Replicating Vanmassenhove et al. 2018 with OpenNMT (provisional title)

Author1, Author2, Author3, Author4
Affiliation1, Affiliation2, Affiliation3
Address1, Address2, Address3
author1@xxx.yy, author2@zzz.edu, author3@hhh.com
{author1, author5, author9}@abc.org

Abstract

In this paper, we reproduce some of the experiments related to neural network training as reported in (Vanmassenhove and Way, 2018). They annotated a sample from the Europarl aligned corpora with Part-of-speech and semantic annotation to train neural networks with the Nematus Neural Machine Translation (NMT) toolkit. Following the original publication, we obtained significantly XXX results than the authors of the original paper. In the second half of the paper, we try to analyze the diffrenbce in the results obtained and suggest some methods to improve the results. Applying these methods, we report a precision of XXX .

Keywords: NMT, replication study, corpus annotation

1. Introduction

Laure: discuss the importance of reproducible results for science and write the outline of the paper and pay tribute to the outline I have "replicated, something like: We have been inspired by the previous editions of the workshops on replicability, especially (Branco et al., 2018) superscript[1]. Possibly mention the Blackbox papers

2. Characterisation of the Original Approach

Nabil: sum up the (Vanmassenhove and Way, 2018) paper.

2.1. Features/ Parameters

Laure: Underline everything that is not clear in the paper to reproduce the experiment.

2.2. Data sources and experiments

Nabil : detail the info from the original paper without paraphrasing them

3. Reimplementing the approach

Laure: Describe what you did, the version you used of OpenNMT and why (need to understand NN + industrial applications with SYSTAN tools).

3.1. Experimental setup

Nabil: be as specific as you can, see initial paper for inspiration.

3.2. Problems with reimplementation

Laure: Discuss every thing that is not explicit and the differences between nematus, default parameters to use, etc.

3.3. results

Make sure we reproduce the expected figures and results. Major reproduction comparables: BLEU scores (tables 1 and 2; plots in figures 2, 3 and 4). Comment the results.

[1] We found Repar et al. outline particularly convincing and this outline closely follows theirs.

4. Adaptation: Experiments for improving the reimplementation

Possible improvements of the original paper: prove the point that more adequately trained data provided better BLEU scores.Discuss alternative enrichment of the training data.

- **upos tags: POS tagging not dependent on a tagset**: Retagging the PoS-tags with upos universal part of speech: do we get better results. The tagset used for the PoS-tagging was not questioned in the original paper.
- **deep parsing**: Use a parser to provide the training with deep parsing information. PoS-tagging only consists in "shallow parsing", syntactic structures are annotated with . Use the SpaCy python library to annotate the data.

In this section we present several methods aiming at improving the results. :

4.1. Manual evaluation?

4.2. Other Scores

Discuss the possibility of extending supertags for French from the original paper, results not reported for French though the Europarl corpus is multilmingual and has a French component. https://github.com/nschneid/ pysupersensetagger

4.3. Syntactic supertags?

The original paper has addressed semantic tags and PoS tags but has eschewed syntactic tags. Discuss the parsing of the data with Spacy for universal dependency tags. Can we train a neural network with parsing information?

5. Conclusions and Future Work

In this paper, we tried to reimplement the approach to training neural networks with annotated data by (Vanmassenhove and Way, 2018). We first replicated it with a different Neural translation toolkit, to check some assumptions/facts

about NN architecture (Laure: discuss Nematus vs OpenNMT). We then tried to improve the quality of the translation by enriching the linguistic annotation of the input, taking into account several layers of annotation. Report previous works on Hebrew/English using morphological information. Discuss status of this information for industrial applications : generic NN engine or corpus-specific??

5.1. Bibliographical References

All bibliographical references within the text should be put in between parentheses with the author's surname followed by a comma before the date of publication,(StrÃ¶tgen and Gertz, 2012). If the sentence already includes the author's name, then it is only necessary to put the date in parentheses: StrÃ¶tgen and Gertz (2012). When several authors are cited, those references should be separated with a semicolon: (StrÃ¶tgen and Gertz, 2012; Castor and Pollux, 1992). When the reference has more than three authors, only cite the name of the first author followed by "et al." (e.g. (Superman et al., 2000)).
Example of a figure enclosed in a box:

Figure 1: The caption of the figure.

Level	Tools
MWU	Analyser
Syntax (Spacy Library)	

Table 1: The alternative levels for the annotation of the training set

6. Acknowledgements

Thanks are die to Benoit CrabbÃ© for advice on parsing and MArie Candito for annotation of phraseological units.

7. Bibliographical References

Branco, A., Calzolari, N., and Choukri, K. (2018). 4real 2018 workshop on replicability and reproducibility of research results in science and technology of language.

Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

StrÃ¶tgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey, may. European Language Resource Association (ELRA).

Superman, S., Batman, B., Catwoman, C., and Spiderman, S. (2000). *Superheroes experiences with books*. The Phantom Editors Associates, Gotham City, 20th edition.

Vanmassenhove, E. and Way, A. (2018). Supernmt: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73.

8. Language Resource References

lrec lrec2020W-xample Nematus tool kit Open NMT Tools to train the data ?