A Case Study: Custom Domain-Specific NMT for English-German Sports Ticker Translation

James Levell
Im Hürdli 2
8152 Opfikon
Switzerland
leveljam@students.zhaw.ch

Ondřej Bojar
Charles University, MFF, ÚFAL
Malostranské nám. 25, Praha 1
Czech Republic
bojar@ufal.mff.cuni.cz

Abstract

This paper summarizes our experiment with neural machine translation (NMT) for a particular domain: sports news tickers. We create our own, domain-specific, system using Marian NMT toolkit for the translation from English into German and compare its performance to currently available online translation systems by Google Translate and DeepL. As a measurement for the performance, three indicators were used: the automatically calculated BLEU score and two techniques of manual judgement of a subset of translations. According to our results, our domain-specific model outperforms Google Translate but DeepL turns out to be better.

1 Introduction

We were contacted by a sports news company which produces sports news tickers in multiple languages. Currently, the sports ticker news are translated manually. Great savings of both money and processing time could be achieved if this translation process was automated. We decided to run a pilot study for English-to-German translation. For the translation model, Marian NMT toolkit (https://marian-nmt.github.io/) was used. The training data was based on publicly available data sources (ACL 2019 , 2019) combined with the domain-specific data received from the customer.

The created model was then compared to two existing and well-known online translation systems: Google Translate (https://translate.google.com/) and DeepL (https://www.deepl.com/translator).

2 Creating a Domain-Specific Translation Model

As the corpus for the training data, publicly available sources were used:

- Europarl (1`838`568 sentence pairs): http://www.statmt.org/europarl/v9/training/europarl-v9.de-en.tsv.gz
- News-commentary (338`285 sentence pairs): http://data.statmt.org/news-commentary/v14/training/news-commentary-v14.de-en.tsv.gz
- Wikititles (1`305`141 sentence pairs): http://data.statmt.org/wikititles/v1/wikititles-v1.de-en.tsv.gz
- ParaCrawl (36`936`714 sentence pairs): https://s3.amazonaws.com/web-language-models/paracrawl/release5/en-de.classify.gz
- Common crawl corpus (2’399’123 sentence pairs): http://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz

2.1 Domain Adaption

To adapt the model for the sports domain, additional sports ticker news in English and German have been added to the training set. This domain-specific training data is very limited in size (37`575 sentence pairs only) but it exactly matches the domain of interest; it is the set of news previously translated by the company. To allow for automatic assessment of translation quality, a separate 3`000 sentence pairs from the

same proprietary source served as a dev set and another 3`000 sentence pairs as a test set.

In the end, the domain-specific sentences made only 0.87% of the whole training corpus.

2.2 Marian NMT toolkit Settings

To simplify the preprocessing, Marian NMT toolkit (marian-nmt, 2020) was used with its integrated sentence splitting option (marian-nmt, 2018).

2.3 Training

In our environment, Marian NMT toolkit training was able to process almost 16k words per second. We stopped the training after 114 hours (approximately 4.7 days) at which point the learning curves seemed to have flattened. In Figure 1, the overall performance during the training can be seen. The performance was measured using the BLEU and the cross-entropy (CE)-means values on the dev set.

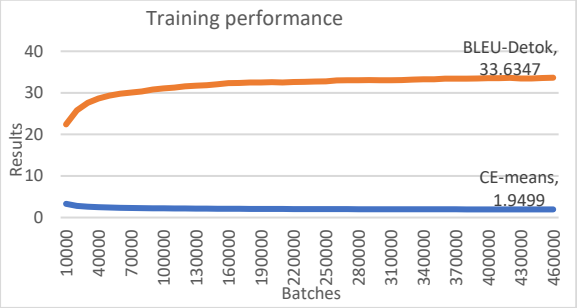

Figure 1 – Dev Set learning curves for the news domain.

3 Comparing translations

To compare the translation systems (Marian NMT toolkit, Google Translator, DeepL), three different methods were used:

- BLEU score: calculated BLEU score for the whole test set for each translation system
- White box judgment: manually judge each translation and select which one is the best for a given English source sentence and a German reference sentence. Only 50 translations have been presented to the judge.
- Black box judgement: manually judge 50 new translations but now without awareness which translation is coming from which origin. Each of the translations is then rated from best to worst (++, +, -).

3.1 BLEU Score

The BLEU scores have been calculated using the moses tool set (moses-smt, 2019).

