-Entity features of the Premise and Hypothesis sentences. The new modified architectures not only perform significantly better than the unmodified architectures on the NC dataset but also performs as well on the existing datasets.

# DEDICATION

I would like to dedicate this work to my parents and all those people who never gave up on me.

ACKNOWLEDGMENTS

I would like to thank Dr. Chitta Baral for his constant support, guidance and motivation throuhgout my masters here at ASU. I wouldn't have pursued thesis if not for Dr. Baral's course on Natural Language Processing which help develop my interest in this field. This work would not have been possible if not for the motivation and knowledge I gathered in this course. I am sure this experience will help pave a successful way in all my future endeavors.

I would like to thank my committee members Dr. Saadat Anwar and Dr. Yezhou Yang who offered their guidance and support whenever I needed. Their valuable suggestions and critical comments led me to improve this work significantly.

I would also like to thank Arindam Mitra, my mentor and collaborator, for helping me out in each step of the way throughout my work in this thesis. He is a perfect mentor who patiently guided me and answered all my questions and concerns throughout this thesis. I would also like to thank Arpit Sharma and Aurgho Bhattacharjee for all their encouragement. They were involved in many brain storming sessions with me which led to the betterment of this work.

Last but not the least, I would like to thank my parents who constantly stood by me and encouraged me to never give up throughout my life. No amount of words or actions will ever be enough to do justice to everything they have done for me.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

# Chapter 1

# INTRODUCTION AND MOTIVATION

Computers or machines have seen an exponential growth in terms of their capability to perform various tasks. Humans have continued to develop powerful computer systems that are powerful not just in terms of speed and time but also in terms of their diverse working domains. A continued effort has been made in the field of Artificial General Intelligence. This effort is a result of our curiosity that led humanity to wonder, "Can a machine think and behave like humans?"(Point, n.d.).

Therefore, the primary objectives for an Artificially Intelligent system would be to be able to do anything a human can do. It includes tasks such as planning, recognizing sounds and objects, speaking, translating etc. This leads us to the broader area of Natural Language Processing (NLP) that this work is about. NLP is that area of AI that deals with language (usually written). Natural Language refers to the spoken and written language by people, while NLP refers to the mechanisms employed in order to extract information from the spoken and written words.

Natural Language Processing encompasses two broader areas, 1) *Natural Language Generation* that deals with empowering systems to formulate phrases that humans might generate and 2) *Natural Language Understanding (NLU)* that deals with ensuring that the system understands the phrases of natural language, what the words in the phrase mean and their intent. In this work, we go deeper into *Natural Language Understanding (NLU).*

NLU interprets the meaning and proper intent of a text. In other words, it is an area that aims at enabling and ensuring whether a system can comprehend

natural language or not. One of the ways of testing whether a system understands or comprehends the natural language is by asking the system questions and evaluating on the answers the system gives. Therefore, a system that understands and comprehends natural language would be a Question-Answering system.

A Question-Answering system would be required to answer a question given a passage. This problem of Question-Answering can therefore be broken down to finding a relation between two pieces of text. One of the two pieces of text can be a statement combining the question and an answer, while the other piece of text could be the given passage. If the first piece of text (a statement combining the question and an answer) can be inferred from the given passage then the system can learn to figure out the correct answer. This is referred to as Natural Language Inference (NLI).

Natural Language Inference also referred to as NLI is an important component and a precursor to solving many Natural Language Processing tasks like Question Answering, Text Summarization, Relation Extraction etc. Having a strong NLI system leads to a system that understands natural language well and therefore brings Humanity closer to building an Artificially Intelligent System. In this work, we see how existing NLI architectures work and what datasets are used to train such systems. We see what are the drawbacks in these architectures and the datasets. Finally this work makes two contribution in the form of two datasets and a new approach for building a better NLI system.

## 1.1 What is Natural Language Inference

The problem of determining an *Entailment*, *Contradiction* or *Neutral* relationship between two sentences is known as Natural Language Inference. The task is to identify

the relationship between a sentence referred to as "Premise" and a sentence referred to as the "Hypothesis". An *Entailment* relation suggests that if the premise is true, then the hypothesis must be true as well. A *Contradiction* relation on the other hand suggests that if the premise if true, then the hypothesis can not be true. A *Neutral* relation suggests that the premise does not provide enough evidence against or for the hypothesis. This means that the hypothesis can be both true and false. In other words, finding out whether the meaning of the hypothesis is entailed by the meaning of premise or not. An example of each of the three cases is shown in Table 1.

| |
|---|
| **premise:** A soccer game with multiple males playing.<br>**hypothesis:** Some men are playing a sport.<br>**label:** Entailment. |
| **premise:** A black race car starts up in front of a crowd of people.<br>**hypothesis:** A man is driving down a lonely road.<br>**label:** Contradiction. |
| **premise:** A smiling costumed woman is holding an umbrella.<br>**hypothesis:** A happy woman in a fairy costume holds an umbrella.<br>**label:** Neutral. |

Table 1. Example premise-hypothesis pairs from SNLI dataset with human-annotated labels.

The field of NLI has seen the release of many large scale datasets that have advanced this research by helping in the creation of various Deep Neural Network architectures. Among these datasets, the Stanford Natural Language Inference (SNLI) (Bowman et al. 2015a) corpus stands out with 570k premise-hypothesis pairs. Release of such a large dataset paved way for the creation of many entailment systems (Parikh et al. 2016), (Chen et al. 2016)) with different deep learning architectures.

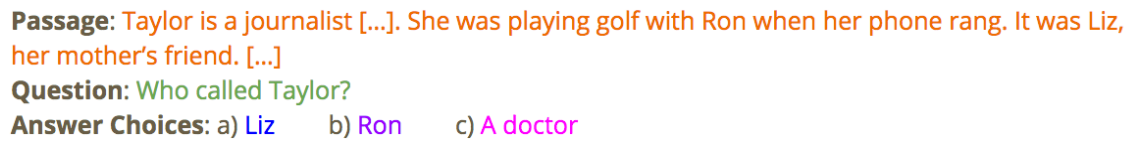



Figure 1. This figure shows an example of a Question Answering problem

## 1.2 Importance of Natural Language Inference

Natural Language Inference is a precursor to various Natural Language Processing tasks. This section describes in what ways different NLP tasks transform the inference needs of an individual application in terms of Natural Language Inference and then use it to solve and improve the end-application performance. This shows how having an effective NLI system is important and essential to solving various Natural Language Tasks.

### 1.2.1 Question Answering

In Question Answering, NLI can be employed to validate or re-rank candidate answers. The main idea is that a candidate answer should be considered correct if and only if the corresponding hypothesized answer statement is entailed by the passage for the question. Figure 1 shows an example of a Question Answering problem. A passage along with a question about it is given. There are three answering choices given as well. The task is to identify the correct answer based on the passage.

The Question Answering problem is transformed in to a Natural Language Inference problem. The passage in the Question Answering problem is considered as the premise. The question and the different answer choices are combined to form declarative statements. The respective declarative statements are considered as differ-


Figure 2. This figure shows the premise and different hypothesis generated for a Question Answering problem

ent hypothesis. Figure 2 shows the premise and different hypothesis for the example mentioned above. A NLI system is then used to find the "Entailment" confidence scores for each *Premise-Hypothesis* pair. The answer choice corresponding to the "hypothesis" that gives the highest entailment confidence score is predicted as the correct answer.

### 1.2.2 Text Summarization

Text Summarization aims at developing a technique to generate concise summary of large volume of texts while stressing on sections that contain useful information without missing out on the overall meaning of the text. In text summarization, Natural Language Inference is used for different types of inference. (Harabagiu, Hickl, and Lacatusu 2007) used their NLI system to compare between six different text summarization techniques by selecting the best summary among the six candidate summaries generated. They used the entailment scores in the following way. First, they used the entailment confidence scores between all pairs of sentences appearing in distinct summaries. This was done to find out their semantic overlap. These semantic overlap entailment scores were used to assess how well the summary sentence captures the semantics of the text to eventually provide each summary an individual score.

Their evaluation showed that using an entailment based summary selection method

| |
|---|
| **premise:** John went to the kitchen.<br>**hypothesis:** Peter went to the kitchen. |
| **premise:** John lent Peter a bicycle.<br>**hypothesis:** Peter lent John a bicycle. |

Table 2. Sample premise-hypothesis pairs where existing models trained on SNLI suffers significantly.

provided the most responsive summary in 86% of the cases. This was important because it was observed that different summarization strategies often gave similar quality summaries while the overall best strategy produced the most responsive summary in only 35% of the cases. Therefor, the author concluded by considering this as an encouraging result, as it showed how an effective Natural Language Inference system is necessary to correctly distinguish even among the similarly responsive summaries.

## 1.3 Motivation for this Research Thesis

Various deep learning architectures have been built in order to understand and solve for Natural Language Inference. While these systems do get good accuracy on the SNLI test data set, there are some inherent drawbacks these systems suffer.

These system lack the understanding of "Entities" and "Roles" . They suffer greatly while trying to solve for cases where the only difference between a premise and a hypothesis sentence is that they refer to different entities. For example, consider the first example in Table 2. Both premise and hypothesis talk about a "Person" entity going to the kitchen, but the premise refers to "John" while hypothesis refers to "Peter". The ESIM (Chen et al. 2016) Model fails to capture this difference and predicts "Entailment" with a confidence of 86.98%. Table 3 shows examples of wrong

predictions for the pairs of sentences where the premise and hypothesis refer to different entities.

These systems are also unable to differentiate between the "Roles" played by an entity in different sentences. For example, consider the second example in Table 2. Both premise and hypothesis talk about the action of "lending" a bicycle, but they differ in terms of who plays the "Role" of an "Agent" and a "Recipient". In premise "John" is the "Agent" and "Peter" is the "Recipient", while these "Roles" of reversed in case of the hypothesis. This makes the true label for this example as "Contradiction", but the ESIM model fails to grasp this subtlety and predicts "Entailment" with a confidence of 96.58%. Table 4 shows examples of wrong predictions for the pairs of sentences where the premise and hypothesis sentences have entities with their roles reversed.

There are two main reasons that the current state of the art systems fail to capture these subtleties of "Entities" and "Roles":

1. The famous large scale datasets like SNLI and MNLI, although cover many premis-hypothesis pairs with different complexities, they do not contain examples that require the understanding of "Entities" and "Roles" to solve for inference.
2. The secondary reason for such a behaviour is the use of vector similarity in the word-to-word attention mechanism employed by the various top performing architectures.

| Differing Entity Type | Premise-Hypothesis Pair | ESIM Entailment Confidence Scores |
|---|---|---|
| Name of a Person | *premise*: **Gary Johnson** stated the technical possibility of concluding the next phase of the agreement by December 2007.<br>*hypothesis*:**Steve Waugh** stated the technical possibility of concluding the next phase of the agreement by December 2007. | 99.76% |
| Name of a City | *premise*: **Hong Kong** stands to benefit more than most from continued global trade liberalisation as trade is the engine of its growth, accounting for nearly three times its gross domestic product.<br>*hypothesis*: **Melbourne** stands to benefit more than most from continued global trade liberalisation as trade is the engine of its growth, accounting for nearly three times its gross domestic product. | 99.99% |
| Date | *premise*: He was first appointed to the nine-member court by President Nixon in **2002**.<br>*hypothesis*: He was first appointed to the nine-member court by President Nixon in **1976**. | 91.0% |
| Date | *premise*: In response to these challenges, King Mohammed in **1928** launched a National Initiative for Human Development, a $ 2 billion program aimed at alleviating poverty and underdevelopment by expanding electricity to rural areas and replacing urban slums with public and subsidized housing, among other policies.<br>*hypothesis*: In response to these challenges, King Mohammed in **1995** launched a National Initiative for Human Development, a $ 2 billion program aimed at alleviating poverty and underdevelopment by expanding electricity to rural areas and replacing urban slums with public and subsidized housing, among other policies.. | 99.99% |
| Cardinal in Numeric | *premise*: Bangladeshi officials say worst hit was the southeastern port city of Chittagong, where **16651** people died after many hillside homes were swept away or collapsed under tons of mud.<br>*hypothesis*: Bangladeshi officials say worst hit was the southeastern port city of Chittagong, where **6948** people died after many hillside homes were swept away or collapsed under tons of mud. | 61.52% |
| Cardinal in Word | *premise*: **Twelve** of those injured later died at a hospital.<br>*hypothesis*: **Seventeen** of those injured later died at a hospital. | 64.64% |

Table 3. Sample premise-hypothesis pairs with different named entity where ESIM model gives wrong predictions

| Premise-Hypothesis Pair | ESIM Entailment Confidence Scores |
|---|---|
| *premise*: **Quinn** lost the most important match of his life to **Pat**.<br>*hypothesis*:**Pat** lost the most important match of his life to **Quinn**. | 48.55% |
| *premise*: **Quinn** hangs **Frankie** rendering him dead.<br>*hypothesis*: **Frankie** hangs **Quinn** rendering him dead. | 60.79% |
| *premise*: **Current Prime Minister Stephen Harper** called **Gray** "an honourable parliamentarian who served his country well".<br>*hypothesis*: **Gray** called **current Prime Minister Stephen Harper** "an honourable parliamentarian who served his country well".. | 54.44% |
| *premise*: **India** rejects the accusation, and calls for **Pakistan** to prosecute militants based there for the 2008 Mumbai attacks, which killed over 150 people..<br>*hypothesis*: **Pakistan** rejects the accusation, and calls for **India** to prosecute militants based there for the 2008 Mumbai attacks, which killed over 150 people. | 56.41% |

Table 4. Sample premise-hypothesis pairs with entities playing different roles where ESIM model gives wrong predictions

### 1.4 Contribution

In order to overcome the challenges mentioned in the previous section, we need a NLI system that captures the notion of "Entities" and "Roles". To make such a NLI system this work's contribution are two folds:

1. This work's contribution includes two new datasets, "NER Changed"(NC) dataset and the "Role-Switched"(RS) dataset. This work shows how existing annotated corpora like bAbI (Weston et al. 2015), AMR (Banarescu et al. 2013), CONLL 2003 (Ratinov and Roth 2009), CONLL 2004 (Carreras and Màrquez 2005), Verbnet (Schuler 2005), PropBank (Palmer, Gildea, and Kingsbury 2005) and QA-SRL (FitzGerald et al. 2018) can be used to automatically create pairs of premise-hypothesis that emphasize on capturing the notion of "Entities" and

"Roles". The two new datasets aim at overcoming the issue mentioned in the first reason by creating pairs of premise and hypothesis that will require the understanding of "Entities" and "Roles" in order to be solved by a NLI system. This work shows how one of the existing architectures (Chen et al. 2016) can understand "Roles" given such examples at train time.

2. Secondly, this work proposes a modification to the existing architectures by introducing a new attention mechanism. The novel attention mechanism overcomes the second reason mentioned above by not only relying on the vector similarity, but by combining it with symbolic similarity. The modification enables the NLI system to capture the notion of "Entities" by making it learn to weigh the attention weights between vector similarity and symbolic similarity. This work shows how the resulting new architectures perform significantly better than the unmodified architectures on the two new datasets, while maintaining the performance on existing datasets.

## 1.5 Structure of this Thesis

In this thesis, rest of the chapters are structured in the following way.

- **Chapter 2**: In this chapter we discuss the datasets, the existing approach for NLI and the issues that arise because of the existing approach.
- **Chapter 3**: In this chapter we discuss how existing datasets and resources are used in this work to create two new labelled NLI datasets. The two datasets are created in order to overcome the lack of examples in the existing NLI datasets that emphasize on capturing the notion of "Entities" and "Roles".
- **Chapter 4**: In this chapter we first formalize the two existing architectures for

NLI: Decomposable Attention Model and Enhanced Sequential Inference Model. We then formalize and propose the novel attention mechanism which enables these architecture to capture the notion of "Entities".

- **Chapter 5**: In this chapter we describe how the proposed architectures are explainable and can be analyzed to understand the decisions made by the proposed attention mechanism. We also discuss the experimental setup employed in this work and finally analyse and see the results for all the experiments.
- **Chapter 6**: In this chapter we finally conclude and detail the areas that can be worked at to improve or enhance the datasets generated, the proposed attention mechanism and experimental analysis done in this work.

# Chapter 2

# RELATED WORKS AND BACKGROUND

In this chapter, we will discuss about the existing datasets and models along with their issues in detail. Section 2.1 will describe about the different datasets that have been created in order to advance the research in NLI. We also discuss about the issues that arise due to the data creation methodology that is employed while creating these datasets. In Section 2.2 we detail the widely adapted approach to NLI which is referred to as the Align and Compare approach. In Section 2.3 and 2.4 we describe two of the models for NLI that have adopted this approach: Decomposable Attention Model (DecAtt) and Enhanced Sequential Inference Model. Section 2.5 explains why the Align and Compare Approach fails to capture the notion of "Entities" and what could be done to overcome this issue.

## 2.1 Important Datasets in NLI

Many large labelled inference data sets have been released so far. (Bowman et al. 2015b) is the first one to address the problem of creating a large amount of labelled inference data. They use crowd sourcing to create a high agreement entailment data set known as Stanford Natural Language Inference (SNLI) (Bowman et al. 2015b) data set. They use the image captions as premise and ask the human workers to formulate three hypothesis for each of the "Entailment", "Contradiction" and "Neutral" scenarios.

Later MultiNLI (Williams, Nangia, and Bowman 2017) was created which covers

multiple genres. SciTail (Khot, Sabharwal, and Clark 2018) Dataset and QNLI Dataset (Demszky, Guu, and Liang 2018) are generated directly from the end task of question-answering. PAWS (Zhang, Baldridge, and He 2019) is also a very recent paraphrase identification dataset.

The release of such large data sets serves as a motivation to create many advanced deep learning architectures (Bowman et al. 2016), (Vendrov et al. 2015), (Mou et al. 2015), (Y. Liu et al. 2016), (Munkhdalai and Yu 2016a), (Rocktäschel et al. 2015), (Wang and Jiang 2015), (Cheng, Dong, and Lapata 2016) (Parikh et al. 2016), (Munkhdalai and Yu 2016b), (Sha et al. 2016), (Paria et al. 2016), (Chen et al. 2016), (Khot, Sabharwal, and Clark 2018), (Devlin et al. 2018), (X. Liu et al. 2019). Among these, the most relevant to this work are the Decomposable Attention(DecAtt) Model (Parikh et al. 2016), Enhanced Sequential Inference Model(ESIM) (Chen et al. 2016) and BERT (Devlin et al. 2018). (Parikh et al. 2016) proposed a very simple model that decomposes the inference problem into sub-problems to be solved separately, while (Chen et al. 2016) explored the use of sequential LSTM-based encoding along with attention to outperform the previous results. Details of the DecAtt and the ESIM models are described in the subsequent sections.

Although these deep learning models perform extremely well on SNLI (Bowman et al. 2015b) and MultiNLI (Williams, Nangia, and Bowman 2017), these models fail when encountering simple adversarial examples. (Kang et al. 2018) shows how the Decomposable Attention Model fails when tested on examples with simple linguistic variations such as negation or re-ordering of words. This work tries to make the NLI system immune to re-ordering of words by training it for such adversarial scenarios. (Gururangan et al. 2018) points to the reasons for such failures. It shows that the bias created as a result of crowd sourcing can be attributed to these failures. They observe

that crowd sourcing results in the creation of certain patterns in the hypothesis which can be exploited by a classifier to easily learn to classify correctly without any help from the premise.

