NL2code

Sai Krishna Karanam karanam.s@husky.neu.edu
Nischal Mahaveer Chand mahaveerchand.n@husky.neu.edu
Varun Sundar Rabindranath rabindranath.v@husky.neu.edu

I. Introduction

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

II. Related Word

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

III. Dataset

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

IV. Models

Building upon Pengcheng's model, we build 4 models, each encodes the input in a slightly different way. First we start by briefly describing Pengcheng's model.

A. Pengcheng's model

Pengcheng recognized that adding syntax information to the model should give better prediction. The model takes as input the raw comment and generates an AST of the corresponding code. The model is an Encoder-Decoder model, with the decoder doing most of the heavy lifting. The encoder comprises of an embedding layer, which produces token specific embeddings, that are feed into a Bidirectional LSTM (BiLSTM). The output of this BiLSTM layer is a query embedding, which is passed to the decoder as input.

The decoder is slightly complicated and works in mysterious ways! Lord Voldomort himself blessed it with his divine wand to produce a magical black box, that generates AST's!

As described, the decoder is already at a state-of-the-art level, and needs no further modifications. All models described hereon use the same decoder architecture as Pengcheng's model.

B. Basic Concat (BC)

Our first model - Basic Concat (BC) model, concatenates the POS tag and phrase tag of the current timestep to the corresponding token embedding; Essentially expanding the input from l to (l + 2) dimensions.

C. Linear Projection (LP)

After experimenting with BC, we decided to embed the POS and Phrase tags in the same way the tokens are embedded, then perform a linear project on the new embedding, like in BiLSTM.

D. Linear Projection Reduced Dimension (LPrd)

We noticed that the embedding dimensions for POS embedding and Phrase embedding are very large for their relatively small vocabulary sizes (14 and 44 respectively). We decided to reduce the embedding dimensions to 8 and 32.

E. Raw Query Independent Preprojection (AdvLP)

Rather then performing a linear projection on the three embeddings together, we tried to preproject the POS embedding and phrase embedding to create a new embedding, augmentation embedding, which we then concatenate with the token embedding and performing another

linear projection to get the final token embeddings. We hypothesized that this would give us better results as we do try to retail as much information as possible from the token embedding.

V. Results

Models		Metrics	
		BLUE	Accu.
Pengcheng		1	1
Base		3	1
NL2code	BC	1	1
	LP	9	1
	LPrd	1	1
	AdvLP	1	1

VI. Conclusion

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.