Translation Tool	BLEU Score	Bleu Details	Brevity Penalty (BP)
Marian NMT toolkit	28.74	61.9 / 36.4 / 24.8 / 17.6	0.912
Google Translate	26.02	58.4 / 31.1 / 19.6 / 12.9	1.000
DeepL	37.49	64.9 / 42.3 / 30.9 / 23.3	1.000

Figure 2 - BLEU score of translation on the entire test set from each translator

Important to note in the above values is the BP penalty the marian NMT toolkit translations received. This indicates that the translations from this model are shorter, compared to human translations.

3.2 White Box Judgement

In our first small manual evaluation, the judge is aware from where the translation originates and can compare all the translations with one another. Then he has to select the best translation.

Translation method	Best translations white box
Marian NMT toolkit	0
Google Translate	0
DeepL	50

Figure 3 - Results of the white box judgement

In this evaluation, DeepL got the full score, while both other remaining systems were considered worse all the time.

3.3 Black box judgement

Using the tool QuickJudge from ÚFAL (UFAL, 2018), new translations are presented to the judge, but this time she is not aware of the origin. The translation presented are different from the previous ones to ensure that the judge does not know the translation system involved.

As seen bellow an example of the view for the judge. The first line presents the reference translation. The judge has to rate the following

translations, all coming from different translation systems, starting from best (++), ok (+) to worse (-). After the annotations have been made, the tool QuickJudge reveals the identity of the originating systems.

document/ original_de.txt	*Robin Haase gewann das bisher einzige Duell im Jahr 2014, als die beiden Spieler auf dem Challenger Tour gegeneinander spielten (Challenger Prostejov).*
-	*Haase gewann 2014 ihr einziges Meeting zurück, als sie sich im Challenger Circuit (Prostejov Challenger).*
+	*Haase gewann ihre einzige Sitzung wieder im Jahr 2014, wenn sie im Circuit der Herausforderer (Prostejov Challenger) erfüllt.*
++	*Ihre einzige Begegnung gewann Haase bereits 2014, als sie sich auf dem Challenger Circuit (Prostejov-Herausforderer) trafen.*

To make the results aggregable the ratings have been numerically converted (++ = 1, + = 0, - = -1)

Translation method	Rating using black box (++, +, -)
Marian NMT toolkit	-16
Google Translate	-20
DeepL	38

Figure 4 - Results of the black box judgement

4 Scores Summary

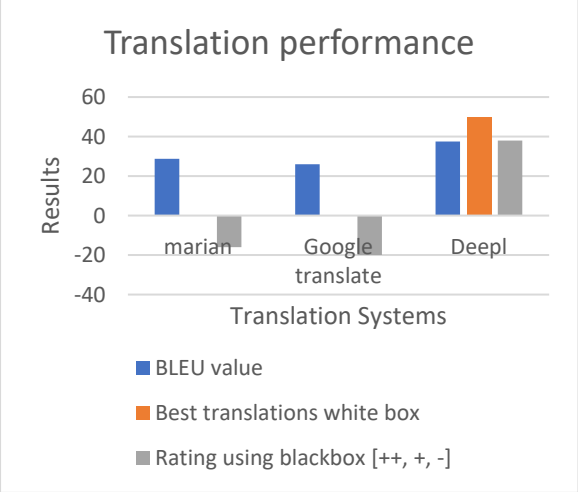

Figure 5 - summary of the translation comparison

It can be seen that DeepL outperforms other systems in all of the three different styles of evaluation. With our model which was created especially for this domain, the results are better than those of the overall market leader Google Translate.

4.1 Manual Assessment of the Translation Quality

In this section, we illustrate where each translation technology is failing.

In the following example, we can see that DeepL also provides translation of sayings, whereas Google and Marian NMT toolkit are too literal.

Original English	Heather Watson is on a roll at Hobart.
German reference	Heather Watson hat beim WTA-Turnier im australischen Hobart einen Lauf.
Translated using Marian NMT toolkit	Heather Watson ist auf einer Rolle bei Hobart.
Translated using Google Translate	Heather Watson ist auf einer Rolle in Hobart.
Translated using DeepL	Heather Watson hat eine Glückssträhne in Hobart.

Figure 6 – Example of a translation error: "too literal translation, missing sayings"

In the next example, DeepL could extract the domain and use that knowledge for the best translation regarding this specific domain. Again, Marian NMT toolkit and Google Translate simply translated the sentence on a word to word basis.

