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

A. Standard Datasets

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.

1) Django: 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.

2) HS: 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.

B. Our 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!

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.

B. Our models

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 with modifications to the encoder and the input data. All models are trained and tested using the dataset decribed in SECTION.

1) Basic Concat (BC): For our first attemp to incorporate syntax information into the encoder, we decided to add (append) the POS and phrase ID of each token to the corresponding token embedding, giving us the Augmented Token Embedding (ATE). The ATE is then feed into a modified BiLSTM layer that took 130 dimension embeddings, rather than the specified 128 dimensions.

Token embedding dimensions: 128
POS and Phrase ID dimensions: 1 each; total 2
Augmented token embedding dimensions: 130

2) Linear Projection (LP): To add some syntactic information over a sequence of tokens, we used an embedding layer for POS and phrase tags. The resulting TE, POS embedding (POSE), and Phrase embedding (PhE) are then concatenated to produce the ATE; which is a (128 * 3) dimension vector. We then apply a linear projection (using a dense layer with (128 * 3) input nodes, 128 output nodes, and the linear actiation function). This new vector is passed to the BiLSTM of Pengcheng's model.

Token embedding dimension: 128
POS embedding dimension: 128
Phrase embedding dimension: 128
Dense = (128 * 3) input nodes, 128 output nodes
ATE = [TE : POSE : PhE] (: is concatenation)
ATE' = Dense(ATE)
QE = BiLSTM(ATE')

3) Linear Projection Reduced Dimension (LPrd): Subsequently, we noticed that the POS and Phrase vocabulary sizes were relatively smaller than token vocabulary size. To avoid redundancy XXX, we changed the embedding dimensions of POSE and PhE to 8 and 32 respectively. The process described in LP is then repeated.

Token embedding dimension: 128
POS embedding dimension: 8
Phrase embedding dimension: 32
Dense = (128 + 8 + 32) input nodes, 128 output nodes
ATE = [TE : POSE : PhE] (: is concatenation)
ATE' = Dense(ATE)
Query Embedding = BiLSTM(ATE')

4) Raw Query Independent Preprojection (AdvLP): Rather than applying one linear projection on ATE, we apply two here, where the first is independent of the input query.

Token embedding dimension: 128
POS embedding dimension: 128
Phrase embedding dimension: 128
preprojector = (128 * 2) input nodes, 128 output nodes
Dense = (128 * 3) input nodes, 128 output nodes
AE = [POSE : PhE]
PAE = Preprojector(AE)
ATE = [TE : PAE]
ATE' = Dense(ATE)
QE = BiLSTM(ATE')

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.

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.