## 2.2 Align and Compare : A General Approach to NLI

A general approach to NLI has been "Align and Compare". This approach has been followed and implemented in various ways in the many existing architectures. In this approach, each token or phrase in Hypothesis is looked at to determine whether or not they are similar with each token or phrase in Premise. In the next step, an alignment between Premise and Hypothesis is selected. This is done by choosing the best corresponding element in Premise for each token in Hypothesis. This way many alignment pairs are created. Based on how many and how strong alignments are generated between the Premise and Hypothesis the classification decision is made.

Consider the Figure 3 which shows an example of different alignments for the premise and hypothesis sentences. One of the alignment pairs is "flute solo" and "music". "flute solo" is a sub-phrase in the hypothesis sentence that aligns with the "music" in the premise sentence. Similarly "Alice" aligns with "someone", "park" aligns with "outside" and "plays" aligns with "playing".

### 2.2.1 Word-to-Word Attention and Softly Aligned Sub-phrase calculation

Among all the major NLI deep learning architectures like DecAtt and ESIM, the Align and Compare approach is implemented using Word-to-Word Attention. The Word-to-Word Attention mechanism leads to the calculation of vectors that represent

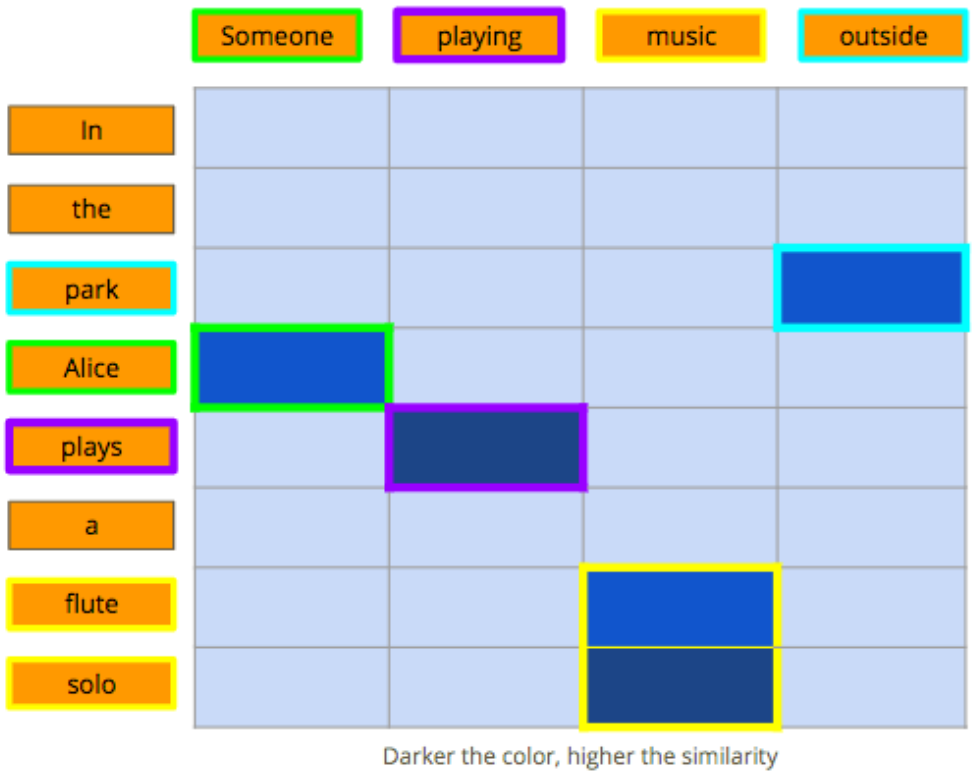



Figure 3. Example of different alignment pairs

Softly Aligned Sub-phrases from the other sentence. Figure 4 shows the attention matrix that is calculated in the Word-to-Word Attention mechanism. It also shows how the Softly Aligned Sub-phrases are calculated.

The Word-to-Word Attention results in an Attention matrix as shown in Figure 4. The attention matrix contains the attention weights for each pair of token in the Premise and Hypothesis sentences. The attention weights are calculated as the Dot Product similarity between the vectors for each pair of token in the Premise and Hypothesis sentences.

The next step is the calculation of Softly Aligned Sub-phrase vectors. These vectors represent a sub-phrase from Premise/Hypothesis that softly aligns (similar/related) to a token or sub-phrase in Hypothesis/Premise. This is done by taking a weighted average

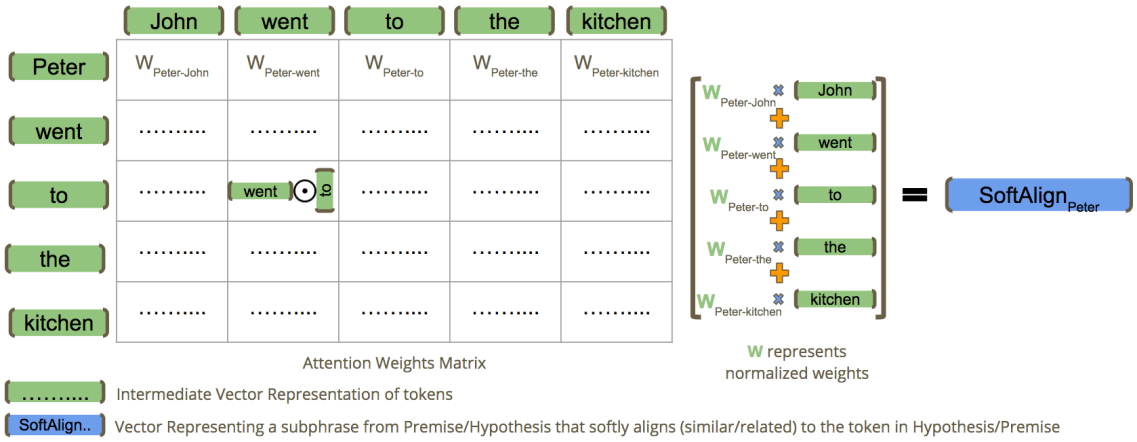



Figure 4. Example of different alignment pairs

of all vectors for the tokens in one sentence weighted by each of their normalized attention weights with the token from the second sentence. Figure 4 shows the calculation of a vector that represents a sub-phrase from the sentence "John went to the kitchen" that softly aligns with the word "Peter" from the sentence "Peter went to the kitchen".

## 2.3 Decomposable Attention Model

The Decomposable Attention(DecAtt) Model consists of three components mainly. These components are referred to as Attend, Compare and Aggregate. The first step is to transform the initial vector embeddings (shown in "Orange" in the Figure 5) into an intermediate vector representation (shown in "Green" in the Figure 5). The next step is the Attend component. Here the soft-aligning of the two sentences is done using a variant of neural attention known as "Word-To-Word" attention as described in the previous section. This decomposes the problem into sub-problem which requires just the comparison of aligned sub phrases which becomes the other major component of the model known as Compare. This comparison produces sets of

vectors for each sentence which contains the comparison information about each word in the sentence with the softly aligned sub-phrase in the other sentence as shown in Figure 5 in "Yellow". The next component of the model is called Aggregate and as the name suggests, this is where the set of vectors produced in the previous component are aggregated to get the final inference vector as shown in step 4 in Figure 5. The final inference vector is the input to a final feed-forward neural network which is used to make the final prediction as shown in step 5 in Figure 5.

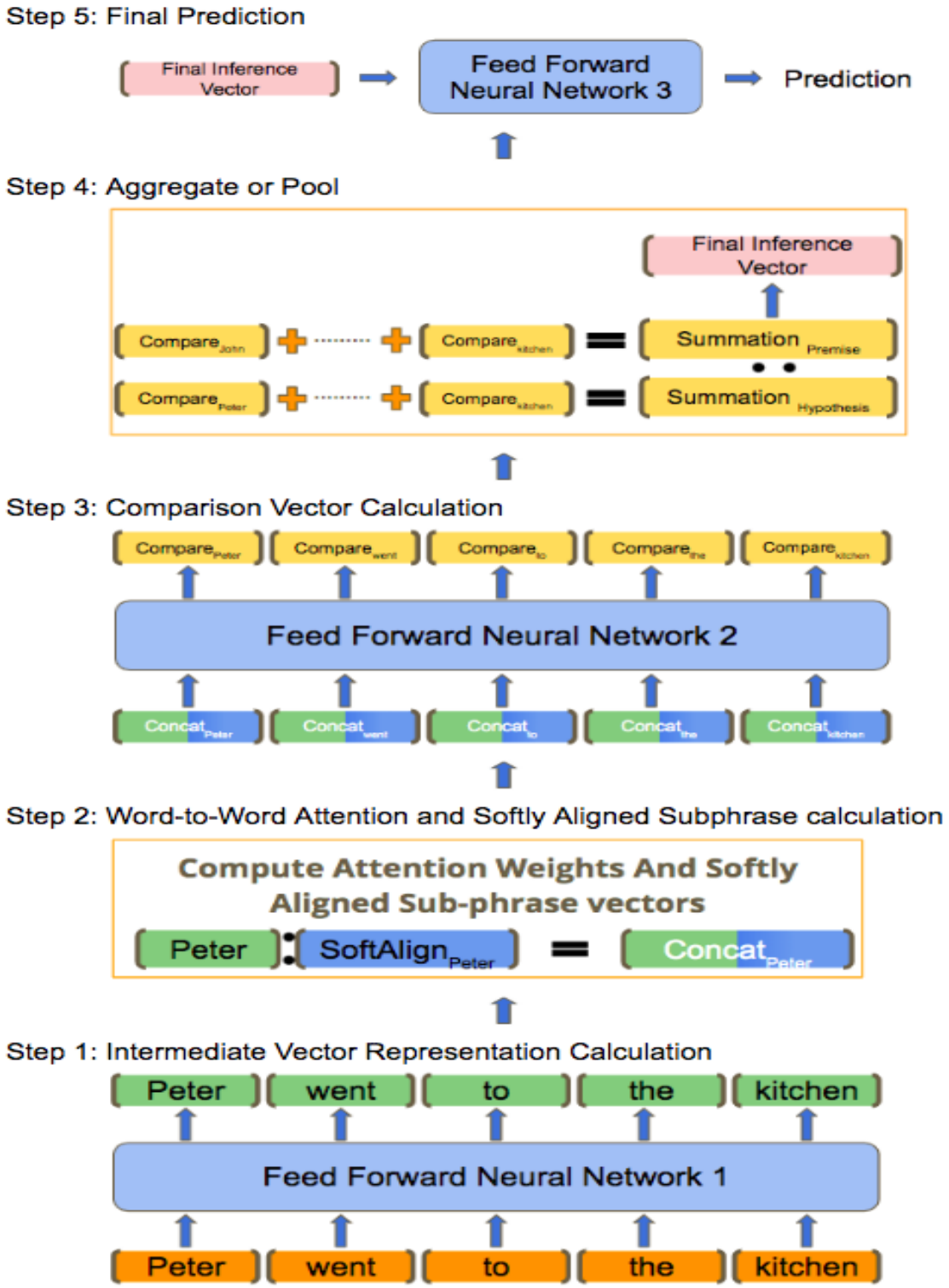



Figure 5. Flow of Decomposable Attention Model

## 2.4 Enhanced Sequential Inference Model

Similar to the Decomposable Attention(DecAtt) Model, the Enhanced Sequential Inference Model(ESIM) also consists of three components. These components are referred to as Locality of Inference, Inference Composition and Pooling which are analogous to Attend, Compare and Aggregate components respectively from the DecAtt Model.

The first step is to transform the initial vector embeddings (shown in "Orange" in the Figure 6 into an intermediate vector representation (shown in "Green" in the Figure 6. Unlike DecAtt model, in ESIM a Bidirectional LSTM is used instead of a feed-forward neural network for this step. Next the Locality of Inference computes the softly aligned subphrases using the BiLSTM encodings of the input Premise and Hypothesis embeddings. In the next step, the local inference information collected in the previous step is enhanced by computing their element wise difference and products. This enhanced representation is then concatenated with the BiLSTM encodings and the subphrases which result in two sets of vectors capturing a high-order interaction between the Premise and Hypothesis sentence as shown in step 2 in Figure 6. These vectors are then used to compute max and average pooling to get the final inference vector as shown in step 4 in Figure 6. The final inference vector is the input to a final feed-forward neural network which is used to make the final prediction as shown in step 5 in Figure 6

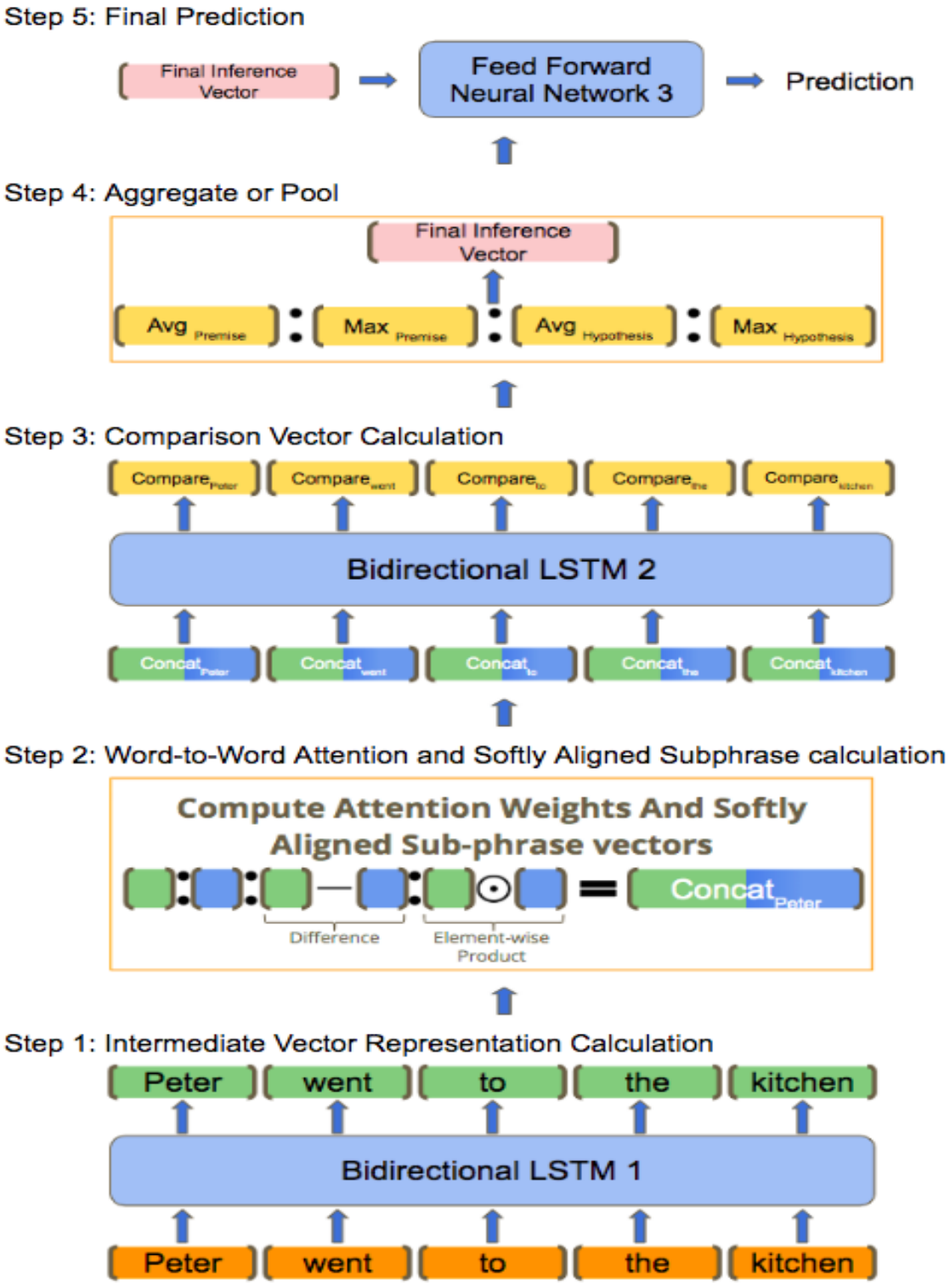



Figure 6. Flow of Enhanced Sequential Inference Model

## 2.5 Issues in the Align and Compare Approach

As described in the previous sections, the Word-to-Word attention mechanism provides a way of finding out the alignment pairs between the Premise and the Hypothesis sentences. Consider the following *Premise-Hypothesis* pair:

| |
|---|
| **Premise** : Peter went to the kitchen |
| **Hypothesis** : John went to the kitchen. |
| **Gold Label**: contradiction |

In this example, apart from the two named entities "Peter" and "John", rest of the sentence is exactly the same. As mentioned in the previous section the attention weights are computed as the vector similarity of two tokens. As expected this will result in very high similarity weights for "went"-"went", "to"-"to" and "kitchen"-"kitchen". This is shown in Figure 7 that shows the attention weights for the example shown above.

If you see this figure you will notice that the attention weights for "Peter"-"John" is also high. The reason for this is that since both "Peter" and "John" are named entities of type "Name of a Person" they are used in a similar context. All the word vectors that are used these days are dependent on the context in which a word is used. Since a "Name of a Person" would be used in similar context, their word vectors are also highly similar.

Having a high attention weight for "Peter"-"John" strongly aligns "Peter" with "John". This would not have happened if the attention weight for "Peter"-"John" pair is a small value. The resulting high value comes due to the strict use of Vector Similarity. In this case, if Symbolic Similarity would have been used, a weight of 0 would have been assigned to "Peter"-"John" pair. In the chapters moving forward, we will see how

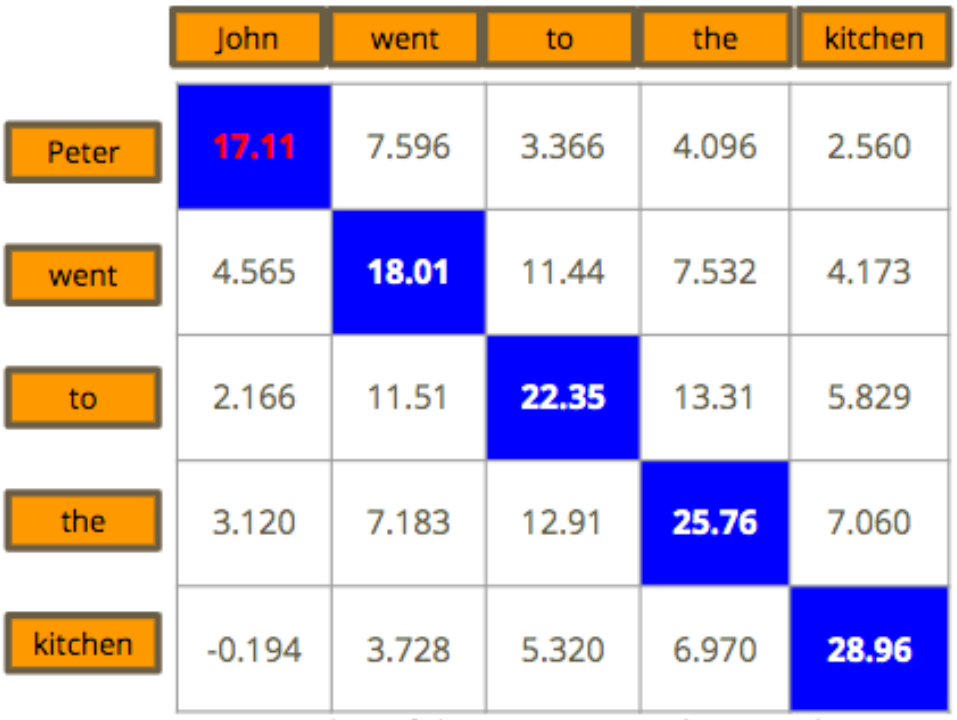



Figure 7. Attention weights of the existing word-to-word attention mechanism (ESIM)

we can modify this attention mechanism to incorporate both Vector Similarity and Symbolic similarity and learn to compute the attention weights as a weighted average of the two.

