UPPSALA UNIVERSITET
Department of Linguistics and Philology

Thesis – Statement of Intent (form with instructions)

Student

Shifei Chen (Shifei.Chen@student.uu.se)

Academic supervisor

Ali Basirat (Ali.Basirat@lingfil.uu.se)

Company/organization, contact person(s) (with e-mails) (if applicable)

None

The title, purpose, and extent of the thesis task

The supposed thesis title is "A Study of the Universal Word Embeddings Applicaiton in a Multilingual Neural Machine Translation System". The topic of the master thesis concerns about the generalizability of neural translation models across languages (from multiple source languages to multiple target languages) with regard to universal word embeddings.

Time management

The thesis will last until mid May (week 20). Here is the proposed detail timeline, with deadline dates in parentheses.

- Background research — week 7-8 (2.23)
- Data preparation — week 9 (3.1)
- Prelimentry experiments — week 10-11 (3.15)
- Reserved for buffering — week 12 (3.22)
- Main experiments — week 13-15 (4.12)
- Reserved for buffering — week 16 (4.19)
- Result analysis — week 17-18 (5.3)
- Overall documentation — week 19-20 (5.17)

Thesis content and structure

The thesis shall consist of 7 chapters, as listed below.

Chapter 1 will descirbe the master thesis briefly, giving out its purpose and the outline of the rest of the work.

Chapter 2 will focus on the background information of the thesis. It should include contents of

- Introduction to machine translation, multilingual translation and NMT (Neural Machine Translation)
- Introduction to word embeddings
- Introduction to universal word embeddings
- Previous work in multilingual NMT systems with word embeddings

Chapter 3-5 should detail the experiments with a multilingual NMT system and universal word embeddings from different propspectives — the data, the methodology and the evaluation metrics. In particular, Chapter 4, which is going to write about the methodology, should include things such as

- The training process of the universal word embeddings
- The design of the multilingual NMT system
- Overall experiment settings (hardware, software, etc.)

Chapter 6 will analysis the experiment results. Finally, everything will be concluded in Chapter 7.

Language

The thesis will be written in English.

Hardware and software

During the experiment we expect to use general Deep Learning software kits and frameworks, such as Python, Tensorflow/PyTorch, etc. The software part should not be the bottleneck.

For the hardware, at the time belong we have access the lab computers at the Department of Linguistics and Philology. If they are not enough we will search for additional resources with the help of my supervisor, including resources from other departments of the university, or other third-party partners.

Rights to results

As no third-party will be involved, the author owns the copyright of the thesis and all of the possible products during the thesis period. However, he will make all of these things available to the supervisor and the examiner during the thesis period, and to the public once the thesis is finished.

The dataset of the thesis, if it is not collected by the author, belongs to its own creater. It should be free for academic use.

Other things?

No