IBM Czech Republic
V Parku 2294/4,
Praha 4 - Chodov

Student Affairs Department
Charles University in Prague
Faculty of Mathematics and Physics
Ke Karlovu 3, Praha 2

Review of the Doctoral Thesis of Mgr. Martin Popel

The Charles University in Prague has invited me to act as an opponent in the Doctoral Thesis defense of Mgr. Martin Popel. In this letter, I state my opinion on his Thesis "Machine Translation Using Syntactic Analysis".

The Thesis is divided into seven chapters. Chapter 1 sketches up the structure of the Thesis, describes motivation and goals. Chapter 2 describes the data, methods of evaluation of results, briefly overviews historical approaches to machine translation. Chapter 3 describes improvements done in the field of tectogrammatical MT. Chapter 4 summarizes author's experience in training NMT Transformer. Chapter 5 describes techniques used to build the best performing En-Cs NMT system, including the innovative approach to using backtranslated data and checkpoint averaging. Chapter 6 contains final evaluation and discussion of results. Chapter 7 concludes the work presented in the Thesis and suggests future research directions. Formal qualities of the Thesis are exceptional. The work is well structured, clearly and concisely written, it has sufficient extent of 160 pages, previous work is properly referenced, author's contribution is clearly stated in an appendix. Experiments are described in sufficient level of detail, properly evaluated and discussed. In addition, epigraphs set at the beginning of each chapter are entertaining.

The focus of the Thesis is split into three MT topics: TectoMT, NMT Transformer, and Backtranslation and Checkpoint Averaging.

The **TectoMT** part introduces the baseline system and describes author's contributions to various components, such as dependency parser influence on translation quality, rule-based block for clitics reordering, context-sensitive translation models, as well as adaptation to new language pairs. This chapter is most true to the Thesis' title.

The **NMT Transformer** part describes author's experiments with the Tensor2Tensor (T2T) implementation of [Vaswani et al., 2017]. The author examines number of metaparameters, such as training data size, model size, maximum sentence length, batch size, learning rate, warmup steps, number of GPU, checkpoint averaging. The role of each parameter is thoroughly explored in a separate subsection containing empirical evidence, discussion, and author's recommendations at the end. This chapter can be viewed as a very useful tutorial for NMT Transformer training.

The **Backtranslation and Checkpoint Averaging** chapter describes novel technique for using monolingual data. Generating synthetic parallel data by backtranslating a monolingual corpus is a common technique. While previous approaches train on a mix of original and synthetic data, the novel method *concatenates* blocks of data of the same type. In addition, the author

finds synergy between concatenation and checkpoint averaging. Additional techniques used to improve translation accuracy are described as well.

I am attaching a list of remarks which occurred to me.

1) The main strength of this work is the usefulness for the community as an NMT cookbook. In this perspective, the Transformer introduction seems too brief, deserves an extension, possibly with figures, so that readers do not need to reach for the original literature. More explanation should be added to such metaparameters that are specific to the actual T2T implementation.
2) Transformer outperforms older RNN auto-regressive approaches. Using a more modern and powerful approach was a wise decision for the evaluation competition. The scientific community would, however, benefit from a back-to-back comparison if carried out with the level of detail presented in this work.
3) I missed characteristics relevant to the deployment (decoding speed, etc.).
4) Model ensembles are often used to improve translation accuracy. This work uses checkpoint averaging (Sections 4.3.10 and 5.5.5). It would be interesting to compare and/or combine these two techniques.
5) There seems to be a contradiction (the author also finds it surprising) between the Figure 4.6 that shows no difference for runs with batches between 1450 and 2000, and Section 4.3.7 that shows Time Until Score improving with growing number of GPUs (equivalent to the effective batch size multiplication). This leaves a reader thinking that either the Figure 4.6 is not correct, or that batch sizes 12k should have been explored (not enough memory), or that multi vs. single GPU computations are not equivalent for the same effective batch sizes. Surprising results should be explained with additional investigation.
   Are the results reproducible? One always gets a different result due to random factors such as initialization and/or data shuffling. It is a common practice to train several candidate models with different seed, and pick the best performing one. Would the Figure 4.6 change had it displayed best out of 5 or 10 runs?
6) My understanding is that the Section 5.3.4 Data Block Size Effect describes different auth:synth ratios. What about the effect of the slice length after which we switch auth/synth? E.g. the *mix* regime uses the shortest possible slices, *concat* regimes slices contain all auth/synth data. What about configurations in between?

I hope that some of the comments above will be addressed during the defense. I believe that the exploration described in this work is sound, the work has enough substance, the proposed method is properly implemented and evaluated. In addition, I am familiar with author's academic work, and his successes in the Czech-English WMT evaluation track.

I believe that the Thesis is a novel contribution to the field of machine translation and that it clearly demonstrates the author's ability to conduct research independently and to present its results.

In Prague, August 30th, 2018

Martin Čmejrek
IBM Watson