# Chapter 3

# DATASET GENERATION

In the previous chapters, we have discussed about the importance of having an effective NLI system. We have seen the drawbacks that the current NLI systems suffer from. We saw that these systems lacked the understanding of "Entities" and "Roles". There were two main reasons for these drawbacks. One of the main reasons is the lack of adversarial examples in the existing benchmarks like SNLI and MNLI datasets. Due to the lack of such examples, the existing models are unable to learn to infer the change in "Entities" and "Roles". This leads to the need of a dataset that contains labelled *Premise-Hypothesis* pairs that could aid the models in learning to capture the notion of "Entities" and Roles" in order to solve for inference. In this chapter, we will see how to create two new datasets that emphasize the importance of "Entities" and "Roles. We will see how the two new datasets are generated using existing datasets and resources.

This work introduces two different kinds of datasets which emphasizes on capturing the notion of "Entities" and "Roles. The first one is referred to as "NER Changed" dataset in this piece of work. It includes examples of *contradiction* labelled *Premise-Hypothesis* pairs with a difference of a named entity like "Name", "Number" or "Dates" between the premise and hypothesis sentences. The entities mentioned are replaced by disjoint set of different named entities. The second kind of dataset is referred to as the "Roles-Switched" dataset. It also includes examples of *contradiction* labelled *Premise-Hypothesis* pairs by reversing/switching the roles between the Premise and Hypothesis sentences. To create *entailment* labelled *Premise-Hypothesis* pairs for the

two datasets, pairs of exact same sentences where Premise is equal to the Hypothesis are created.

The two datasets does not contain any *neutral* labelled *Premise-Hypothesis* pairs because this work follows the assumptions made for the creation of the SNLI dataset.

## 3.1 NER Changed Dataset

Among the two datasets that are generated in this work, the first one is refered to as the "NER Changed" dataset. This dataset contains examples of premise-hypothesis pairs that require the model to understand the notion of entities in order to solve for inference. This section describes how existing annotated corpora is used to create the "NER Changed" dataset. This work used the following four different fully annotated corpora:

- **bAbI** (Weston et al. 2015)
- **AMR** (Banarescu et al. 2013)
- **CoNLL 2003** (Ratinov and Roth 2009)
- **GMB(Groningen Meaning Bank)** (Bos et al. 2017)

All of these corpora consisted of annotations which helped in identifying different kinds of entities like "Name" of a person, city, country or location, "Numbers" and "Dates". In general the following steps are followed for all the annotated corpora considered in this work:

1. "Candidate" sentences are identified from the corpus. The candidate sentences are those that contain one or more entities mentioned above.

2. Once the candidate sentences are identified, with the help of annotations we replace the "Entities" in consideration with a placeholder token. The placeholder tokens are *PersonX*, *PersonY* for "Name" of a person, *CityX*, *CityY* for "Name" of a city, *CountryX*, *CountryY* for "Name" of a country, *NumberX* for "Number" entities and *DateX* for "Date" entities.
3. Using such placeholder tokens, many unique template sentences are generated.
4. For each such template sentence, a list of unique replacement options are used to replace the placeholder tokens to create various *contradiction* labelled *Premise-Hypothesis* pairs.
5. *entailment* labelled *Premise-Hypothesis* pairs are created where the hypothesis is exactly equal to the premise.

| | |
|---|---|
| **Train Names:** | 'Casey', 'Riley', 'Jessie', 'Jackie', 'Avery', 'Jaime', 'Peyton', 'Kerry', 'Jody' |
| **Dev Names:** | 'Kendall', 'Peyton', 'Skyler' |
| **Test Names:** | 'Frankie', 'Pat', 'Quinn' |

Table 5. List of 15 Gender Neutral names split for Train, Dev and Test sets

### 3.1.1 Details of creation of Dataset using bAbI Corpus

The bAbI dataset set is a result of the bAbI project of Facebook AI research. The goal of this project and the dataset is to promote research in the area of automatic text understanding and reasoning. This dataset contained a small list of names shown below, that have been used repeatedly throughout the corpus.

| |
|---|
| 'Mary', 'John', 'Daniel', 'Sandra', 'Bill', 'Fred', 'Julie', 'Yann', 'Antoine', 'Jason', 'Sumit', 'Antoine', 'Jeff' |

A simple keyword search for these names resulted in a total 30814 single named "candidate" sentences and 4770 double named "candidate" sentences. A singe named "candidate" sentence refers to a sentence where these names appear only once throughout the sentence, while a double named sentence refers to a sentence where two of these names appear.

For all the single named "candidate" sentences, the names were replaced with the placeholder token **PersonX**. This resulted in a total of 398 unique single named template sentences. An example of a single named "candidate" sentence and the resulting template sentence is shown below.

| |
|---|
| **Single named "candidate" sentence** : "Mary moved to the hallway." |
| **Template sentence** : "**personX** moved to the hallway." |

After creating such single named template sentences, the list of 15 gender neutral names mentioned in Table 5 is used to replace the placeholder token **PersonX** in all the template sentences. These placeholder token replaced template sentences are used to create the *Premise-Hypothesis* pairs. The *Premise-Hypothesis* pairs with the same sentences but different names for the placeholder tokens **PersonX** are labelled as *contradiction*, while the *Premise-Hypothesis* pairs with same sentences and same names for the placeholder token **PersonX** are labelled as *entailment*. For the template sentence mentioned above, one of the *contradiction* and *entailment* labelled *Premise-Hypothesis* pairs are shown below.

| | |
|---|---|
| **Premise** : "**Kendall** moved to the hallway." | **Premise** : "**Kendall** moved to the hallway. " |
| **Hypothesis** : "**Peyton** moved to the hallway." | **Hypothesis** : "**Kendall** moved to the hallway." |
| **Gold Label**: *contradiction* | **Gold Label**: *entailment* |

The double named "candidate" sentences without 'and' keyword immediately in

between the two names were filtered out. Such sentences could have been where the two names play different roles. Such cases are included in the "Role-Switched" dataset. After applying this filter, the two names were replaced with two different placeholder tokens **PersonX** and **PersonX**. This way 30 unique double named template sentences were created. All of these templates resulted in "entailment" labelled *Premise-Hypothesis* pairs.

This way the bAbI dataset is used to create a total of 29166 automatically labelled *Premise-Hypothesis* pairs. The set of replacement names, in this case the set of gender neutral names and the set of template sentences are kept disjoint for the train, test and dev split.

### 3.1.2 Details of creation of Dataset using AMR Corpus

The Abstract Meaning Representation (AMR) bank consists of a set of English sentences. For each of these sentences a simple, readable semantic representation is also provided. An example of such sentence and its representation pair is shown in Figure 8. This dataset is manually constructed by human annotators.

Contrary to bABi corpus, the AMR corpus has sentences that are more complex. This provides our dataset some variety in terms of the complexity of sentences being included. The AMR corpus has annotations describing "Names" of persons, city and countries. An example of the annotation describing the name of a city is shown in Figure 8. The "candidate" sentences in this case are the ones that contain at least one mention of a person, city or country. The AMR annotations aid in getting such "candidate" sentences. Replacing the mentions of the three kinds of named entities, person, city and country with their placeholder tokens **PersonX**, **CityX**,**CountryX**

```
# ::id PROXY_AFP_ENG_20071219_0023.29 ::date 2013-06-30T22:31:51 ::snt-type body ::annotator SDL-
AMR-09 ::preferred
# ::snt Teheran defied international pressure by announcing plans to produce more fuel for its nuclear program.
# ::save-date Mon Jul 8, 2013 ::file PROXY_AFP_ENG_20071219_0023_29.txt
(d / defy-01
    :ARG0 (c / city :name (n / name :op1 "Teheran"))
    :ARG1 (p / pressure-01
        :ARG1 c
        :mod (i / international))
    :instrument (a / announce-01
        :ARG0 c
        :ARG1 (p2 / plan-01
            :ARG0 c
            :ARG1 (p3 / produce-01
                :ARG0 c
                :ARG1 (f / fuel
                    :quant (m / more))
                :ARG3 (p4 / program
                    :mod (n2 / nucleus)
                    :poss c)))))
```

Figure 8. Annotations provided in the AMR corpus

respectively provides 945 unique template sentences. An example of a "candidate" sentence and the corresponding template sentence is shown below.

**A "candidate" sentence** : "Teheran defied international pressure by announcing plans to produce more fuel for its nuclear program."

**Template sentence** : "**CityX** defied international pressure by announcing plans to produce more fuel for its nuclear program."

A list of names of persons, city and country extracted from the AMR corpus is used to replace the respective named entity in the template sentences to automatically create *contradiction* labelled *Premise-Hypothesis* pairs. Example of a *contradiction* and *entailment* labelled *Premise-Hypothesis* pair created from the template sentence is shown below.

**Premise** : "**Dublin** defied international pressure by announcing plans to produce more fuel for its nuclear program.".

**Hypothesis** : "**Shanghai** defied international pressure by announcing plans to produce more fuel for its nuclear program."

**Gold Label**: *contradiction*

**Premise** : "**Dublin** defied international pressure by announcing plans to produce more fuel for its nuclear program."

**Hypothesis** : "**Dublin** defied international pressure by announcing plans to produce more fuel for its nuclear program."

**Gold Label**: *entailment*

In this manner 46664 *Premise-Hypothesis* pairs are created using the AMR corpus. Just like with bAbI corpus, the set of replacement names and the set of template sentences are kept disjoint for the train, dev and test split.

### 3.1.3 Details of creation of Dataset using CoNLL 2003 Shared task NER data

The shared taks of CoNLL-2003 deals with language-independent named entitiy recognition. The dataset released for this task concentrated on four types of entities: persons, locations, organization and names of miscellaneous entities that do no belong to the other three groups. This corpus contains sentences from Reuters Corpus. These sentences are fully annotated for the Named-Entity Recognition (NER) task. The dataset contains files with four columns separated by a single space. The first column represents the words in the sentence. Each word has been put in a separate line. An empty space represents the end of one sentence. The second item in each line represents a part of speech (POS) tag, the third represents a syntactic chunk and the fourth represents the named-entity tag. An example of the annotation describing the name of a location is shown in Figure 9.

These annotations made it very easy to extract the "candidate" sentences for template creation. Just like with the AMR corpus, the "candidate" sentences are the ones that contain at least one mention of a person or a location. Unlike the AMR

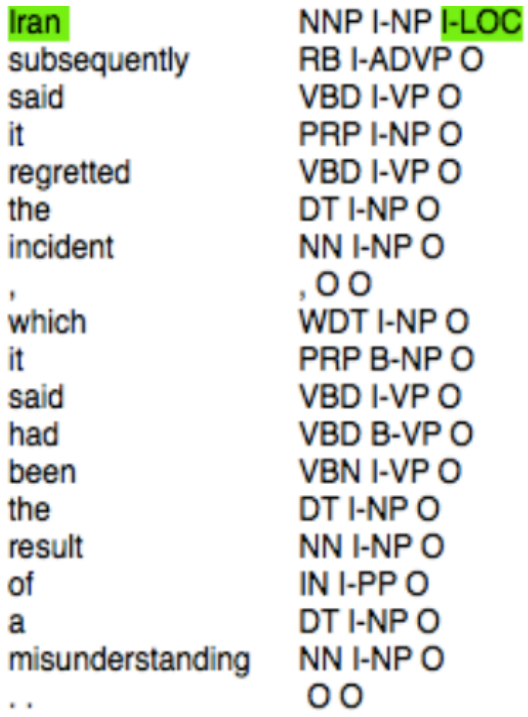



Figure 9. Annotations provided in the CoNLL 2003 corpus

corpus, this dataset's annotations did not distinguish between a city and a country. They were instead clubbed into one single category called location. Based on these "candidate" sentences 3910 template sentences are created. The placeholder tokens used here are **PersonX** for person entity and **LocationX** for location entity. Example of such a template sentence along with the "candidate" sentence it was created from is shown below.

**A "candidate" sentence** : "New Delhi subsequently said it regretted the incident, which it said had been the result of a misunderstanding."

**Template sentence** : "**LocationX** subsequently said it regretted the incident, which it said had been the result of a misunderstanding."

Exact same steps as that with AMR corpus are employed from this point on to automatically generate the *Premise-Hypothesis* pairs. Example of a *contradiction*

and *entailment* labelled *Premise-Hypothesis* pair created from the above mentioned template sentence is shown below.

| |
|---|
| **Premise** : "**Dublin** subsequently said it regretted the incident, which it said had been the result of a misunderstanding.". |
| **Hypothesis** : "**Shanghai** subsequently said it regretted the incident, which it said had been the result of a misunderstanding." |
| **Gold Label**: *contradiction* |

| |
|---|
| **Premise** : "**Dublin** subsequently said it regretted the incident, which it said had been the result of a misunderstanding." |
| **Hypothesis** : "**Dublin** subsequently said it regretted the incident, which it said had been the result of a misunderstanding." |
| **Gold Label**: *entailment* |

This led to the creation of 63333 *Premise-Hypothesis* pairs and as with the other corpora the disjointedness of the set of replacement names and templates was taken care of while splitting the data into train, dev and test sets.

### 3.1.4 Details of creation of Dataset using GMB(Groningen Meaning Bank) dataset

Up till now, our focus has only been on creating pairs of Premise-Hypothesis sentences where the entity changed was "Name" of a person, city or country. This work uses the GMB corpus to create the "NER Changed" dataset for the "date" and "number" named entity. The annotations consists of Part-Of Speech(POS) tags and Named-Entity tags. Here are the following tags in the dataset - geo:Geographical Entity, org:Organization, per:Person, gpe:Geopolitical Entity, tim:Time, indicator art:Artifact, eve:Event and nat:Natural Phenomenon

Using these two types of tags it becomes easy to idenfity the different types of "Date" and "Number" entities. The following two subsections will describe the different types of "Date" and "Number" entities and how they are identified to create template sentences and finally the labelled *Premise-Hypothesis* pairs.

#### 3.1.4.1 Identifying and Changing "Date" Entity

The different types of "Date" entity identified in the GMB corpus are:

1. "Year"
2. "Month"
3. "Day of the week"

Based on the NER annotations (Figure **??**) provided all of these are grouped under "tim" tag. Sentences with at least one mention of "tim" tag are shortlisted. For the sentences thus shortlisted the POS tags are considered to differentiate further among the three "Date" entity types. For the type "Year" the POS tag is always "CD"(*cardinal number*). After this a simple check for the length of the token being equal to four helps in identifying the sentences with "Year" type of "Date" entities. The POS tag for "Month" and "Day of the week' entity type is always "NNP" (*Proper noun, singular*). Once this POS filter is added, a simple keyword search for the names of the twelve months and for the names of the seven days of a week helps in getting the "candidate" sentences for "Month" and "Day of the week" types.

This is how a total of 2037 and 8819 "candidate" sentences for "Year" and "Month"/"Day of the week" were shortlisted. Examples of each of these are shown below:

| In | IN | O | The | DT | O | On | IN | O |
|---|---|---|---|---|---|---|---|---|
| 2005 | CD | B-tim | spokesman | NN | O | Friday | NNP | B-tim |
| , | , | O | says | VBZ | O | , | , | O |
| the | DT | O | a | DT | O | five | CD | O |
| government | NN | O | formal | JJ | O | soldiers | NNS | O |
| passed | VBD | O | agreement | NN | O | were | VBD | O |
| a | DT | O | on | IN | O | killed | VBN | O |
| controversial | JJ | O | the | DT | O | when | WRB | O |
| hydrocarbons | NNS | O | project | NN | O | dozens | NNS | O |
| law | NN | O | will | MD | O | of | IN | O |
| that | WDT | O | be | VB | O | militants | NNS | O |
| imposed | VBD | O | signed | VBN | O | stormed | VBD | O |
| . | . | . | in | IN | O | a | DT | O |
| . | . | . | June | NNP | B-tim | military | JJ | O |
| . | . | . | when | WRB | O | checkpoint | NN | O |
| | | | Indonesian | JJ | B-gpe | in | IN | O |
| | | | President | NNP | B-per | Orakzai | NNP | B-geo |
| | | | Susilo | NNP | I-per | | | |
| | | | Bambang | NNP | I-per | | | |
| | | | Yudhoyono | NNP | I-per | | | |
| | | | is | VBZ | O | | | |
| | | | scheduled | VBN | O | | | |
| | | | to | TO | O | | | |
| | | | visit | VB | O | | | |
| | | | Moscow | NNP | B-geo | | | |

Figure 10. Annotations provided in the GMB corpus

**"Year" "candidate" sentence** : "In **2005**, the government passed a controversial hydrocarbons law that imposed significantly higher royalties and required foreign firms then operating under risk-sharing contracts to surrender all production to the state energy company in exchange for a predetermined service fee."

**"Month" "candidate" sentence** : "The spokesman says a formal agreement on the project will be signed in **June** when Indonesian President Susilo Bambang Yudhoyono is scheduled to visit Moscow."

**"Day" "candidate" sentence** : "On **Friday**, five soldiers were killed when dozens of militants stormed a military checkpoint in Orakzai."

Two random disjoint sets, each consisting of twenty, 4 digit numbers ranging from 1900 to 2019 is created to serve as "Year" replacement options for premise and

hypothesis sentences respectively. The replacement options for "Months" and "Day of week" are the standard list of month names and days of the week.

Using the above mentioned replacement options a total of 112372 labelled *Premise-Hypothesis* pairs are created. Among these 34870 are with respect to "Year" type of "Date" entity, 14645 are with respect to "Month" type of "Date" entity and 62857 are with respect to "Day of week" type of "Date" entity. Example of a *contradiction* labelled *Premise-Hypothesis* pair created for the three types of "Date" entity is shown below. As always the *entailment* labelled *Premise-Hypothesis* pairs are created by keeping the premise and the hypothesis as same.

**Premise** : "In **1985**, the government passed a controversial hydrocarbons law that imposed significantly higher royalties and required foreign firms then operating under risk-sharing contracts to surrender all production to the state energy company in exchange for a predetermined service fee.'

**Hypothesis** : "In **2014**, the government passed a controversial hydrocarbons law that imposed significantly higher royalties and required foreign firms then operating under risk-sharing contracts to surrender all production to the state energy company in exchange for a predetermined service fee."

**Gold Label**: *contradiction*

**Premise** : "The spokesman says a formal agreement on the project will be signed in **February** when Indonesian President Susilo Bambang Yudhoyono is scheduled to visit Moscow."

**Hypothesis** : "The spokesman says a formal agreement on the project will be signed in **November** when Indonesian President Susilo Bambang Yudhoyono is scheduled to visit Moscow."

