Grammar error correction using deep networks.

Supervisor:  Dr. Lizhen Qu
Student:    Kiarie Ndegwa
Presentation overview:

A. Motivation.
B. Problem definition.
C. Background.
D. Related work and benchmarks.
E. Implementation overview.
F. Results.
G. Conclusion.
H. Future work.
What I want you to take home from this talk.

• Recognize main stream uses of deep learning technologies in natural language processing tasks.

• Understand the problem being attempted to be solved.

• Learn what the basic tool kit of deep learning algorithms are in natural language processing.

• Understand the architecture implemented during this project.
Language is really hard to analytically model and models are really hard to use effectively. Just ask Naom Chomsky.
Collocation errors are one of the hardest grammar error types out there.

kingdom come
every which way
by and large
in short
kick the bucket
spill the beans
attorney general
congressman at large
San Francisco etc

There as many, as there are lexical terms within the English language!
Consistently one of the lowest scoring grammar corrections in NLP challenges.

Most importantly, one of the hardest aspects for ESL speakers to grasp.
Can we solve collocation errors using deep networks?
NLP Deep networks are already in mass use

- Google’s newest translation algorithm
- Google’s smart reply.
- Microsoft transcript algorithm
- Personal assistants.
- Genome sequencing
- Speech synthesis
Background

Underlying components of these applications

Language models using word embeddings

[... 0.2, 0.3, 0.6, 0.9, 0.5, 0.4, 0.6 ... ]
Underlying components of these applications

Recurrent neural networks

Bi-directional recurrent networks.

Sequence to sequence recurrent networks
Attention mechanisms, the new kid on the block
Related work and benchmarks

Final explored solutions

• CAMB winners of the CONLL2014 shared task, achieving a f0.5 score of 15.4 on collocation error correction:

• Recent results from another team at Cambridge University. Work published in June 2016.
Sequence to sequence LSTM network with one of or all the options investigated and optimized, using perplexity as a metric:

- Global soft attention
- Bi-directional network.
- Using various sized Glove language models.
- LSTM
- Reversing sentences
Final explored solutions

Cambridge equivalent architecture:
- Bi-directional neural network.
- Dropout 0.2.
- ADAM optimization (hyper-parameter set to 0.2)
- 200d hidden layer.
- 2 hidden layers on encoder and decoder.

This project’s architecture:
- Global soft attention on a bi-directional neural network.
- Dropout 0.3
- SGD (no momentum) optimization
- 2 hidden layers on encoder and decoder.
- 1000d hidden layer
Data sets employed

**NUCLE data set stats:**
- 5305 collocation errors, 11% of errors in data set.
- 7190 vocabulary size
- Holdout cross validation:
  - 4190 training source–target pairs
  - 3675 training set,
  - 408 validation,
  - 107 test set

**Brown data set stats:**
- 50,007 training source–target pairs
- 1,000,000 vocabulary size
**Training times and general stats**

**This project's algorithm:**
- Number of weights: -- 50M+ weights
- Training times: -- Roughly 10 hrs
- Perplexity on NUCLE2014: -- 5.5
- Pre-training perplexity on Brown -- Training times: 24hrs -- Perplexity: 1.03

**Cambridge Uni algorithm:**
- Number of weights: -90M+ weights
- Training times: -- roughly 18 hrs
- Perplexity on NUCLE2014: --7.4
- Pre-training perplexity on Brown -- Training times: 17hrs -- Perplexity: 1.13
<table>
<thead>
<tr>
<th>Incorrect text</th>
<th>'Corrected' text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therefore, it would have never met the KLM.</td>
<td>Therefore, it would have never aided the authority.</td>
</tr>
<tr>
<td>personal one, regardless of the genetic testing result of another family member.</td>
<td>personal one, regardless of the genetic testing result of another family articles.</td>
</tr>
<tr>
<td>Contrarily, the main source of energy which ...</td>
<td>Consequently, the main source of energy which ...</td>
</tr>
<tr>
<td>To certain extent, ...</td>
<td>To a certain extent, ...</td>
</tr>
<tr>
<td>... a wild debate on the banning of the use of surveillance ...</td>
<td>... a difficult debate on the effectiveness of the use of surveillance ...</td>
</tr>
<tr>
<td>... most of our electricity was generated from coal, fuel or oil refineries.</td>
<td>... most of our electricity was generated from coal, fuel or oil simultaneously.</td>
</tr>
<tr>
<td>... as compared to other corresponding energy source.</td>
<td>... as compared to other insufficient energy source.</td>
</tr>
<tr>
<td>... more research should be carried out on the nuclear power plant.</td>
<td>... more research should be implemented on the nuclear power plant.</td>
</tr>
<tr>
<td>... no matter what the sequence of innovation-development process ...</td>
<td>... no matter what the intention of innovation-development process ...</td>
</tr>
</tbody>
</table>
Project recap

- Tested attention scheme
- Cleaned up a tricky data set.
- Made easy to understand pre-processing package
- Learnt the basics of Torch
- Learnt how to use the BRAGG super cluster CSIRO computational resources.
Future work

• Implement pre-training process successfully
• Language model re-ranker during beam search phase on decoder side
• Larger data sets, such as the FCE data set
I CAN HAZ QUESTIONS?