Original English	The German was extremely dominant on his serve last week at Doha and he continues at serve in the same level this week in Auckland.
German reference	Gojowczyk hat schon letzte Woche in Doha extrem gut aufgeschlagen und konnte seinen Service in Auckland bisher auf demselben Niveau halten.
Translated using Marian NMT toolkit	Der Deutsche war äußerst dominant an seiner letzten Woche in Doha und er weiterhin in der gleichen Zeit in Auckland.
Translated using	Der Deutsche war extrem dominant auf seinem letzte Woche in Doha dienen und er weiterhin in der

Google Translate	gleichen Ebene in dieser Woche in Auckland dienen.
Translated using DeepL	Der Deutsche war bei seinem Aufschlag in der vergangenen Woche in Doha extrem dominant und er setzt seinen Aufschlag in dieser Woche in Auckland auf gleichem Niveau fort.

Figure 7 - Example of a translation error: "missing domain-specific knowledge"

The following and last example illustrates the ability of DeepL to know when it does not have to translate a word. This is the case for "Denglish" words. So, words which are English but have been adopted to the German language.

Original English	Agut won 70% of his first serve points and saved four of the six break points chances he faced against Johnson in his second-round match.
German reference	Agut gewann dabei 70% seiner ersten Aufschlagpunkte und nutzte vier seiner sechs Breakchancen.
Translated using Marian NMT toolkit	Agut gewann 70% seiner ersten Punkte und rettete vier der sechs Bruchpunkte, die er gegen Johnson in seinem zweiten Spiel.
Translated using Google Translate	Agut hat 70% seiner ersten Punkte dienen und gespeichert vier der sechs Haltepunkte Chancen er gegen Johnson in seinem Zweitrundenspiel gegenüber.
Translated using DeepL	Agut gewann 70% seiner ersten Aufschlagspunkte und rettete vier der sechs Breakballchancen, die er gegen Johnson in seinem Zweitrundenspiel hatte.

Figure 8 - Example of a translation error: "translated Denglish words"

4.2 Summary of Results

It can be seen that DeepL is a powerful translation solution. The translated sentences are very fluent and feel natural compared to the bumpy translation provided by Google Translator and Marian NMT toolkit. As seen in the given translation example, DeepL seems well aware of sayings and English words, which should not be translated into German. This kind of awareness makes it a better and more natural translator.

The downsides of DeepL are the higher costs if you want it to use it on a larger scale (Deepl, 2020). The free version of DeepL allows a maximum of 5`000 characters to be translated per request. If you want to translate documents or integrate it into an app, additional costs for the usage of their API must be considered. Comparing it to the totally free version of Google Translate, which is even implemented in Google Sheets (Google, 2020), this can be quite a restriction.

Needless to say, we find it rather surprising, that most widely translator Google Translate is performing quite poorly in this test. For example, comparing the BLEU values of Google Translate 26.02 and DeepL 37.49.

5 Conclusion and Future Work

In this paper, we demonstrated that a domain-specific translation model can outperform the well-known translation system Google Translate. Furthermore, we documented that the performance of DeepL is remarkable even in the specific domain of sports news, a rather arbitrary test case for DeepL.

To further improve our model, firstly, the training data size for the domain adaption should be increased. Currently, only 0.87% of the training data is domain-specific. Enlarging the data promises to reduce the number of mis-translated words and also brings the chance of better coverage of saying. Lastly, it has to be verified if an increase in the training time of the model could still improve its performance.

References

ACL 2019 . (2019, August 1). FOURTH CONFERENCE ON MACHINE TRANSLATION (WMT19) . Retrieved from Shared Task: Machine Translation of News: http://www.statmt.org/wmt19/translation-task.html

Brownlee, J. (2017, 11 20). A Gentle Introduction to Calculating the BLEU Score for Text in Python. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

Deepl. (2020, March 25). DeepL. Retrieved from DeepL Pro: faster, safer, better: https://www.deepl.com/pro.html

Google. (2020, March 25). Support Google. Retrieved from GOOGLETRANSLATE: https://support.google.com/docs/answer/3093331?hl=de

marian-nmt. (2018, October 24). Github. Retrieved from SentencePiece: https://github.com/marian-nmt/sentencepiece

marian-nmt. (2020, January 21). MarianNMT. Retrieved from Documentation : https://marian-nmt.github.io

moses-smt. (2019, June 8). Github. Retrieved from Moses, the machine translation system: https://github.com/moses-smt/mosesdecoder

UFAL. (2018, February 7). Github. Retrieved from QuickJudge: https://github.com/ufal/quickjudge