**Gold Label**: *contradiction*

**Premise** : "On **Sunday**, five soldiers were killed when dozens of militants stormed a military checkpoint in Orakzai."

**Hypothesis** : "On **Tuesday**, five soldiers were killed when dozens of militants stormed a military checkpoint in Orakzai."

**Gold Label**: *contradiction*

#### 3.1.4.2 Identifying and Changing "Number" Entity

To identify the "Number" entity both the NER and POS annotations are looked at (Figure 11). For a "Number" entity the NER annotation is "O" while the POS annotation is "CD"(*cardinal number*). Therefore, the "candidate" sentences are identified that contain at least one token who NER annotation is "O" and the POS annotation is "CD"(*cardinal number*). The "candidate" sentences identified so far are further divided to represent the following two types of "Number" entity:

1. "Cardinal in Numerics"
2. "Cardinal in Words"

The segregation of "candidate" sentences into the "candidate" sentences for the two types of "Number" entity is done using a simple check for whether the "Number" entity token is a digit or not.

As replacement options for the "candidate" sentences of "Cardinal in Numerics" type, two random disjoint sets, each consisting of 30 random numbers ranging from 10 to 20000 are used. Similarly two more random disjoint sets are created which contain only 2 digit numbers. These are automatically converted to word and are used as replacement options for the "candidate" sentences of "Cardinal in Words" type.

| | | | | | |
|---|---|---|---|---|---|
| Australia | NNP | B-geo | That | DT | O |
| has | VBZ | O | agreement | NN | O |
| about | IN | O | ended | VBD | O |
| 1,300 | CD | O | the | DT | O |
| troops | NNS | O | two | CD | O |
| in | IN | O | decade-long | JJ | O |
| Iraq | NNP | B-geo | civil | JJ | O |
| as | IN | O | war | NN | O |
| part | NN | O | between | IN | O |
| of | IN | O | the | DT | O |
| the | DT | O | government | NN | O |
| U.S.-led | JJ | O | in | IN | O |
| coalition | NN | O | Khartoum | NNP | B-geo |
| | | | and | CC | O |
| | | | southern | JJ | O |
| | | | rebels | NNS | O |

Figure 11. Annotations provided in the GMB corpus

This results in a total of 90690 labelled *Premise-Hypothesis* pairs. These included 54500 pairs for "Cardinal in Numerics" type and 36190 pairs for "Cardinal in Words" type. Example of a *contradiction* labelled *Premise-Hypothesis* pair created for the two types of "Number" entity is shown below. The *entailment* labelled *Premise-Hypothesis* pairs are created by keeping the premise and the hypothesis as same.

**Premise** : "Australia has about **14061** troops in Iraq as part of the U.S. led coalition.'

**Hypothesis** : "Australia has about **8958** troops in Iraq as part of the U.S. led coalition."

**Gold Label**: *contradiction*

**Premise** : "That agreement ended the **eleven** decade-long civil war between the government in Khartoum and southern rebels."

**Hypothesis** : "That agreement ended the **two** decade-long civil war between the government in Khartoum and southern rebels."

**Gold Label**: *contradiction*

## 3.2 Role-Switched Dataset

Up till now, we have seen how to automatically create dataset that emphasizes on the importance of "Entities". This section describes the process of creating the Role-Switched Dataset. This dataset emphasizes on capturing the notion of "Roles". It contains examples where the difference in the pair of sentences lies in the roles played by the two persons mentioned in the sentences even though they participate in the same event (verb). To create this dataset, this work uses the following resources and corpora:

1. VerbNet (Schuler 2005)
2. PropBank (Palmer, Gildea, and Kingsbury 2005)
3. QA-SRL (FitzGerald et al. 2018)
4. CoNLL 2004 Shared SRL Task (Carreras and Màrquez 2005)
5. CoNLL 2003 Shared NER Task (Ratinov and Roth 2009)

| | | | | |
|---|---|---|---|---|
| accompany | admit | alert | ask | banish |
| bill | buy | choke | complain | confess |
| conspire | cure | defend | deport | dress |
| drive | drown | force | groom | jump |
| lend | manage | murder | order | persuade |
| poison | provide | ride | row | steal |
| strangle | tell | wag | walk | waltz |

Table 6. List of 35 Verbs shortlisted from VerbNet

### 3.2.1 Details of creation of Dataset using VerbNet

VerbNet (Schuler 2005) lexicon contains a list of VerbNet classes for many types of verbs. Each class mentions the restrictions that define the kind of thematic roles that can be allowed as arguments. As an example, let's consider the VerbNet class "give-13.1" for the verb "give". This class mentions that "Agent", "Theme" and "Recipient" are only roles it can have. It also mentions the restrictions on "Agent" and "Recipient". Both can be either an "Animate" or an "Organization" type of entity.

Using this information 35 VerbNet classes are shortlisted that accepts the same kind of entities for different roles. Table **??** shows all of these 35 verbs. One such class is "give-13.1" because its two different roles "Agent" and "Recipient" can only have the same kind of entity, "Animate" or an "Organization".

VerbNet also provides a list of "member" verbs for each of the VerbNet classes. The "member" verbs can be considered as synonyms. For example, "lend", "loan", "pass" and "peddle" are few of the member verbs for the class "give-13.1". Using the "member" verbs for each VerbNet class shortlisted so far, a list of around 600 "interesting" verbs is generated. Finally the annotated sentences for these verbs are extracted from VerbNet to generate the template sentences.

As an example, consider a sentence from VerbNet and the corresponding template mentioned below. The verb used here is "lent" which is a member verb for "give-13.1"

VerbNet class. Using the QA-SRL parser, the entities for the two roles are identified to create the template sentences.

| |
|---|
| **A "candidate" sentence** :"**They** lent **me** a bicycle." |
| **Template sentence** : "**PersonX** lent **PersonY** a bicycle." |

Note that the example sentences provided by VerbNet for every VerbNet class are not for every member verb. Some of the examples might not contain the required slots of **PersonX** and **PersonY** as well. For these reasons only 87 templates are generated. For the templates generated so far, the member verbs are used to create more templates. The template sentences from this corpus are very simple and thus automatically converting them to different tenses is also very easy. Therefore converted all the template sentences into Present tense in 3rd person and Future tense to get a list of 1611 unique templates. For example, the template sentence "**PersonX** lent **PersonY** a bicycle." is used to create two more template sentences in Present tense in 3rd person and Future tense as shown below:

| |
|---|
| **Template in Present tense in 3rd person** :"**PersonX** lends **PersonY** a bicycle." |
| **Template in Future tense** : "**PersonX** will lend **PersonY** a bicycle." |

For all such template sentences, the list of gender neutral names mentioned in Table 5 are used to create 134821 labelled *premise-hypothesis* pairs. An example of a *contradiction* and an *entailment* labelled *premise-hypothesis* pair is shown below:

| |
|---|
| **Premise** : **Kendall** lent **Peyton** a bicycle. |
| **Hypothesis** : **Peyton** lent **Kendall** a bicycle. |
| **Gold Label**: *contradiction* |

| |
|---|
| **Premise** : **Kendall** lent **Peyton** a bicycle. |
| **Hypothesis** : **Kendall** lent **Peyton** a bicycle. |
| **Gold Label**: *entailment* |

| | | | | | |
|---|---|---|---|---|---|
| lead | admit | advise | cheat | bill | say |
| banish | hire | give | stole | tell | |

Table 7. List of 11 Verbs shortlisted from PropBank

### 3.2.2 Details of creation of Dataset using PropBank

PropBank (Proposition Bank) is a large corpus of sentences with annotations for propositions and predicate argument relations. It also provides a mapping to VerbNet. This work uses these mappings to VerbNet in order to extract example sentences from PropBank for the shortlisted VerbNet Classes. Not all the extracted example sentences are good enough to create the desired template sentences. For example:

| |
|---|
| "The Beatles give way to baseball in the Nipponese version." |

Therefore, we manually look at these examples and shortlist 13 example sentences to manually create 13 template sentences. These sentences contain 11 verbs as shown in Table 7. One example from the 13 template sentences is " **PersonX** led **PersonY** from the frying pan to the fire." Just like with the VerbNet template sentences, more PropBank template sentences are created by adding sentences for different tenses. Since these template sentences are more complicated than VerbNet template sentences manual supervision is required. This way a total of 89 template sentences are created.

The list of gender neutral names mentioned in Table 5 are used as replacement options to create 9648 labelled *premise-hypothesis* pairs. An example of a *contradiction* and an *entailment* labelled *premise-hypothesis* pair is shown below:

| |
|---|
| **Premise** : **Kendall** led **Peyton** from the frying pan to the fire. |
| **Hypothesis** : **Peyton** led **Kendall** from the frying pan to the fire. |
| **Gold Label**: *contradiction* |

| |
|---|
| **Premise** : **Kendall** led **Peyton** from the frying pan to the fire. |
| **Hypothesis** : **Kendall** led **Peyton** from the frying pan to the fire. |
| **Gold Label**: *entailment* |

### 3.2.3 Details of creation of Dataset using QA-SRL

The QA-SRL (FitzGerald et al. 2018) dataset is a very large corpus for Semantic Role Labeling. It contains the semantic role labeling annotations in a Question and Answering format. It contains 250,000 question-answer pairs for over 64,000 sentences. For each sentence's each verbal predict a question-answer pair where each answer is a set of contiguous spans from the sentence.

Therefore I shortlist the QA-SRL sentences that contain "Who" questions for the verbs shown in Table 6. The assumption here was that switching the answers to these questions could help in generating the *premise-hypothesis* pairs. Consider the example given below from the QA-SRL dataset for the verb "protect" which is the member verb for the VerbNet class "defend-85". The role switched sentence here is used as the hypothesis generating a contradiction label for the premise(original sentence).

**Verb** : protect

**Sentence** : "In Germany, the Emperor had repeatedly protected Henry the Lion against complaints by rival princes or cities especially in the cases of Munich and Lbeck."

**Who protected someone?**: *the Emperor*

**Who did someone protect?**: *Henry the Lion*

**Role Switched Sentences**: In Germany, *Henry the Lion* had repeatedly protected *the Emperor* against complaints by rival princes or cities especially in the cases of Munich and Lbeck.

This leads to the creation of 349 labelled *premise-hypothesis* pairs. An example of such a "contradiction" labelled *premise-hypothesis* pair is shown below:

**Premise** : Many kinds of power plant have been used to drive propellers.

**Hypothesis** : Propellers have been used to drive many kinds of power plant.

**Gold Label**: contradiction

More sentences with two "Who" questions for the given verb were shortlisted. The assumption that switching the answers to these questions could help in generating the *premise-hypothesis* pairs doesn't hold true for few of the cases as shown below:

**Verb** : rally

**Sentence** : "Sombat Boonngam-anong, a social activist who has made efforts to rally the Thai people in continued protest, offered his own interpretation of the gesture on Facebook."

**Who is rallying someone?**: *Sombat Boonngam-anong , a social activist*

**Who is someone rallying?**: *the Thai people*

**Role Switched Sentences**: *the Thai people* who has made efforts to rally *Sombat Boonngam-anong , a social activist* in continued protest, offered his own interpretation of the gesture on Facebook.

**Verb** : welcome

**Sentence** : "Fraser has welcomed Vietnamese immigrants to Australia during his term."

**Who is rallying someone?**: *Vietnamese immigrants*

**Who is someone rallying?**: *Fraser*

**Role Switched Sentences**: *Vietnamese immigrants* has welcomed *Fraser* to Australia during his term.

Manual supervision was required to remove such cases. After removing such cases 109 templates are automatically created. We then use the list of gender neutral names mentioned in Table 5 to create 6636 labelled *premise-hypothesis* pairs. An example of "contradiction" labelled pair for the template sentence "**PersonX** asked that she be allowed to inform **PersonY** before the news was released." is shown below:

**Premise** : **Kendall** asked that she be allowed to inform **Peyton** before the news was released.

**Hypothesis** : **Peyton** asked that she be allowed to inform **Kendall** before the news was released.

**Gold Label**: contradiction

| | | | | | | |
|---|---|---|---|---|---|---|
| being an agent of | met criticism from | succeed | flip-flop | urge | meet | attack |
| call | introduce | talk | serve | lead | manage | relay |

Table 8. List of 14 Verbs from CoNLL 2004

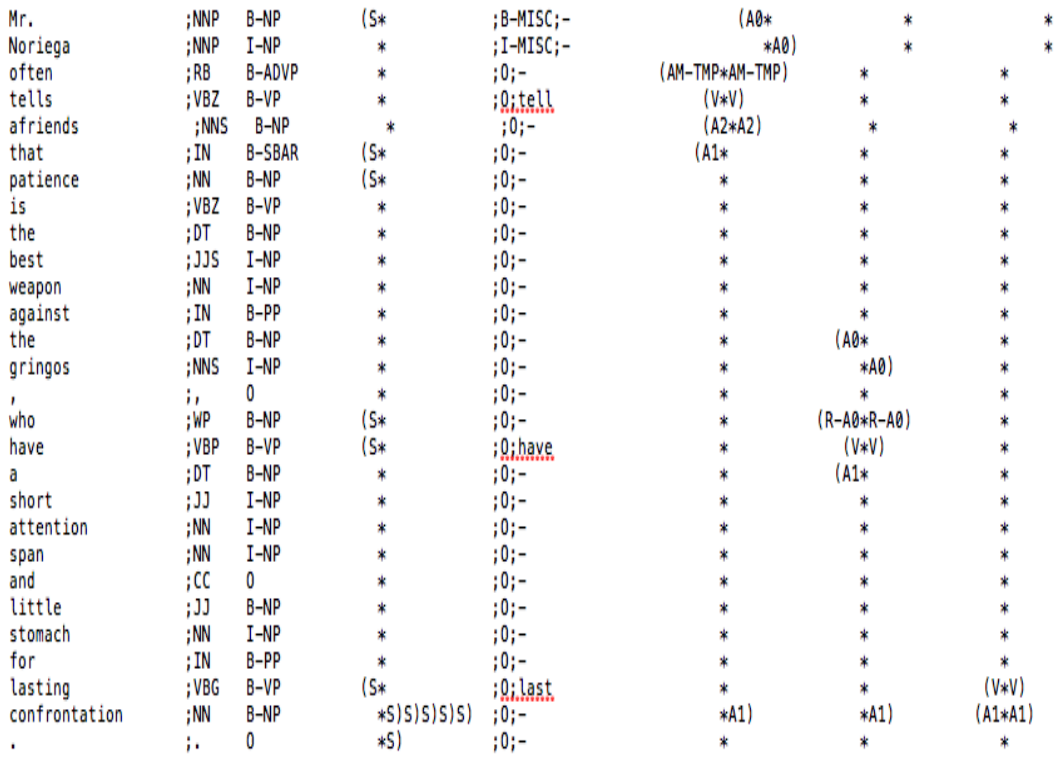



Figure 12. Annotations provided in the CoNLL 2004 corpus

### 3.2.4 Details of creation of Dataset using CoNLL 2004

The goal of CoNLL 2004 shared task was of Semantic Role Labeling. It aimed at creating a machine learning system that could recognize the arguments of a verb and assign them with their correct semantic role. As a part of the annotations this dataset contained part of speech tags, NER tags and syntactic parse trees for each of the sentence. Figure 12 shows an example of these annotations.

In this work, the NER tags are used. We shortlist sentences with mentions of 2 persons. This results in roughly 150 sentences. Manually going through each of these sentences, we filtered out few sentences for which switching the roles won't create meaningfull sentences. The sentences thus shortlisted included roughly 14

verbs with some of them being phrasal verbs as shown in Table 8. Using the VerbNet member verbs for these verbs led to the creation of 71 template sentences. These template sentences are more complex as compared to the VerbNet template sentences. Examples of such template sentences are shown below:

| |
|---|
| And, according to one dealer, Mr. **PersonX** had a penchant for introducing Mr. **PersonY** with the phrase : " He can buy anything. " |
| Col. North conveyed the request to his superiors and to Assistant Secretary of State **PersonX**, who will deliver it to Secretary of State **PersonY**. |

The list of gender neutral names shown in Table 5 is used to fill in the template sentence and 6396 labelled *premise-hypothesis* pairs are created. An example of such a "contradiction" labelled *premise-hypothesis* pair is shown below:

| |
|---|
| **Premise** : Col. North conveyed the request to his superiors and to Assistant Secretary of State **Kendall**, who will deliver it to Secretary of State **Peyton**. |
| **Hypothesis** : Col. North conveyed the request to his superiors and to Assistant Secretary of State **Peyton**, who will deliver it to Secretary of State **Kendall**. |
| **Gold Label**: contradiction |

### 3.2.5 Details of creation of Dataset using CoNLL 2003

The shared taks of CoNLL-2003 is about language-independent named entitiy recognition. There are four types of entities tagged in the dataset released for this task: persons, locations, organization and names of miscellaneous entities that do no belong to the other three groups. These sentences are fully annotated for the Named-Entity Recognition (NER) task. The dataset contains files with four columns separated by a single space. The first column represents the words in the sentence.

| watch | replace | beat | outplay | upset |
|---|---|---|---|---|
| concede | tell | describe | meet | send |
| lead | eject | name | endorse | bowl |
| appoint | chip | make | add | welcome |
| propose | ahead | miss | collide | lose |
| gainst | play | force | tested against | eliminated by |
| led against | victory over | furious with | running second | called up |
| did not name | towering over | is the main political rival | is a friend of | act as assistant to |
| won any specific commitment | drew the ire of | was secretary of state in the Republican administration | fought off | dived to his right to save |
| knocked out | nominated by | fired two shots that killed | has been | takes on |

Table 9. List of 55 Verbs from CoNLL 2003

Each word has been put in a separate line. An empty space represents the end of one sentence. The second item in each line represents a part of speech (POS) tag, the third represents a syntactic chunk and the fourth represents the named-entity tag. An example of the annotation describing the name of a location is shown in Figure 9.

The NER tags present in the annotations provided in dataset are used to extract candidate sentences. To shortlist the "candidate" sentences we look at sentences with mentions of 2 persons. This results in roughly 200 sentences. Not all of these "candidate" sentences are desirable because switching roles might lead to a grammatically incorrect sentence. In some cases switching roles might lead a sentence that might not contradict the orginal sentence. Therefore we manually go through them to filter out such sentences. The new subset of "candidate" sentences included roughly 55 verbs with some of them being phrasal verbs as shown in Table 8. A total of 234 template sentences are created using the VerbNet member verbs for the verbs involded. Just like for the template sentences from CoNLL 2004 corpus, these template

sentences are also more complex as compared to the VerbNet template sentences. Examples of such template sentences are shown below:

| |
|---|
| **PersonX**, who led a rebel group against the military regime of **PersonY** in the late 1980s, is now a successful businessman with mining and logging interests. |
| **PersonX** has decided not to endorse **PersonY** as the presidential candidate of the Reform Party, CNN reported late Tuesday. |

As done with the other resources used in this work, the list of gender neutral names shown in Table 5 is used to fill in the template sentence and 14688 labelled *premise-hypothesis* pairs are created. An example of such a "contradiction" labelled *premise-hypothesis* pair is shown below:

| |
|---|
| **Premise** : **Kendall** has decided not to endorse **Peyton** as the presidential candidate of the Reform Party, CNN reported late Tuesday.. |
| **Hypothesis** : **Peyton** has decided not to endorse **Kendall** as the presidential candidate of the Reform Party, CNN reported late Tuesday.. |
| **Gold Label**: contradiction |

# Chapter 4

# MODEL AND PROPOSED ATTENTION MECHANISM

In the previous chapters, we have discussed about the issues that the current NLI systems suffer from. We saw that these systems lacked the understanding of "Entities" and "Roles". There were two main reasons for these drawbacks. In the Data Generation chapter, we tackled the first reason of this drawback. This reason was the lack of adversarial examples in the existing datasets like SNLI and MNLI. To overcome this, the Data Generation chapter showed how to generate labelled premise-hypothesis pairs that could help in overcoming this drawback.

