MAMA AI
Revoluční 764/17
Praha 1 - Old Town

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

Review of the Doctoral Thesis of Mgr. Ivana Kvapilíková

The Charles University in Prague has invited me to act as an opponent in the Doctoral Thesis defense of Mgr. Ivana Kvapilíková. In this letter, I state my opinion on her Thesis "Towards Machine Translation Based on Monolingual Texts".

The Thesis is divided into eight chapters. Chapter 1 sketches up the structure of the Thesis, describes motivation and goals. Chapter 2 describes the data, describes world languages from the perspective of resource richness as well as training data sources, and defines the extent of the work. Chapter 3 describes fundamentals of NLP techniques in semantic embeddings, transformer language modeling, machine translation. Chapter 4 overviews existing approaches to unsupervised machine translation from the model and data perspectives. Chapter 5 describes the field of parallel corpus mining. Chapter 6 overviews unsupervised machine translation methodology. Chapter 7 brings a list of 6 groups of experiments, each group structured into its data - method - experiment - discussion subsections. Chapter 8 contains final evaluation and discussion of results, and suggests future research directions.

Formal qualities of the Thesis are notably pleasing. The work is well structured, clearly and concisely written, it has sufficient extent of 154 pages, previous work is properly referenced, author's contribution is clearly stated. Figures taken over from other publications are properly sourced. Experiments are described in sufficient level of detail, properly evaluated and discussed.

The focus of the Thesis is split into two directions: (1) methods of obtaining parallel data if there is no parallel data, (2) unsupervised machine translation models and training strategies.

The three (the third one not explicitly claimed) main contributions of the work to the field of UMT seem to be the following:

1. A **novel** method for obtaining training data using a synthetic corpus of 10k parallel sentences obtained by UMT to fine tune XLM cross-lingual LM (5.3). The method brings significant and stable improvement on BUCC, News corpus mining tasks and corpus deshuffling tasks.
2. A **novel** combination of Masked LM + Denoising Autoencoding in UMT pre-training (7.3). The method improves translation quality in many low-resource translation tasks.
3. A **practical cookbook and a diary** of training recipes collected over several years of active competition in WMT evaluation campaigns.

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

1) A formal one: Reviewer's eyes would appreciate more detailed and explicit declarations of the novel contributions to the field, preferably in the earlier sections of the thesis.

2) Some questions arose while reading the experimental part:
    a. 7.1.: Was the discussion of results divergence sufficient? Is it what others observe as well when doing so many iterations?
    b. 7.3.3.: Stable improvement from the novel method thested, except for EN-KK in both directions. Despite the remedy offered in 7.4, a discussion of the root cause would be appreciated.
    c. Section 7.1 looks different from 7.2. and the following subsections with Figures and captions taken over from WMT reports, but this is mainly a formalistic nibble.

3) Except for using GPT as one of translation baselines there are no mentions of LLMs. What is the relevance of LLMs for the UMT field? For example using the zero shot capabilities for the UMT usecase. https://arxiv.org/pdf/2306.11372.pdf.

4) What is the impact of the work on current practical usecases, such as improving LLMs performance on low-resource languages?

5) A single github repository would be a nice entry point for anyone aiming at reproducing the results.

I hope that some of the comments above will be addressed during the defense. However, the exploration described in this work is sound, the work has enough substance, proposed methods are properly implemented and evaluated.

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

In Prague, January 28th, 2024

Martin Čmejrek
MAMA AI