University of Zurich^UZH

Institute of Computational Linguistics

Andreasstrasse 15
CH-8050 Zürich-Oerlikon
Switzerland

Prof. Dr. Martin Volk
Institute of
Computational Linguistics

Phone +41 44 635 43 17
volk@cl.uzh.ch
https://www.cl.uzh.ch

Andreas Schaufelbühl
Binzmühlestrasse 14
8050 Zürich

Matrikel-Nr. 12-918-843
andreas.schaufelbuehl@uzh.com

April 10, 2019

Master's Thesis Specification

Morphological Inflection of Terminology for Constrained Neural Machine Translation

Introduction

Machine translation (MT) plays an important role in industrial settings where large volumes of text need to be translated with limited time and budget, as it has the potential to make manual translation more efficient (**??**). The usefulness of machine translated text, however, is strictly bound to good translation quality. Consequently, a lot of effort is invested in improving the state-of-the-art of MT—most recently by investigating models based on neural networks (neural machine translation, NMT)—, with the goal of producing naturally-sounding, accurate translations (**???**).

While NMT systems now achieve remarkable translation quality for well-resourced languages (e.g., **??**), including specific, user-defined words in the output is an open problem. Consider the translation of the following English sentence into German, taken from the product description of a vacuum cleaner:[1]

> The floorhead is easy to manoeuvre and you always achieve optimal cleaning results.

The online translation system DeepL[2] translates the English word "floorhead" into "Bodenkopf". To avoid such mistranslations (even in manual translation), many companies maintain so-called terminology databases. These databases contain translations of single words or multi-word expressions in their basic form, such as:

> floorhead → Bodendüse

? propose a method to integrate such terminology into MT, essentially constraining the model to include the target side (here: "Bodendüse") as-is in the sequence of translated output words. The method is computationally expensive and thus slow in practice, and **?** present an alternative approach to reduce translation time at comparable quality.

[1] `https://www.miele.in/domestic/1785.htm?info=200046164-ZST`
[2] `https://www.deepl.com`

However, consider the case where a translated term needs morphological inflection:

> The floorheads are easy to manoeuvre and you always achieve optimal cleaning results.

Since terminology databases do not usually contain inflected forms or morphological information (such as "Bodendüse" is a noun), and since constrained decoding methods (**??**) place translated terminology without any modification in the output, the resulting translations may be grammatically incorrect:

> Die Bodendüse sind leicht zu manövrieren und Sie erzielen immer optimale Reinigungsergebnisse.

The goals of this master's thesis

We investigate methods for including lexical constraints (i.e., user-defined terminology) with correct morphological inflection in NMT. The research questions that guide this thesis are:

- *RQ1: Given the basic form of a word in the target language, can we generate its inflected form based on the source text to be translated?*
- *RQ2: Does the inclusion of inflected words in constrained NMT improve translation quality?*

We plan to focus on the translation of texts from the IT domain from English into German.

Task description

The main tasks of this thesis are:

1. Read up on the current state of the art in industry and research in the areas relevant to the thesis, including NMT, lexically constrained decoding, and morphological inflection.
2. Get familiar with currently existing open-source frameworks for multi-source NMT.
3. Define and implement an evaluation framework.
4. Define, implement, and refine a model for morphological inflection of terminology in NMT.
5. Evaluate the interplay of the model with an existing constrained decoding framework.
6. Writing an academic report explaining and summarising the results from the work on items 1 to 6.

Deliverables

The major milestones of the project are as follows:

When	What
1^{st} and 2^{st} week	State of the Art review is finished. Open Source tools are examined in-depth, so the student understands them.
3^{st} and 4^{st} week	Basis tool for development is extended with the testing environment and ready for implementation of the model. The model is planned more in detail.
2^{st} month	First approach of proof-of-concept is implemented.
3^{nd} month	Evaluation and adjustment of first version model is done.
4^{rd} month	Second approach with experiments is implemented.
5^{th}month	The proof-of-concept implementation is finished.
6^{th}month	Thesis is written, proof-of-concept is fully functional, documented and delivered.

Provided resources

There is no need of special provided resources to realise this thesis.

General thesis guidelines

The typical rules of academic work must be followed. **?** describes a number of guidelines which must be followed. At the end of the thesis, a final report has to be written. The report should be clearly organised, follow the usual academic report structure, and has to be written in English. As implementing software is also part of this thesis, state-of-the-art design, coding, and documentation standards for the software have to be obeyed.

Advisors:

Professor:
Prof. Dr. Martin Volk

Responsible assistant:
Samuel Läubli

Signatures:

Andreas Schaufelbühl Prof. Dr. Martin Volk