The other reason for these drawbacks was the strict use of vector similarity in the Word-To-Word Attention mechanism of the existing architectures. Now we will see how to address this reason and see how we can informatively move away from strict use of vector similarity and propose a novel Word-To-Word Attention mechanism that relies on both Vector Similarity and Symbolic Similarity.

This chapter describes the attention mechanism that is used by the DecAtt DBLP:journals/corr/ParikhT0U16 and the ESIM DBLP:journals/corr/ChenZLWJ16 model. Later in this section, the modification that helps these models to perform better on NER CHANGED dataset is also described. Section 4.1 formalizes the DecAtt and ESIM models describing the input and the attention mechanism employed in these models. Section 4.2 describes the proposed attention mechanism that incorporates both Vector and Symbolic similarity and is used to enable DecAtt and ESIM model in capturing the notion of "Entities". In Section 4.3 we describe rest of the steps which

remain unchanged between the new and proposed architectures for both DecAtt and ESIM.

## 4.1 Formalizing DecAtt and ESIM models

Let a and b be two input sentences of length $l_a$ and $l_b$ such that a = ($a_1$,$a_2$,...,$a_{l_a}$) and b = ($b_1$,$b_2$,...,$b_{l_b}$) where each $a_i$ and $b_j \in \mathrm{R}^{\mathrm{d}}$ is a word vector embedding of dimensions $d$. Here a is the premise and b is the hypothesis.

The first step in both Decomposable attention(DecAtt) model and Enhanced Sequential Inference Model(ESIM) is to form the soft alignments between the subphrases of premise and hypothesis sentences. In other words, we try to find the subphrase in the premise that is (softly) aligned with the hypothesis and vice versa. This step is referred to as the Attend step and Locality of inference step in DecAtt and ESIM model respectively.

This is done by computing the unnormalized attention weights $e_{ij}$. Both Decomposable attention(DecAtt) model and Enhanced Sequential Inference Model(ESIM) computes $e_{ij}$ in a similar manner with slight difference between the two. DecAtt Model uses a feed-forward neural network ($F$) with ReLU activations to transform the input word embeddings and then computes the dot product of thus transformed vectors as described in Equation 4.1. ESIM on the other hand transforms the input word embeddings using a BiLSTM (Equation 4.2 and 4.3) and uses the hidden state tuples (Figure 13) of words from the two sentences to compute the attention weights as described in Equation 4.4.

$$e_{\mathrm{ij}} = F(a_i)^T F(b_j) \tag{4.1}$$

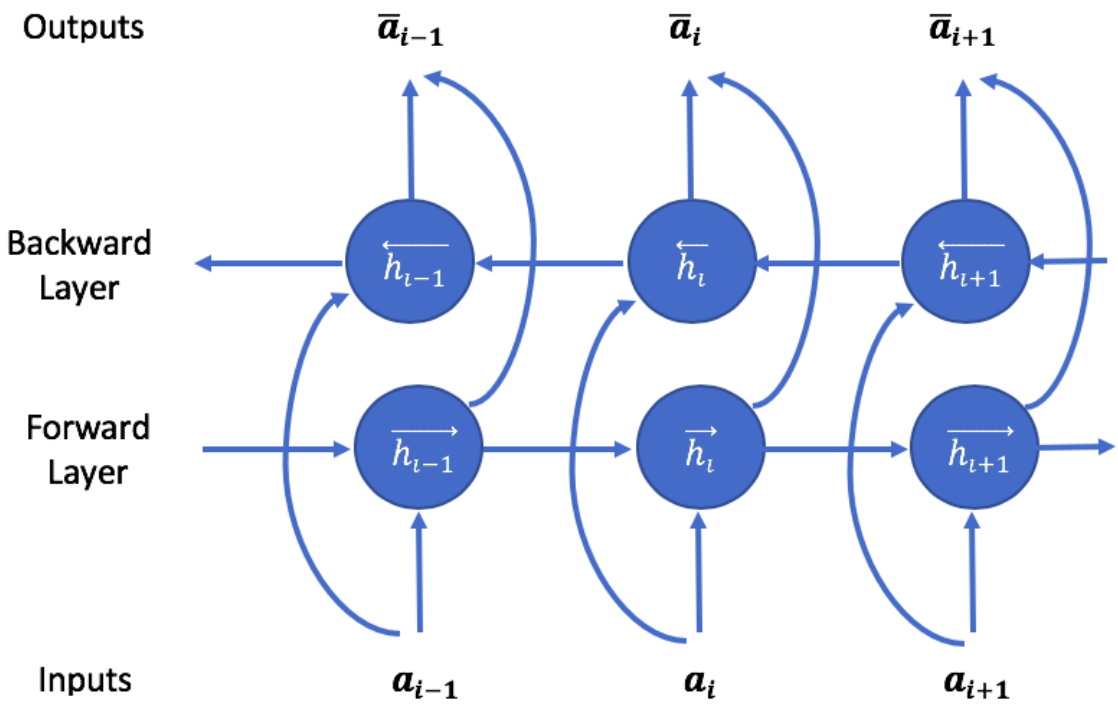



Figure 13. This figure shows the pictorial representation of a Bidirectional LSTM

$$\bar{a}_{\mathrm{i}} = BiLSTM\,(a,i) \forall i \in [1, ..., l_{\mathrm{a}}] \tag{4.2}$$

$$\bar{b}_{\mathrm{j}} = BiLSTM\,(b,j) \forall j \in [1, ..., l_{\mathrm{b}}] \tag{4.3}$$

$$e_{\mathrm{ij}} = (\bar{a}_{\mathrm{i}})^T \bar{b}_{\mathrm{j}} \tag{4.4}$$

## 4.2 A Novel Word-To-Word Attention Mechanism

First let us see why the existing attention mechanism fails to detect a change in an entity between the premise and hypothesis sentence. Let $_{\mathrm{i}}$ represent the word

| | John | went | to | the | kitchen |
|---|---|---|---|---|---|
| Peter | 17.11 | 7.596 | 3.366 | 4.096 | 2.560 |
| went | 4.565 | **18.01** | 11.44 | 7.532 | 4.173 |
| to | 2.166 | 11.51 | **22.35** | 13.31 | 5.829 |
| the | 3.120 | 7.183 | 12.91 | **25.76** | 7.060 |
| kitchen | -0.194 | 3.728 | 5.320 | 6.970 | **28.96** |

Figure 14. Attention weights of the existing word-to-word attention mechanism (ESIM)

embedding for the name "John" while $_j$ represent the word embedding for the name "Peter". Since names are used in a similar context, the vectors $_i$ and $_j$ will be highly similar. Now consider a case where the remainder of the premise and hypothesis sentences are entirely same. For example:

**Premise** : "**John** went to the kitchen."

**Hypothesis** : "**Peter** went to the kitchen."

The attention weights for the pair mentioned in the previous section would result in a diagonal matrix as shown in Figure 14. This usually happens when the premise is the same as hypothesis. This leads to a wrong "entailment" prediction for such pairs of premise and hypothesis.

This is where we introduce symbolic similarity in the attention mechanism to

deal with this issue. We learn a weight $\lambda$ which decides how much weight should be given to vector similarity and the symbolic similarity ($sym_{ij}$) while calculating the new unnormalized attention weights $e'_{ij}$

$$e'_{\mathrm{ij}} = \lambda_{\mathrm{ij}} e_{\mathrm{ij}} + (1 - \lambda_{\mathrm{ij}}) sym_{\mathrm{ij}} \tag{4.5}$$

$sym_{ij}$ represents the symbolic similarity which is assigned 0 if the string representing $\mathrm{a_i}$ is not equal to the string representing $\mathrm{b_j}$. If the two string matches, then a weight $w$ which is a hyper-parameter, is assigned.

$\lambda_{\mathrm{ij}}$ is calculated using a feed-forward neural network whose output is calculated as a single neuron with a custom activation function which is a variant of LeakyRelu as described in equation 6. We will refer to this feed-forward neural network as the *Lambda Layer*.

$$1 - LeakyReLU(1 - LeakyReLU(W_\lambda X_\lambda)) \tag{4.6}$$

$\mathrm{W}_\lambda$ represents the weight vector for the lambda layer which is learnt from the training data. $\mathrm{X}_\lambda$ is the input to the lambda layer which is a 16 dimensional sparse feature vector. It encodes the NER (Named Entity Recognition) information for the pair of words in the two sentences. The NER information is grouped in to 4 categories. The four categories are "Name", "Numeric", "Dates" and "Others". To identify these categories for each word in the sentence, Spacy and Stanford NER tagger is used.

Each dimension of the input feature vector represents the pair-wise combination of the 4 categories. Table 10 shows what the 16 dimensions represent.

We then create a 16 dimensional sparse vector for each $\mathrm{a_i}$-$\mathrm{b_j}$ pair. This vector is sparse as it contains a value 1 for whichever pair-wise combination of the 4 categories the pair $\mathrm{a_i}$-$\mathrm{b_j}$satisfies and a value of 0 otherwise. For the example mentioned above

| | | | |
|---|---|---|---|
| Dimension1: "Names-Names" | Dimension2: "Names-Dates" | Dimension3: "Names-Numeric" | Dimension4: "Names-Others" |
| Dimension5: "Date-Names" | Dimension6: "Dates-Dates" | Dimension7: "Dates-Numeric" | Dimension8: "Dates-Others" |
| Dimension9: "Numeric-Names" | Dimension10: "Numeric-Dates" | Dimension11: "Numeric-Numeric" | Dimension12: "Numeric-Others" |
| Dimension13: "Others-Names" | Dimension14: "Others-Dates" | Dimension15: "Others-Numeric" | Dimension16: "Others-Others" |

Table 10. Meaning of each dimension of the 16 dimensional feature vector

with premise as "**John** went to the kitchen." and hypothesis as "**Peter** went to the kitchen." a total of 25 feature vectors will be created, each of 16 dimensions. The feature vectors for few of the word-pairs "Kendal" and "Peyton" and the 16 dimension which are shown in Table 11.

| **Dimension Number** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| "John-Peter" | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "John-went" | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "went-kitchen" | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |

Table 11. Meaning of each dimension of the 16 dimensional feature vector

In this work we use the proposed word to word attention mechanism to calculate the new attention weights $e'_{ij}$. Figure 15 shows the new attention weights $e'_{ij}$. As opposed to Figure 14, the new attention weights for "Peter-John" are smaller as it should be since they are different named entities.

## 4.3 DecAtt and ESIM Continued

Using the proposed mechanism instead of the original mechanism results in the creation of two models, "Lambda DecAtt" which is the variant of DecAtt model and "Lambda ESIM" which is the variant of ESIM model. We collectively refer to these

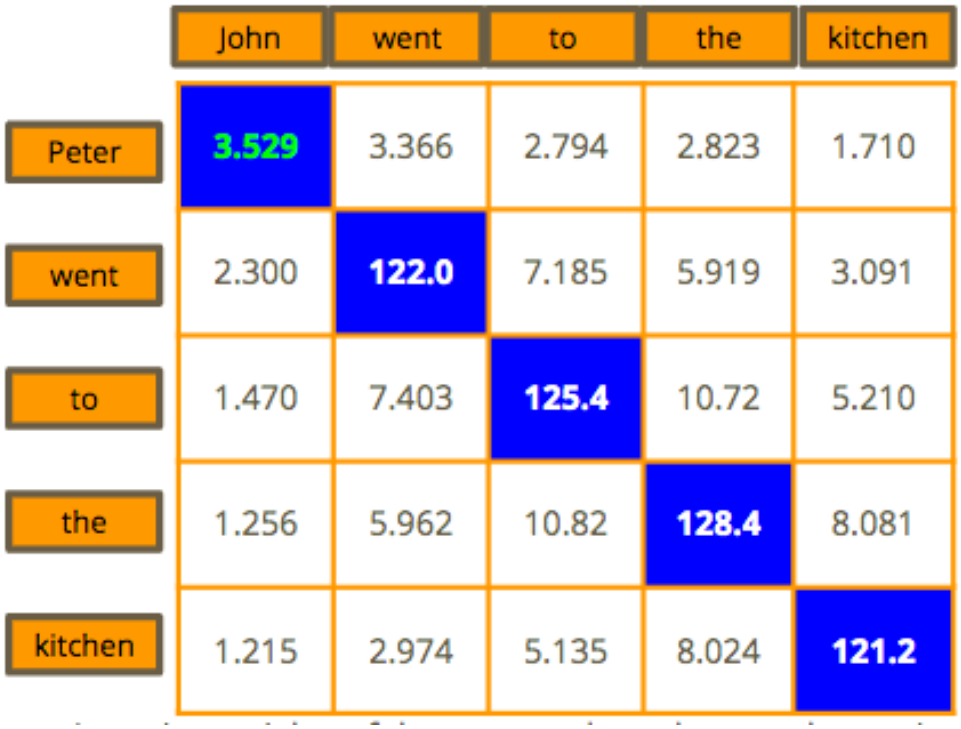



Figure 15. Attention weights of the proposed word-to-word attention mechanism (Lambda ESIM)

variants as "Lambda" architectures or "Lambda" models. Rest of the steps in these "Lambda" models are the same as their original and unmodified counterparts.

The attention weights e'$_{ij}$ thus calculated are unnormalized. Therefor these attention weights are normalized as shown in the Equation 4.7 and 4.8. This step remains the same for both "Lambda DecAtt" and "Lambda ESIM" models. Here $\beta_i$ refers to the subphrase in $\bar{b}$ (*Premise Encoding*) that is softly aligned to $\bar{a}_i$ (*Encoding of the $i^{th}$ token of the Hypothesis*) and vice versa.

$$\beta_i = \sum_{j=1}^{l_b} \frac{\exp(e'_{ij})}{\sum_{k=1}^{l_b} \exp(e'_{kj})} \bar{b}_j \forall i \in [1, ..., l_a] \tag{4.7}$$

$$\alpha_j = \sum_{i=1}^{l_a} \frac{\exp(e'_{ij})}{\sum_{k=1}^{l_a} \exp(e'_{kj})} \bar{a}_i \forall j \in [1, ..., l_b] \tag{4.8}$$

The next step in the "Lambda DecAtt" and the original "DecAtt" model is referred to as the Compare step. Here the aligned phrases are separately compared using a function G which is a feed-forward network (Equation 4.9 and 4.10) to produce two sets of comparison vectors $\{\mathrm{V_{1,i}}\}^{\mathrm{l_a}}{}_{\mathrm{i=1}}$ and $\{\mathrm{V_{2,j}}\}^{\mathrm{l_b}}{}_{\mathrm{j=1}}$. Each set of comparison vectors are aggregated over by summation as shown in Equation 4.11 and 4.12 and the output $\mathrm{V_1}$ and $\mathrm{V_2}$ is fed through a final classifier H. H is a feed forward network followed by a linear layer as shown in Equation 4.13. $\hat{y} \in \mathcal{R}^{\mathrm{C}}$ represents the predicted scores for each class.

$$V_{1,i} = G([\bar{a}_i, \beta_{\mathrm{i}}])\forall i \in [1, ..., l_{\mathrm{a}}] \tag{4.9}$$

$$V_{2,j} = G([\bar{b}_j, \alpha_{\mathrm{j}}])\forall j \in [1, ..., l_{\mathrm{b}}] \tag{4.10}$$

$$V_1 = \sum_{i=1}^{l_{\mathrm{a}}} V_{1,i} \tag{4.11}$$

$$V_2 = \sum_{j=1}^{l_{\mathrm{b}}} V_{2,j} \tag{4.12}$$

$$\hat{y} = H([V_1, V_2]) \tag{4.13}$$

In contrast, the next step in "Lambda ESIM" and the original "ESIM" involves the enhancement of the subphrases calculated by Equations 4.7 and 4.8 also referred as the local inference information. This is done by computing element wise difference and product for the tuple $< \bar{a}, \beta >$ and $< \bar{b}, \alpha >$. These are then concatenated with the original vectors $\bar{a}$ and $\beta$ ,or $\bar{b}$ and $\alpha$ respectively as shown in Equations 4.14 and 4.15.

$$m_{\mathrm{a}} = [\bar{a}; \beta; \bar{a} - \beta; \bar{a} \odot \beta] \tag{4.14}$$

$$m_{\mathrm{b}} = [\bar{b}; \alpha; \bar{b} - \alpha; \bar{b} \odot \alpha] \tag{4.15}$$

$$V_{1,\mathrm{ave}} = \sum_{i=1}^{l_{\mathrm{a}}} \frac{V_{1,i}}{l_{\mathrm{a}}}, V_{1,\mathrm{max}} = \max_{1 \leq i \leq l_{\mathrm{a}}} V_{1,i} \tag{4.16}$$

The next step here is referred to as the Inference Composition step where the enhanced local inference information is fed to a BiLSTM to produce two sets of vectors $\{\mathrm{V_{1,i}}\}^{\mathrm{l_a}}{}_{\mathrm{i=1}}$ and $\{\mathrm{V_{2,j}}\}^{\mathrm{l_b}}{}_{\mathrm{j=1}}$ similar to those in the DecAtt models. Although unlike the Aggregate step of the "Lambda DecAtt" and "DecAtt" models, instead of summation, the Pooling step of the "Lambda ESIM" and "ESIM" models computes both max and average pool as shown in Equations 4.16 and 4.17. These are then concatenated as shown in Equation 4.18 and fed to a final multilayer perceptron(MLP) classifier.

$$V_{2,\mathrm{ave}} = \sum_{j=1}^{l_{\mathrm{b}}} \frac{V_{2,j}}{l_{\mathrm{b}}}, V_{2,\mathrm{max}} = \max_{1 \leq j \leq l_{\mathrm{b}}} V_{2,j} \tag{4.17}$$

$$V = [V_{1,\mathrm{ave}}; V_{1,\mathrm{max}}; V_{2,\mathrm{ave}}; V_{2,\mathrm{max}}] \tag{4.18}$$

# Chapter 5

# EXPERIMENTS AND ANALYSIS

This section describes all the experiments conducted in this work along with the findings of each experiment. This section also shows the learnt weights of the lambda layer of the new attention mechanism proposed in this work. These weights provide an explanation as to how well or how poorly each of the different experiments were able to capture the notion of "Entities".

In Section 5.1 we discuss how the learnt values of the weights of Lambda Layer can help in explaining what the model has learnt. Section 5.2 describes the experimental setup employed in this work. Section 5.3 to 5.10 details the findings and analysis of each of the experiments conducted in this work.

## 5.1 Importance of Learnt Lambda Weights

As explained in the previous chapter each dimension of the lambda layer represents a pair-wise combination of the 4 categories of entities considered in this work ("Names", "Dates", "Numeric" and "Others"). The input to this layer is a one hot vector representative of the kind of "Entities" the words in a pair belongs to. Therefore the weights learnt by this layer can be easily used to see if the model has learned to choose between Symbolic and Vector similarity or not while calculating the attention weights of different pairs of words based on the kind of entities these words belong to.

By looking at the weights learnt by the lambda layer, we can find out what the model has learned. If the weights for the dimensions representing "Names"-"Names",

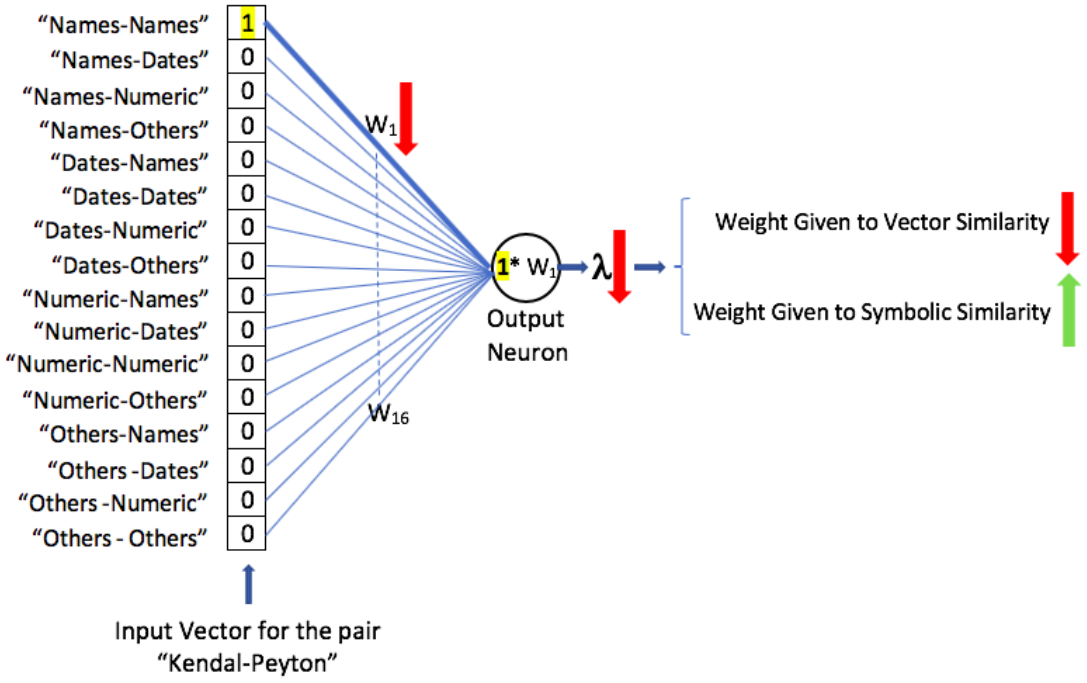



Figure 16. This figure shows the input to the Lambda layer for the word pair "Kendall" and "Peyton". It also shows how the weight of the corresponding dimension effects the weight given to the Vector and Symbolic similarity. The Output Neuron shown here has a variant of LeakyReLU as its activation function as described in previous chapter.

"Dates"-"Dates" and "Numeric"-"Numeric" pair-wise combinations are small then it is indicative of the model learning to give more weight to Symbolic similarity over Vector similarity. A higher weight for these dimensions will indicate that the model has learnt to give more weight to Vector similarity over Symbolic Similarity. We expect the weight for dimension representing "Others"-"Others" pair-wise combination to have a high value. This is because if the two words in a pair belong to "Others" category, we would prefer the model to give higher weight to Vector similarity over Symbolic similarity.

Consider an example with premise as "Kendall moved to the hallway." and

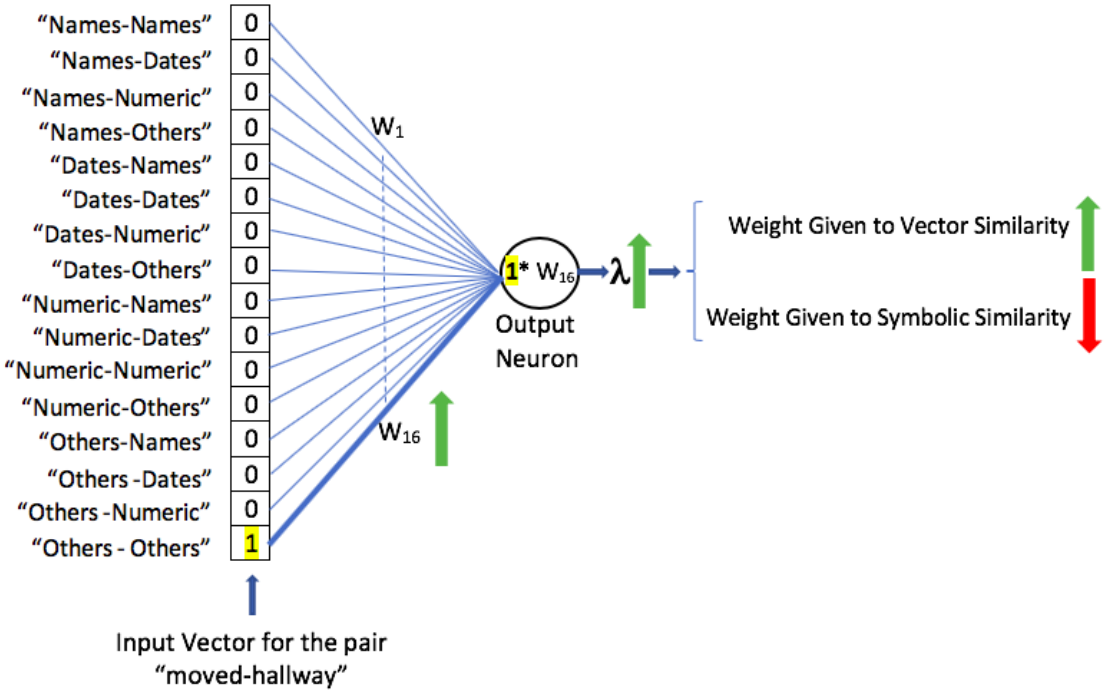



Figure 17. This figure shows the input to the Lambda layer for the word pair "moved" and "hallway". It also shows how the weight of the corresponding dimension effects the weight given to the Vector and Symbolic similarity. The Output Neuron shown here has a variant of LeakyReLU as its activation function as described in previous chapter.

hypothesis as "Peyton moved to the hallway.". Here the words "Kendall" and "Peyton" belong to the entity of type "Names" while all the other words belong the "Others" entity type. Figure 16 shows the input vector to the lambda layer for the pair "Kendall" and "Peyton". These words belong to "Names" entity type and therefor as shown in the Figure 16 we would prefer the model to learn a small value for weight $W_1$. This will result in more weight being given to Symbolic similarity and less weight being given to Vector similarity while calculating the attention weights for this pair.

Let us also look at a case where both the words in a pair belong to the entity type "Others". The word pair "moved"-"hallway" is an example of such a scenario. Here

we would want the model to learn a high value as the weight for the corresponding dimension of the lambda layer. Figure 17 shows how having a high value for the weight $W_{16}$ will result in less weight being given to Symbolic similarity and more weight being given to Vector similarity while calculating the attention weights for this pair.

## 5.2 Experimental Setup

For the experiments, the "NER Changed" and "Roles-Switched" dataset is split in Train, Dev and Test sets each containing 289K, 26.5k, 26.6K and 129K, 8.5k, 9k "premise-hypothesis" pairs respectively. This split is used to evaluate the performance of a total of 5 models. Among these, there are three existing models DecAtt (Parikh et al. 2016), ESIM (Chen et al. 2016) and BERT (Devlin et al. 2018). The other two models are the ones that include this works modification to the attention mechanism. We will refer to them as Lambda DecAtt and Lambda ESIM. The results are shown in Table 12 and Table 13.

| Exp | Data Sets | | DecAtt | | ESIM | | L DecAtt | | L ESIM | | BERT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) |
| 1 | SNLI | NC | 84.58 | 33.80 | 89.78 | 56.11 | 85.10 | **53.69** | 90.10 | **53.95** | 88.45 | 55.56 |
| 2 | SNLI | RS | 84.58 | 49.96 | 89.78 | 50.46 | 85.10 | **49.98** | 90.11 | 54.14 | 88.45 | 49.74 |
| 3 | SNLI | SNLI | 84.58 | 85.01 | 89.78 | 87.96 | 85.10 | **84.58** | 90.10 | **87.88** | 88.45 | 89.16 |
| 4 | SNLI + NC | NC | 88.52 | 82.91 | 91.21 | 69.77 | 88.56 | **97.81** | 92.14 | **99.13** | 85.19 | 69.31 |
| 5 | SNLI + NC | SNLI | 88.52 | 83.50 | 91.21 | 85.09 | 88.56 | 82.94 | 92.14 | 87.10 | 85.19 | **88.26** |
| 6 | SNLI + RS | RS | 78.96 | 49.93 | 89.66 | 93.72 | 78.98 | 49.94 | 90.23 | **90.11** | 82.33 | 49.78 |
| 7 | SNLI + RS | SNLI | 78.96 | 85.31 | 89.66 | 86.99 | 78.98 | 84.52 | 90.23 | 87.70 | 82.33 | **88.47** |
| 8 | SNLI + RS + NC | NC | 84.18 | 80.72 | 92.63 | 75.64 | 84.77 | **95.32** | 92.75 | **98.91** | 81.26 | 69.00 |
| 9 | SNLI + RS + NC | SNLI | 84.18 | 83.71 | 92.63 | 87.03 | 84.77 | 84.24 | 92.75 | 87.28 | 81.26 | **88.08** |
| 10 | SNLI + RS + NC | RS | 84.18 | 50.11 | 92.63 | 84.45 | 84.77 | 50.08 | 92.75 | **87.92** | 81.26 | 49.80 |

Table 12. Table shows the train and test set accuracy for all the experiments involving SNLI dataset. Here, L DecAtt and L ESIM refers to the Lambda DecAtt and Lambda ESIM models. NC refers to NER-Changed dataset, RS refers to the Role-Switched dataset. Each row of this table represents an experiment. The Second and Third columns of each row represents the train set and the test set used for that experiment. Rest of the columns show the train and the test accuracy (Acc) in percentages for all the five models. In our experiments, we have used the *bert-large-uncased* model.

### 5.3 Experiment 1, 2 and 3

Here the 5 models are trained on only SNLI train set. As shown in the Table 12, all the models perform poorly on the "NER Changed" and "Roles-Switched" datasets. The reason for this behavior is the lack of examples that emphasize on capturing notion of "Entities" and "Roles" in the SNLI datastet. Although capable of learning these notions, the Lambda DecAtt and Lambda ESIM models still perform poorly due to the lack of such examples at train time.

| Exp | Data Sets | | DecAtt | | ESIM | | L DecAtt | | L ESIM | | BERT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) | Train Acc(%) | Test Acc(%) |
| 10 | MNLI | NC | 74.47 | 61.16 | 83.71 | 70.35 | 74.00 | **78.76** | 85.20 | **68.15** | 81.87 | 54.19 |
| 11 | MNLI | RS | 74.47 | 50.08 | 83.71 | 50.16 | 74.00 | 50.12 | 85.20 | 50.64 | 81.87 | 50.18 |
| 12 | MNLI + NC | NC | 84.40 | 85.56 | 88.82 | 75.58 | 83.50 | **97.94** | 88.30 | **95.12** | 80.13 | 68.03 |
| 13 | MNLI + NC | MNLI MisM | 84.40 | 71.49 | 88.82 | 75.59 | 83.50 | 70.08 | 88.30 | 74.29 | 80.13 | 79.96 |
| 14 | MNLI + NC | MNLI M | 84.40 | 71.76 | 88.82 | 76.75 | 83.50 | 69.95 | 88.30 | 75.17 | 80.13 | 79.74 |
| 15 | MNLI + RS | RS | 69.75 | 50.08 | 85.01 | 50.51 | 63.40 | 50.13 | 84.12 | 50.12 | 75.09 | 49.53 |
| 16 | MNLI + RS | MNLI MisM | 69.75 | 71.58 | 85.01 | 75.75 | 63.40 | 70.85 | 84.12 | 74.51 | 75.09 | 80.90 |
| 17 | MNLI + RS | MNLI M | 69.75 | 71.72 | 85.01 | 76.65 | 63.40 | 71.03 | 84.12 | 74.65 | 75.09 | 80.56 |
| 18 | MNLI + RS + NC | NC | 74.90 | 60.25 | 90.09 | 75.33 | 78.30 | 96.17 | 89.79 | 91.91 | 76.91 | 68.53 |
| 19 | MNLI + RS + NC | RS | 74.90 | 50.08 | 90.09 | 51.18 | 78.30 | 69.87 | 89.79 | 53.35 | 76.91 | 50.27 |
| 20 | MNLI + RS + NC | MNLI MisM | 74.90 | 64.37 | 90.09 | 75.45 | 78.30 | 69.97 | 89.79 | 75.72 | 76.91 | 80.75 |
| 21 | MNLI + RS + NC | MNLI M | 74.90 | 64.56 | 90.09 | 77.29 | 78.30 | 50.11 | 89.79 | 76.48 | 76.91 | 80.74 |

Table 13. Table shows the train and test set accuracy for all the experiments involving MNLI dataset. Here, L DecAtt and L ESIM refers to the Lambda DecAtt and Lambda ESIM models. NC refers to NER-CHANGED dataset, RS refers to the ROLE-SWITCHED dataset, MNLI MisM refers to MNLI MISMATCHED test set and MNLI M refers to MNLI MATCHED test set. Each row of this table represents an experiment. The Second and Third columns of each row represents the train set and the test set used for that experiment. Rest of the columns show the train and the test accuracy (Acc) in percentages for all the five models. In our experiments, we have used the *bert-large-uncased* model.

The Table 14 shows the weights learnt by the Lambda Layer of Lambda DecAtt and Lambda ESIM models for this experiment. Ideally, the weights learnt for the dimension 1, 5 and 9 should be small as this indicates a lower weight being given to the vector similarity and a higher weight being given to the symbolic similarity. But we don't see this here because the examples that could help the Lambda Layer learn

| Weight Vector Dimensions | Dimension Meaning | Lambda DecAtt (Learnt Weight) | Lambda ESIM (Learnt Weight) |
|---|---|---|---|
| 1 | "Names-Names" | **0.5418** | **0.342** |
| 2 | "Names-Dates" | 0.309 | 0.353 |
| 3 | "Names-Num" | 0.2571 | 0.164 |
| 4 | "Names-Others" | 0.713 | 0.374 |
| 5 | "Date-Names" | 0.409 | 0.511 |
| 6 | "Dates-Dates" | **0.359** | **0.287** |
| 7 | "Dates-Num" | 0.374 | 0.466 |
| 8 | "Dates-Others" | 0.566 | 0.350 |
| 9 | "Num-Names" | 0.474 | 0.501 |
| 10 | "Num-Dates" | 0.522 | 0.413 |
| 11 | "Num-Num" | **0.635** | **0.444** |
| 12 | "Num-Others" | 0.522 | 0.351 |
| 13 | "Others-Names" | 0.528 | 0.378 |
| 14 | "Others-Dates" | 0.709 | 0.327 |
| 15 | "Others-Num" | 0.243 | 0.325 |
| 16 | "Others-Others" | **0.869** | **0.372** |

Table 14. Learnt weights by the Lambda Layer for Lambda DecAtt and Lambda ESIM models when trained on SNLI. A higher value indicates more weight being given to Vector Similarity, while a smaller value indicates more weight being given to Symbolic Similarity.

such weights are absent during train time. This leads to the Lambda Layer learn mostly junk values which in turn leads to the poor performance of Lambda DecAtt and Lambda ESIM models on the "NER Changed" dataset.

The corresponding weights learnt by the Lambda layer of Lambda ESIM model are smaller than that of the Lambda DecAtt. This is mainly due to the difference in the weight assigned to the symbolic similarity. As described in the previous chapter the weight assigned when two strings are equal is a hyper-parameter. All the experiments in this work use a value of '30' and '200' for this hyper-parameter in the Lambda DecAtt and Lambda ESIM models respectively. This is the reason why a value of 0.372 is learnt by the lambda layer of Lamba ESIM model for dimension 16. This value is high enough to provide a high weight for Vector similarity. In contrast, the

lambda layer of Lambda DecAtt model learns a value of 0.869 for dimension 16 which eventually provide a high weight for Vector similarity over Symbolic similarity.

## 5.4 Experiment 4 and 5

Here the 5 models are trained on SNLI and "NER Changed" train set. As shown in the Table 12 the three existing models did not perform well as compared with Lambda DecAtt and Lambda ESIM models. The Lambda Models also perform equally well on SNLI dataset as compared to their unmodified counterparts.

Although the examples that require understanding of "Entities" are shown at train time, the attention mechanism of DecAtt (Parikh et al. 2016) and ESIM (Chen et al. 2016) models is unable to capture the notion of "Entities". While the Lambda DecAtt and Lambda ESIM models perform extremely well giving **97.81**% and **99.13**% accuracy. This shows how well the modification in attention mechanism proposed in this work adapts and learn to understand the notion of "Entities".

The Table 15 shows the weights learnt by the Lambda Layer of Lambda DecAtt and Lambda ESIM models for this experiment. Consider the learnt weight for dimension 1 that represents a pair of words from premise and hypothesis being "Names" is extremely small(-0.022 and 0.138 for Lambda DecAtt and Lambda ESIM respectively). This indicates and explains that the two models have learnt to give more weight to Symbolic Similarity for such a case while calculating the attention for inference. It also learns to give more weight to Vector Similarity in case of when for example the pair of words are not a Named Entity as shown by the high value learnt for dimension 16.

| Weight Vector Dimensions | Dimension Meaning | Lambda DecAtt (Learnt Weight) | Lambda ESIM (Learnt Weight) |
|---|---|---|---|
| 1 | "Names-Names" | **-0.022** | **0.138** |
| 2 | "Names-Dates" | 0.312 | 0.040 |
| 3 | "Names-Num" | 0.483 | 0.282 |
| 4 | "Names-Others" | 0.616 | 0.513 |
| 5 | "Date-Names" | 0.439 | 0.098 |
| 6 | "Dates-Dates" | **-0.032** | **-0.123** |
| 7 | "Dates-Num" | 0.470 | 0.330 |
| 8 | "Dates-Others" | 0.715 | 0.525 |
| 9 | "Num-Names" | 0.484 | 0.296 |
| 10 | "Num-Dates" | 0.400 | 0.144 |
| 11 | "Num-Num" | **0.310** | **0.394** |
| 12 | "Num-Others" | 0.558 | 0.478 |
| 13 | "Others-Names" | 0.607 | 0.393 |
| 14 | "Others-Dates" | 0.690 | 0.465 |
| 15 | "Others-Num" | 0.468 | 0.302 |
| 16 | "Others-Others" | **0.811** | **0.451** |

Table 15. Learnt weights by the Lambda Layer for Lambda DecAtt and Lambda ESIM models when trained on SNLI and "NER Changed". A higher value indicates more weight being given to Vector Similarity, while a smaller value indicates more weight being given to Symbolic Similarity.

### 5.5 Experiment 6 and 7

Here the 5 models are trained on SNLI and "Roles-Switched" train set. Here we observe that the DecAtt and Lambda DecAtt models perform poorly while ESIM and Lambda ESIM captures the change in "Roles" between the premise and hypothesis. This behavour can be attributed to the use of BiLSTMs for encoding the input in ESIM models. The DecAtt models lack this and thus are not able to capture the sequence information which is important to capture the difference in the "Roles" played by certain entities in the premise and hypothesis.

## 5.6 Experiment 8, 9 and 10

Here the 5 models are trained on SNLI, "NER Changed" and "Roles-Switched" train set. We observe that among the 5 models, Lambda ESIM performs the best for both 'NER Changed" and "Roles-Switched" test sets. Lambda DecAtt is able to capture the notion of "Entities" but fails to understand "Roles". As mentioned in the previous sections, this can be attributed to the lack of BiLSTMs while encoding the input for DecAtt. A BiLSTM transformation is performed over the input embedding by both the original ESIM and our Lambda ESIM model. The original DecAtt and our Lambda DecAtt model does not perform such a transformation. This could be the reason behind their poor performance on the ROLE-SWITCHED test set.

Both the Lambda DecAtt and Lambda ESIM models also perform better on SNLI as compared to the original DecAtt and ESIM models. BERT although performs slightly better on SNLI as compared to the Lambda models, it fails miserably on the "NER Changed" and "Roles-Switched" test sets. This shows the inability of BERT to capture the subtle differences in "Entities" and "Roles".

The Table 16 shows the weights learnt by the Lambda Layer of Lambda DecAtt and Lambda ESIM models for this experiment. The Lambda weight learnt for dimension 1 is really small(-0.00005 and 0.2572 for Lambda DecAtt and Lambda ESIM respectively). Since this dimension represents that the pair of words from premise and hypothesis are names, small value here suggests the model learns to give more weight to Symbolic Similarity while calculating the attention. This explains how the Lambda models are able to capture the notion of "Entities" and outperform their unmodified counterparts and the BERT model. Tables 17 and 18 shows some examples for which the Original ESIM predicts wrongly and the Lambda ESIM model predicts correctly.

| Weight Vector Dimensions | Dimension Meaning | Lambda DecAtt (Learnt Weight) | Lambda ESIM (Learnt Weight) |
|---|---|---|---|
| 1 | "Names-Names" | **-0.00005** | **0.2572** |
| 2 | "Names-Dates" | 0.2609 | 0.2583 |
| 3 | "Names-Num" | 0.4158 | 0.1885 |
| 4 | "Names-Others" | 0.5019 | 0.5026 |
| 5 | "Date-Names" | 0.2962 | 0.2639 |
| 6 | "Dates-Dates" | **0.2293** | **-0.0361** |
| 7 | "Dates-Num" | 0.4192 | 0.3815 |
| 8 | "Dates-Others" | 0.6136 | 0.5796 |
| 9 | "Num-Names" | 0.4378 | 0.1209 |
| 10 | "Num-Dates" | 0.4451 | 0.2182 |
| 11 | "Num-Num" | **0.8047** | **0.4585** |
| 12 | "Num-Others" | 0.6540 | 0.5248 |
| 13 | "Others-Names" | 0.5510 | 0.3682 |
| 14 | "Others-Dates" | 0.5589 | 0.4826 |
| 15 | "Others-Num" | 0.4797 | 0.4371 |
| 16 | "Others-Others" | **0.8007** | **0.4446** |

Table 16. Learnt weights by the Lambda Layer for Lambda DecAtt and Lambda ESIM models when trained on SNLI, "NER Changed" and "Roles-Switched". A higher value indicates more weight being given to Vector Similarity, while a smaller value indicates more weight being given to Symbolic Similarity.

### 5.7 Experiment 11 and 12

Here the 5 models are trained on only MNLI train set. As shown in the Table 13, all the models perform poorly on the "NER Changed" and "Roles-Switched" datasets. This behavior is also seen when the models are trained only on the SNLI dataset. The lack of examples that emphasize on capturing notion of "Entities" and "Roles" in the MNLI datastet can be attributed as the reason for such poor performance. Similar to experiment 1 and 2, the Lambda DecAtt and Lambda ESIM models also perform poorly due to the lack of such examples at train time.

The Table 19 shows the weights learnt by the Lambda Layer of Lambda DecAtt and Lambda ESIM models for this experiment. As explained in experiment 1 and 2, a small weight for the dimension 1, 5 and 9 indicates a lower weight being given to

| Premise-Hypothesis Pair | ESIM Entailment Scores | L-ESIM Contradiction Scores |
|---|---|---|
| *premise*: **Gary Johnson** stated the technical possibility of concluding the next phase of the agreement by December 2007.<br>*hypothesis*:**Steve Waugh** stated the technical possibility of concluding the next phase of the agreement by December 2007. | 99.76% | 96.83% |
| *premise*: **Hong Kong** stands to benefit more than most from continued global trade liberalisation as trade is the engine of its growth, accounting for nearly three times its gross domestic product.<br>*hypothesis*: **Melbourne** stands to benefit more than most from continued global trade liberalisation as trade is the engine of its growth, accounting for nearly three times its gross domestic product. | 99.99% | 99.97% |
| *premise*: He was first appointed to the nine-member court by President Nixon in **2002**.<br>*hypothesis*: He was first appointed to the nine-member court by President Nixon in **1976**. | 91.0% | 99.79% |
| *premise*: In response to these challenges, King Mohammed in **1928** launched a National Initiative for Human Development, a $ 2 billion program aimed at alleviating poverty and underdevelopment by expanding electricity to rural areas and replacing urban slums with public and subsidized housing, among other policies.<br>*hypothesis*: In response to these challenges, King Mohammed in **1995** launched a National Initiative for Human Development, a $ 2 billion program aimed at alleviating poverty and underdevelopment by expanding electricity to rural areas and replacing urban slums with public and subsidized housing, among other policies.. | 99.99% | 99.99% |
| *premise*: Bangladeshi officials say worst hit was the southeastern port city of Chittagong, where **16651** people died after many hillside homes were swept away or collapsed under tons of mud.<br>*hypothesis*: Bangladeshi officials say worst hit was the southeastern port city of Chittagong, where **6948** people died after many hillside homes were swept away or collapsed under tons of mud. | 61.52% | 99.99% |
| *premise*: **Twelve** of those injured later died at a hospital.<br>*hypothesis*: **Seventeen** of those injured later died at a hospital. | 64.64% | 99.99% |

Table 17. Sample premise-hypothesis pairs with different named entity where ESIM model gives wrong predictions (Confidence Scores) and Lambda ESIM (L-ESIM) model gives correct predictions.

| Premise-Hypothesis Pair | ESIM Entailment Scores | L-ESIM Contradiction Scores |
|---|---|---|
| *premise*: **Quinn** lost the most important match of his life to **Pat**.<br>*hypothesis*:**Pat** lost the most important match of his life to **Quinn**. | 48.55% | 77.22% |
| *premise*: **Quinn** hangs **Frankie** rendering him dead.<br>*hypothesis*: **Frankie** hangs **Quinn** rendering him dead. | 60.79% | 61.52% |
| *premise*: **Current Prime Minister Stephen Harper** called **Gray** "an honourable parliamentarian who served his country well".<br>*hypothesis*: **Gray** called **current Prime Minister Stephen Harper** "an honourable parliamentarian who served his country well".. | 54.44% | 64.90% |
| *premise*: **India** rejects the accusation, and calls for **Pakistan** to prosecute militants based there for the 2008 Mumbai attacks, which killed over 150 people..<br>*hypothesis*: **Pakistan** rejects the accusation, and calls for **India** to prosecute militants based there for the 2008 Mumbai attacks, which killed over 150 people. | 56.41% | 59.86% |

Table 18. Sample premise-hypothesis pairs with entities playing different roles where ESIM model gives wrong predictions (Confidence Scores) and Lambda ESIM (L-ESIM) model gives correct predictions.

the vector similarity and a higher weight being given to the symbolic similarity. But we don't see this here because the examples that could help the Lambda Layer learn such weights are absent during train time. This leads to the Lambda Layer learn mostly junk values which in turn leads to the poor performance of Lambda DecAtt and Lambda ESIM models on the "NER Changed" dataset.

## 5.8 Experiment 13, 14 and 15

Here the 5 models are trained on MNLI and "NER Changed" train set. As shown in the Table 12 the three existing models did not perform well as compared with

| Weight Vector Dimensions | Dimension Meaning | Lambda DecAtt (Learnt Weight) | Lambda ESIM (Learnt Weight) |
|---|---|---|---|
| 1 | "Names-Names" | **0.509** | **0.309** |
| 2 | "Names-Dates" | 0.497 | 0.288 |
| 3 | "Names-Num" | 0.548 | 0.435 |
| 4 | "Names-Others" | 0.753 | 0.411 |
| 5 | "Date-Names" | 0.534 | 0.267 |
| 6 | "Dates-Dates" | **0.651** | **0.323** |
| 7 | "Dates-Num" | 0.626 | 0.432 |
| 8 | "Dates-Others" | 0.818 | 0.420 |
| 9 | "Num-Names" | 0.537 | 0.347 |
| 10 | "Num-Dates" | 0.701 | 0.420 |
| 11 | "Num-Num" | **0.721** | **0.356** |
| 12 | "Num-Others" | 0.766 | 0.415 |
| 13 | "Others-Names" | 0.654 | 0.353 |
| 14 | "Others-Dates" | 0.670 | 0.350 |
| 15 | "Others-Num" | 0.614 | 0.361 |
| 16 | "Others-Others" | **0.816** | **0.409** |

Table 19. Learnt weights by the Lambda Layer for Lambda DecAtt and Lambda ESIM models when trained on SNLI. A higher value indicates more weight being given to Vector Similarity, while a smaller value indicates more weight being given to Symbolic Similarity.

Lambda DecAtt and Lambda ESIM models. The Lambda Models provide comparable performance on MNLI test set as compared to their unmodified counterparts.

In this experiment, we observed similar behavior to experiment 3 and 4. Although the examples that require understanding of "Entities" are shown at train time, the attention mechanism of DecAtt (Parikh et al. 2016) and ESIM (Chen et al. 2016) models is unable to capture the notion of "Entities". While the Lambda DecAtt and Lambda ESIM models perform extremely well giving **97.94**% and **95.12**% accuracy. This shows how well the modification in attention mechanism proposed in this work adapts and learn to understand the notion of "Entities".

## 5.9 Experiment 16 , 17 and 18

Here the 5 models are trained on MNLI and "Roles-Switched" train set. Here we observe that the performance on the "Roles-Switched" test set is always significantly better when combining the "Roles-Switched" train set with the SNLI train set instead of MNLI train set.

## 5.10 Experiment 19 , 20, 21 and 22

Here the 5 models are trained on MNLI, "NER Changed" and "Roles-Switched" train set. Here we observe that the performance on the "NER Changed" and "Roles-Switched" test set is always significantly better when combining these two train set with the SNLI train set instead of MNLI train set.

# Chapter 6

# CONCLUSION AND FUTURE DIRECTION

In this chapter we discuss the conclusion (Section 6.1) and the directions to take in the future. Section 6.2.1 details how FrameNet (Section 6.2.1.1) and CoNLL 2005 (Section 6.2.1.2) can be used to expand and enrich the "Role Switched" dataset. Section 6.2.2 describes how the proposed attention mechanism can be optimized (Section 6.2.2.1) and suggests a non-trivial task of improving BERT (Section 6.2.2.2). Section 6.2.3 discusses the scope and importance of Ablation Study for this work.

## 6.1 Conclusion

This works shows how to use the existing annotated corpora to generate NLI datasets that emphasize on capturing the notion of "Entities" and "Roles". Existing datasets like bAbI, AMR, CoNLL 2003 and GMB were used to automatically create a fully annotated NLI dataset refered to as "NER Changed" dataset. This dataset contained 351.1K labelled *premise-hypothesis* pairs of sentences that emphasize on understanding the notion of "Entities". This work also involved the creation of the "Roles-Switched" dataset. Existing datasets and resources like VerbNet, PropBank, QA-SRL, CoNLL 2004 and CoNLL 2005 were used to create 146.5K labelled *premise-hypothesis* pairs of sentences that emphasize on capturing the notion of "Roles" in order to solve for inference.

Based on these dataset's performance on the existing architectures, this work shows how these architectures perform poorly for simple examples that require under-

standing of "Entities" and "Roles". This work shows the inability of existing attention mechanisms to capture these notions and proposes a novel attention mechanism. The proposed new architecture significantly helps to capture the notion of entities and roles. Furthermore, the performance does not drop on the existing testbeds when the new attention mechanism is used, which shows the generality of the proposed attention mechanism.

The Lambda ESIM model which is the ESIM model with the proposed attention mechanism proves to be the best performing model. When trained on SNLI, "NER Changed" and "Roles-Switched" train sets the Lambda ESIM models achieves a 98.91% accuracy on "NER Changed" test set beating Original ESIM by 13%, Original DecAtt by 18% and BERT by 29%. On the "Roles-Switched" test set this model achieves a 87.92% accuracy beating Original ESIM by 3.5%, Original DecAtt by 37% and BERT by 38%. This model also gives a 87.28% accuracy on SNLI test set beating Original ESIM by 0.2%, Original DecAtt by 3.5%. This is comparable to BERT's SNLI test set accuracy of 88.08%.

## 6.2 Future Direction

There are a few things that were not done as part of this work which leaves room for some future work. There can be more work done with respect to the three major aspects of this work:

- Dataset Generation,
- Models and Approach,
- Ablation Study (Experiment and Analysis)

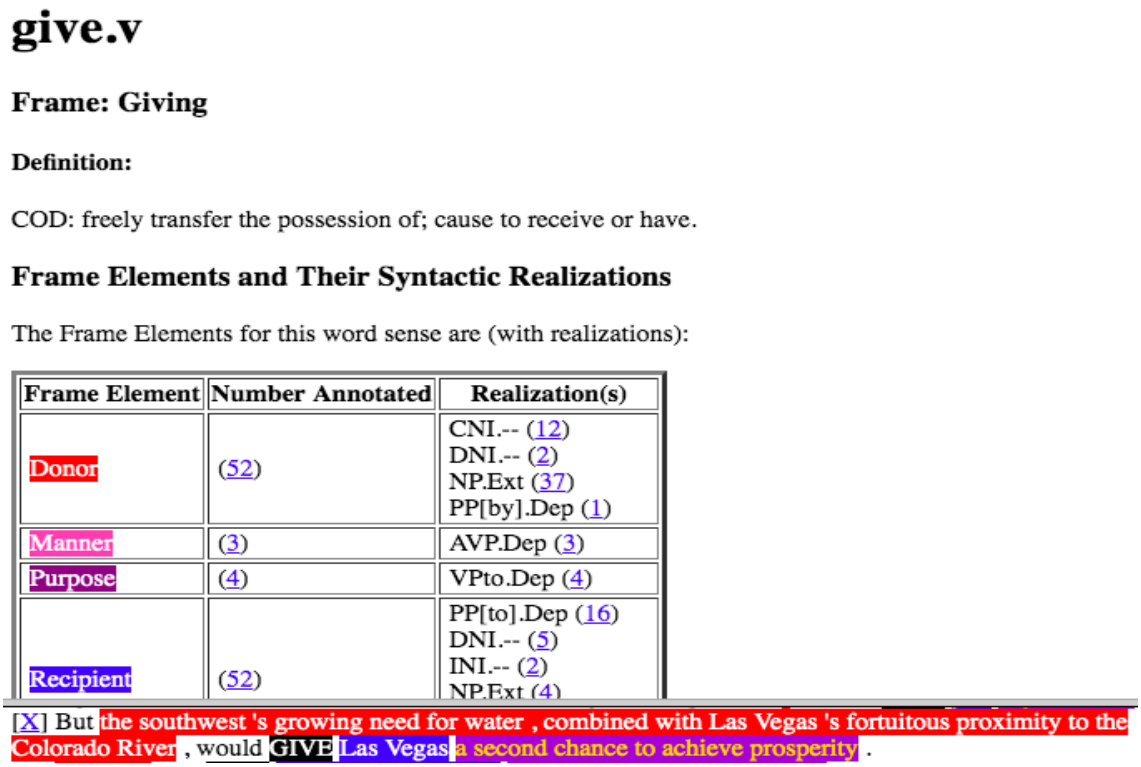
### 6.2.1 Future Direction: Dataset Generation

This section describe what other existing resources can be used to generate more and varied datasets for NLI.

#### 6.2.1.1 Dataset Generation using FrameNet

FrameNet (**Baker:1998:BFP:980451.980860**) can also serve as a good resource for creating more varied kinds of examples for "Role-Switched" dataset. It is a lexical database with more than 200,000 manually annotated sentences. Each of these sentences are linked to more than 1,200 semantic frames which can be understood as a description of a type of event, relation, or entity and the participants in a sentence. Each Frame contain Frame Elements along with their syntactic realizations. The words that evoke a Frame are known as Lexical Units (LUs).

Consider the figure 18. It shows an example of Frame : "Giving" containing the Lexical Unit "give". It contains Frame Elements such as Donor, Recipient etc. The same figure shows an examples for which the Donor is *"the southwest's growing need for water , combined with Las Vegas's fortuitous proximity to the Colorado River"* and the Recipient is *"Las Vegas"*. Switching the Donor and Recipient here will give us a more convoluted Role-Switched pair. In this work, The "Role-Switched" pairs majorly consists of entities being switched, but using FrameNet will give us pairs where phrases and entities both gets switched.



# give.v

**Frame: Giving**

**Definition:**

COD: freely transfer the possession of; cause to receive or have.

**Frame Elements and Their Syntactic Realizations**

The Frame Elements for this word sense are (with realizations):

| Frame Element | Number Annotated | Realization(s) |
|---|---|---|
| Donor | (52) | CNI.-- (12)<br>DNI.-- (2)<br>NP.Ext (37)<br>PP[by].Dep (1) |
| Manner | (3) | AVP.Dep (3) |
| Purpose | (4) | VPto.Dep (4) |
| Recipient | (52) | PP[to].Dep (16)<br>DNI.-- (5)<br>INI.-- (2)<br>NP.Ext (4) |

[X] But the southwest 's growing need for water , combined with Las Vegas 's fortuitous proximity to the Colorado River , would GIVE Las Vegas a second chance to achieve prosperity .

Figure 18. Annotations provided in the FrameNet Lexical Database

#### 6.2.1.2 Dataset Generation using other Semantic Role Labeling Datasets like CoNLL 2005

The Shared Tasks of CoNLL 2005 aims at creating a system that can recognize semantic roles for the English Language based on PropBank predicate-argument structures. Figure 19 shows an example of annotations provided in this corpus. Unlike CoNLL 2004, the 2005 dataset it sufficiently larger and provide complete syntactic trees given by several alternative parsers. Due to the availability of multiple parser's parse trees from multiple parsers it becomes more likely to get more candidate sentences.

If you recall, in this work VerbNet was used to shortlist certain kinds of verbs. Dataset like QA-SRL was queried to find sentences that contain the shortlisted verbs from VerbNet. The same approach appeared inefficientå for the CoNLL 2004 dataset

| | | | | | |
|---|---|---|---|---|---|
| When | - | (AM-TMP*) | * | (AM-TMP* | * |
| Disney | - | (A0*) | (A0*) | * | * |
| offered | offer | (V*) | * | * | * |
| to | - | (A1* | * | * | * |
| pay | pay | * | (V*) | * | * |
| Mr. | - | * | (A2* | * | * |
| Steinberg | - | * | *) | * | * |
| a | - | * | (A3* | * | * |
| premium | - | * | *) | * | * |
| for | - | * | * | * | * |
| his | - | * | (A1* | * | * |
| shares | - | *) | *) | *) | * |
| , | - | * | * | * | * |
| the | - | * | * | (A0* | * |
| New | - | * | * | * | * |
| York | - | * | * | * | * |
| investor | - | * | * | *) | * |
| did | - | * | * | * | * |
| n't | - | * | * | (AM-NEG*) | * |
| demand | demand | * | * | (V*) | * |
| the | - | * | * | (A1* | (A0* |
| company | - | * | * | * | *) |
| also | - | * | * | * | (AM-DIS*) |
| pay | pay | * | * | * | (V*) |
| a | - | * | * | * | (A3* |
| premium | - | * | * | * | *) |
| to | - | * | * | * | * |
| other | - | * | * | * | (A2* |
| shareholders | - | * | * | *) | *) |
| . | - | * | * | * | * |

Figure 19. Annotations provided in the CoNLL 2005 corpus

because although it did contain the sentences with our shortlisted verbs, it did not contain the annotations for many of the verbs in a sentence that could have helped us in identifying the phrases or entities for the roles that verb takes. The reason for this was lack of varying parse trees from different parsers which is not the case in CoNLL 2005.

### 6.2.2 Future Direction: Models and Approach

In this work, a novel attention mechanism is proposed giving rise to a variant of Decomposable Attention Model and Enhanced Sequential Sequential Model known as Lambda Decomposable Attention Model and Lambda Enhanced Sequential Sequential

Model. Other than these models, experiments were also performed on BERT which is the current state of the art model. Considering all this there are two main Future works in this area. The next two subsections will discuss these future works.

#### 6.2.2.1 Optimizing Lambda Layer

This work shows the drawbacks of the existing attention mechanism employed by many existing architectures. To overcome these drawbacks, a novel attention mechanism is proposed which weighs vector and symbolic similarity in order to calculate the new attention weights. An additional layer, known as Lambda Layer is used to learn lambda weights that suggest how to weigh between vector and symbolic similarity.

The input to the Lambda Layer is a 16 Dimensional feature vector. Each dimension of this vector is a pair-wise combination of four types of entities ("Names","Dates", "Numeric" and "Other"). Therefore, this vector requires NER tags to be computed using some NER tagger like Stanford NER Tagger or Spacy NER Tagger. Therefore, the future work here could be to directly use the word vectors as input to the lambda layer. The input word vectors can be concatenated in some form which will increase the input dimensions of the Lambda Layer. This approach will also require to add a couple of extra hidden layers to the lambda layer.

This optimization will come at a cost of explainability of the model. The existing learnt weights of the Lambda Layer layer help in explaining how the decision of made by the model relates to the entity type of the pair of words in question. This explainability will be lost in using the word vectors directly but will provide huge gains in the training time.

#### 6.2.2.2 Making BERT Understand the Notion of "Entities" and "Roles"

The BERT model is the current state of the art architecture on the SNLI and MNLI datasets. At the time of this work, BERT model is considered to be the answer to NLI. In this work, BERT model was used as a part of all the experiments to see if it is able to capture the notion of "Entities" and "Roles". The experiments showed the inability of BERT model. It was failed miserably when understanding of "Entities" and "Roles" was required to make correct inference. Even on exposing the adversarial examples at train time, the BERT model continued to perform poorly.

Since, BERT model performs extremely well with SNLI and MNLI datasers that contain varied generic cases, it will be useful to enable BERT in understanding the "Entities" and "Roles". This will lead to a model excelling not just in the complex cases shown in SNLI and MNLI but even in simple cases as shown in this work.

The feasibility of improving BERT is questionable. A major difference between DecAtt and ESIM as compared to BERT is the kind of similarity used in the attention mechanisms of DecAtt and ESIM as compared to BERT. In DecAtt and ESIM, Dot product vector similarity is used, on the other hand in BERT there many distance and similarity measures like Edit Distance etc are used. A careful analysis of all the measures used needs to be done first in order to understand what needs to be changed at the ground level to enable BERT in capture the notion of "Entities" and "Roles". The other point that needs to be considered is that BERT is a huge model with 24 attention mechanisms stacked on top of each other. This work only tackled DecAtt and ESIM which contain just one attention mechanism. Therefor once the kind of change is figured out, BERT would need to be trained from scratch. Since it is a very heavy model a large amount of computing resources will be required.

### 6.2.3 Future Direction: Ablation Study (Experiment and Analysis)

In this work, all the experiments that were performed included exposing the full train set of "NER Changed" and "Role Switched" datasets. These experiments showed how the existing architectures fail in solving for examples where the notion of "Entities" and "Roles" needs to be captured. These experiments also showed how the proposed change in the attention mechanism can help in empower the models and enable them to capture the notion of "Entities" and "Roles". With all these experiments, there is still room for an Ablation Study. An Ablation Study in machine learning and deep learning refers to removing certain parts or features of a network to better understand the network's behaviour.

In our case, it is really important to understand how much data is needed for these architectures to be able to learn and understand the notion of "Entities" and "Roles". Therefor performing iterations of the same experiments done in this work, with each iteration containing a larger and larger subset of the datasets created in this work could help us in performing this kind of Ablation Study. This could provide us with statistics that could aid in analysing how much data is sufficient for these models to capture the notion of "Entities" and "Roles".

This will be important as the size of existing benchmark datasets are also in the range of 300k-400k pairs of labelled *Premie-Hypothesis* sentences. Therefore training a model by combining the full "NER Changed" and "Role Switched" datasets with the existing benchmark datasets like SNLI or MNLI required significant computing resources and more importantly time. Therefore, it will be useful to see how much data is sufficient as it will help in saving computing resources and time.

# REFERENCES


Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. "Abstract meaning representation for sembanking." In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse,* 178–186.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. "The Groningen Meaning Bank," 463–496. June. doi:10.1007/978-94-024-0881-2_18.

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

———. 2015b. "A large annotated corpus for learning natural language inference." *CoRR* abs/1508.05326. arXiv: 1508.05326. http://arxiv.org/abs/1508.05326.

Bowman, Samuel R., Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. "A Fast Unified Model for Parsing and Sentence Understanding." *CoRR* abs/1603.06021. arXiv: 1603.06021. http://arxiv.org/abs/1603.06021.

Carreras, Xavier, and Llus Màrquez. 2005. "Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling." In *Proceedings of the Ninth Conference on Computational Natural Language Learning,* 152–164. CONLL '05. Ann Arbor, Michigan: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1706543.1706571.

Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. "Enhancing and Combining Sequential and Tree LSTM for Natural Language Inference." *CoRR* abs/1609.06038. arXiv: 1609.06038. http://arxiv.org/abs/1609.06038.

Cheng, Jianpeng, Li Dong, and Mirella Lapata. 2016. "Long Short-Term Memory-Networks for Machine Reading." *CoRR* abs/1601.06733. arXiv: 1601.06733. http://arxiv.org/abs/1601.06733.

Demszky, Dorottya, Kelvin Guu, and Percy Liang. 2018. "Transforming Question Answering Datasets Into Natural Language Inference Datasets." *CoRR* abs/1809.02922. arXiv: 1809.02922. http://arxiv.org/abs/1809.02922.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

FitzGerald, Nicholas, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. "Large-Scale QA-SRL Parsing." *CoRR* abs/1805.05377. arXiv: 1805.05377. http://arxiv.org/abs/1805.05377.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. "Annotation Artifacts in Natural Language Inference Data." *CoRR* abs/1803.02324. arXiv: 1803.02324. http://arxiv.org/abs/1803.02324.

Harabagiu, Sanda, Andrew Hickl, and Finley Lacatusu. 2007. "Satisfying information needs with multi-document summaries." Text Summarization, *Information Processing Management* 43 (6): 1619–1642. doi:https://doi.org/10.1016/j.ipm.2007.01.004.

Kang, Dongyeop, Tushar Khot, Ashish Sabharwal, and Eduard H. Hovy. 2018. "AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples." *CoRR* abs/1805.04680. arXiv: 1805.04680. http://arxiv.org/abs/1805.04680.

Khot, Tushar, Ashish Sabharwal, and Peter Clark. 2018. "SciTail: A Textual Entailment Dataset from Science Question Answering." In *AAAI*.

Liu, Xiaodong, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. "Multi-Task Deep Neural Networks for Natural Language Understanding." *CoRR* abs/1901.11504. arXiv: 1901.11504. http://arxiv.org/abs/1901.11504.

Liu, Yang, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. "Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention." *CoRR* abs/1605.09090. arXiv: 1605.09090. http://arxiv.org/abs/1605.09090.

Mou, Lili, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. "Recognizing Entailment and Contradiction by Tree-based Convolution." *CoRR* abs/1512.08422. arXiv: 1512.08422. http://arxiv.org/abs/1512.08422.

Munkhdalai, Tsendsuren, and Hong Yu. 2016a. "Neural Semantic Encoders." *CoRR* abs/1607.04315. arXiv: 1607.04315. http://arxiv.org/abs/1607.04315.

———. 2016b. "Neural Tree Indexers for Text Understanding." *arXiv e-prints,* arXiv:1607.04492 (July): arXiv:1607.04492. arXiv: 1607.04492 [cs.CL].

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. "The Proposition Bank: An Annotated Corpus of Semantic Roles." *Computational Linguistics* 31 (1): 71–106. doi:10.1162/0891201053630264. eprint: https://doi.org/10.1162/0891201053630264.

Paria, Biswajit, K. M. Annervaz, Ambedkar Dukkipati, Ankush Chatterjee, and Sanjay Podder. 2016. "A Neural Architecture Mimicking Humans End-to-End for Natural Language Inference." *CoRR* abs/1611.04741. arXiv: 1611.04741. http://arxiv.org/abs/1611.04741.

Parikh, Ankur P., Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. "A Decomposable Attention Model for Natural Language Inference." *CoRR* abs/1606.01933. arXiv: 1606.01933. http://arxiv.org/abs/1606.01933.

Point, Tutorial. n.d. "Artificial Intelligence - Overview." https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_overview.

Ratinov, Lev, and Dan Roth. 2009. "Design Challenges and Misconceptions in Named Entity Recognition." In *Proc. of the Conference on Computational Natural Language Learning (CoNLL).* June. http://cogcomp.org/papers/RatinovRo09.pdf.

Rocktäschel, Tim, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2015. "Reasoning about Entailment with Neural Attention." *arXiv e-prints,* arXiv:1509.06664 (September): arXiv:1509.06664. arXiv: 1509.06664 [cs.CL].

Schuler, Karin Kipper. 2005. "Verbnet: A Broad-coverage, Comprehensive Verb Lexicon." AAI3179808. PhD diss.

Sha, Lei, Baobao Chang, Zhifang Sui, and Sujian Li. 2016. "Reading and Thinking: Reread LSTM Unit for Textual Entailment Recognition." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* 2870–2879. Osaka, Japan: The COLING 2016 Organizing Committee, December. https://www.aclweb.org/anthology/C16-1270.

Vendrov, Ivan, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. "Order-Embeddings of Images and Language." *arXiv e-prints,* arXiv:1511.06361 (November): arXiv:1511.06361. arXiv: 1511.06361 [cs.LG].

Wang, Shuohang, and Jing Jiang. 2015. "Learning Natural Language Inference with LSTM." *CoRR* abs/1512.08849. arXiv: 1512.08849. http://arxiv.org/abs/1512.08849.

Weston, Jason, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. "Towards ai-complete question answering: A set of prerequisite toy tasks." *arXiv preprint arXiv:1502.05698.*

Williams, Adina, Nikita Nangia, and Samuel R. Bowman. 2017. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference." *CoRR* abs/1704.05426. arXiv: 1704.05426. http://arxiv.org/abs/1704.05426.

Zhang, Yuan, Jason Baldridge, and Luheng He. 2019. "PAWS: Paraphrase Adversaries from Word Scrambling." *arXiv e-prints,* arXiv:1904.01130 (April): arXiv:1904.01130. arXiv: 1904.01130 [cs.CL].

# APPENDIX A

# CODE AND DATA REPOSITORY

- Google Drive Link to the Data and Saved Model
- Github Repository for the Code

# APPENDIX B

## SNAPSHOT OF "NER CHANGED" DATASET

```
{
  "sentence1": "Iranian President Mohmoud Ahmadinejad said
  thursday Iran is still considering the incentives package.",
  "premise": Iranian President Mohmoud Ahmadinejad said
  thursday Iran is still considering the incentives package.",
  "premiseUF": [[0,0,0,1],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  ,[0,0,0,1],[0,1,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "sentence2": "Iranian President Mohmoud Ahmadinejad said
  monday Iran is still considering the incentives package.",
  "hypothesis": "Iranian President Mohmoud Ahmadinejad said
  monday Iran is still considering the incentives package.",
  "hypothesisUF": [[0,0,0,1],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  ,[0,0,0,1],[0,1,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "gold_label": "contradiction"
}
{
  "sentence1": "Iranian President Mohmoud Ahmadinejad said
  thursday Iran is still considering the incentives package.",
  "premise": "Iranian President Mohmoud Ahmadinejad said
  thursday Iran is still considering the incentives package.",
  "premiseUF": [[0,0,0,1],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  ,[0,0,0,1],[0,1,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "sentence2": "Iranian President Mohmoud Ahmadinejad said
  thursday Iran is still considering the incentives package.",
  "hypothesis": "Iranian President Mohmoud Ahmadinejad said
  thursday Iran is still considering the incentives package.",
  "hypothesisUF": [[0,0,0,1],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  ,[0,0,0,1],[0,1,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "gold_label": "entailment"
}
{
  "sentence1": "Alan Spoon said the population problem
  was one of the major problems currently faced by the
  international community.",
  "premise": "Alan Spoon said the population problem
  was one of the major problems currently faced by the
  international community.",
  "premiseUF": [[1,0,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1],
```

```
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,1,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "sentence2": "Raul Reyes said the population problem
  was one of the major problems currently faced by the
  international community.",
  "hypothesis": "Raul Reyes said the population problem
  was one of the major problems currently faced by the
  international community.",
  "hypothesisUF": [[1,0,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,1,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "gold_label": "contradiction"
}
{
  "sentence1": "Alan Spoon said the population problem
  was one of the major problems currently faced by the
  international community.",
  "premise": "Alan Spoon said the population problem
  was one of the major problems currently faced by the
  international community.",
  "premiseUF": [[1,0,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,1,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "sentence2": "Alan Spoon said the population problem
  was one of the major problems currently faced by the
  international community.",
  "hypothesis": "Alan Spoon said the population problem
  was one of the major problems currently faced by the
  international community.",
  "hypothesisUF": [[1,0,0,0],[1,0,0,0],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,1,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]],
  "gold_label": "entailment"
}
{
  "sentence1": "Washington said North Korea agreed to sampling
  during a meeting with Hill in Septmenber, but Pyongyang
  denies it.",
```

```
  "premise": "Washington said North Korea agreed to sampling
  during a meeting with Hill in Septmenber, but Pyongyang
  denies it.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[1,0,0,0],[0,0,0,1],[0,1,0,0],
  [0,0,0,1],[1,0,0,0],[0,0,0,1],[0,0,0,1]],
  "sentence2": "Washington said North Korea agreed to sampling
  during a meeting with Hill in March, but Pyongyang
  denies it.",
  "hypothesis": "Washington said North Korea agreed to sampling
  during a meeting with Hill in March, but Pyongyang
  denies it.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[1,0,0,0],[0,0,0,1],[0,1,0,0],
  [0,0,0,1],[1,0,0,0],[0,0,0,1],[0,0,0,1]],
  "gold_label": "contradiction"
}
{
  "sentence1": "Washington said North Korea agreed to sampling
  during a meeting with Hill in Septmenber, but Pyongyang
  denies it.",
  "premise": "Washington said North Korea agreed to sampling
  during a meeting with Hill in Septmenber, but Pyongyang
  denies it.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[1,0,0,0],[0,0,0,1],[0,1,0,0],
  [0,0,0,1],[1,0,0,0],[0,0,0,1],[0,0,0,1]],
  "sentence2": "Washington said North Korea agreed to sampling
  during a meeting with Hill in Septmenber, but Pyongyang
  denies it.",
  "hypothesis": "Washington said North Korea agreed to sampling
  during a meeting with Hill in Septmenber, but Pyongyang
  denies it.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[1,0,0,0]
  [0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1],[0,0,0,1]
  ,[0,0,0,1],[0,0,0,1],[1,0,0,0],[0,0,0,1],[0,1,0,0],
  [0,0,0,1],[1,0,0,0],[0,0,0,1],[0,0,0,1]],
  "gold_label": "entailment"
}
```

# APPENDIX C

## SNAPSHOT OF "ROLE-SWITCHED" DATASET

```
{
  "sentence1": "Peyton slaughtered Skyler without any fear.",
  "premise": "Peyton slaughtered Skyler without any fear.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1]],
  "sentence2": "Skyler slaughtered Peyton without any fear.",
  "hypothesis": "Skyler slaughtered Peyton without any fear.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
     [0,0,0,1],[0,0,0,1]],
  "gold_label": "contradiction"
}
{
  "sentence1": "Peyton slaughtered Skyler without any fear.",
  "premise": "Peyton slaughtered Skyler without any fear.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1]],
  "sentence2": "Peyton slaughtered Skyler without any fear.",
  "hypothesis":"Peyton slaughtered Skyler without any fear.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
  [0,0,0,1],[0,0,0,1]],
  "gold_label": "entailment"
}
{
  "sentence1": "Kendall retaliated against Skyler.",
  "premise": "Kendall retaliated against Skyler.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[0,0,0,1],[1,0,0,0]],
  "sentence2": "Skyler retaliated against Kendall.",
  "hypothesis": "Skyler retaliated against Kendall.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[0,0,0,1],[1,0,0,0]],
  "gold_label": "contradiction"
}
{
  "sentence1": "Kendall retaliated against Skyler.",
  "premise": "Kendall retaliated against Skyler.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[0,0,0,1],[1,0,0,0]],
  "sentence2": "Kendall retaliated against Skyler.",
  "hypothesis": "Kendall retaliated against Skyler.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[0,0,0,1],[1,0,0,0]],
  "gold_label": "entailment"
}
{
  "sentence1": "Peyton drove Skyler to home.",
```

```
  "premise": "Peyton drove Skyler to home.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
  [0,0,0,1]],
  "sentence2": "Skyler drove Peyton to home.",
  "hypothesis": "Skyler drove Peyton to home.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
     [0,0,0,1],
  "gold_label": "contradiction"
}
{
  "sentence1": "Peyton drove Skyler to home.",
  "premise": "Peyton drove Skyler to home.",
  "premiseUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
  [0,0,0,1]],
  "sentence2": "Peyton drove Skyler to home.",
  "hypothesis": "Peyton drove Skyler to home.",
  "hypothesisUF": [[1,0,0,0],[0,0,0,1],[1,0,0,0],[0,0,0,1],
     [0,0,0,1],
  "gold_label": "entailment"
}
```