Feature Identification in Corpus Linguistics:

A Case Study in Improving Machine Translation

Stephen Braich

Department of Computer Science and Engineering
Portland State University
sbraich@pdx.edu

Abstract

One of the challenges in adopting machine translation has been how to implement post-editor (translator) feedback. Machine translation engineers are given access to corpora to train translation models, but have struggled to make improvements suggested by the translators that use them. This study explores using NLP to identify types of grammatical features in corpora that are poorly translated. We setup a simple scenario creating conditions where a specific grammatical feature, the passive voice, is poorly translated, how we can identify this feature in corpora, and how augmenting our training corpus with passive voice phrases improves machine translation quality of this feature.

1. Introduction

The last two decades has seen rapid growth in the adoption of machine translation (MT). In this paper we address a problem faced by translation companies who have struggled to adopt MT in their translation workflow: How to implement feedback from translators (known as post-editors) Approaches to machine translation[1] (MT) can be grouped into three types: theory-based, corpus-based, and a hybrid of the two (Dipper, 2008). In this paper, the focus will be on the corpus-based approach and how to make improvements to MT using grammatical feature identification in our corpus.

1.1 Prior Research

Efforts to improving machine translation post-editing (MTPE) has been led by TAUS, an industry organization supported by large corporate translation buyers, such as Google, Microsoft, Dell, NIKE, and eBay. They have several programs to train post-editors (translators), and have published whitepapers on improving the MTPE workflow. While one can find many publications on how to report errors in MTPE (Beregovaya, 2014), there has been little in the way of guidance on how to correct such errors in translation models.

[1] Machine translation (MT) automates the translation of text from a source language (like English), into that of a target language (like French)

1.2 Overview of Machine Translation

The first machine translation engines were rules-based (RbMT) (Slocum, 1985). This theory-based method requires linguistic expertise in the source and target languages. Linguistic rules need to be encoded to translate source text into the target translation. Many organizations that choose to invest in machine translation often require it for multiple languages. It is difficult to find personnel with the technical expertise as well as the linguistic knowledge in all the languages that an organization requires. This requirement has been a contributing factor in the emergence and popularity of the corpus-based approach (Dipper, 2008).

Corpus-based methods include statistical machine (SMT) and neural machine translation (NMT). These methods do not require any linguistic knowledge of the source or target language (although it is helpful in evaluation). This 'lifting' of the language barrier has enabled organizations to hire software engineers to build MT solutions for a wide variety of languages for which there exists parallel corpora[2].

Statistical machine translation was the first corpus-based approach and was popularized by the Moses SMT open-source project developed by Philipp Kohen of the University of Edinburg (Koehn, 2010). However, it wasn't until the advent of neural machine translation, which uses a large artificial neural network, that MT began to approach the quality of human translation (Klein, 2017). These large neural networks required a very large parallel corpus. The task of gathering such a large parallel corpus required for building a neural network presents many challenges, including identification and alignment of parallel texts, and finding relevant parallel corpora for a specific domain.

1.2 Parallel Corpora as Training Assets for Machine Translation

Parallel corpora in the translation industry began to be stored as translation memories that could be looked up when similar documents required translation. Fuzzy matching logic gained traction as translators began working on similar content. Over time, an organization collected these 'memories' of translation and stored them in a central repository. In the early 1990s, the first statistical machine translation engines were designed by IBM using translation memories by which a machine could generate statistical models to predict translation. By the year 2000, many websites were multilingual reaching an international audience.

The world wide web started to become a became a parallel corpus unto itself. Organizations started to buy and trade their translation memory and fee-based and open-source platforms on the web began to emerge where parallel corpora could be downloaded. TAUS, the Translation Automation Users Society

[2] A parallel text is a text placed alongside its translation or translations. Parallel text alignment is the identification of the corresponding sentences in both halves of the parallel text. The Loeb Classical Library and the Clay Sanskrit Library are two examples of dual-language series of texts. The most famous example is the Rosetta Stone.
Source: https://en.wikipedia.org/wiki/Parallel_text

started in 2004 and became a portal where translation memories could be traded or purchased for a fee. The Opus project is a large repository on the web where parallel corpora can be downloaded for free. It is on this website where this project has found most of its data. As more and more parallel corpora becomes freely available, the problem of gathering a corpus for MT, becomes one of selecting the most relevant corpora to your machine translation needs.

Being able to respond to post-editor feedback by identifying corpora that can correct deficiencies in translation models and weight a corpus accordingly will be demonstrated in this paper.

2. Methodology

The corpus-driven approach to building a machine translation model requires that we build a bilingual corpus to train a machine on past translations. This is how the machine 'learns' to translate. The Opus project (open parallel corpora) provides a large repository of parallel corpora for training translation models. A list of the largest sources of this content is listed in Table 1.

Corpus	Domain / Register	Documents	sentences	English tokens	Croatian tokens
OpenSubtitles v2018	Movie & TV Subtitles	46239	37.5M	305.7M	243.4M
TildeMODEL v2018	Misc. EU Documentation	5	0.7M	133.9M	15.2M
ParaCrawl v5	EU Funded Web Crawl Project	38	1.9M	49.5M	46.3M
JW300 v1	Jehovah's Witness Parallel Corpus	14473	1.1M	19.5M	17.7M
DGT v2019	EU Directorate General of Translation	4193	0.7M	17.3M	14.3M
Tatoeba v20190709	Free translation repo	1	2.4k	11.0M	33.7k
SETIMES v2	SE European Times	1	0.2M	4.9M	4.6M
wikimedia v20190628	Wikipedia Translations	1	2.5k	7.7M	0.5M
QED v2.0a	Qatar Educational Domain	1736	0.2M	3.7M	3.0M
hrenWaC v1	English - Croatian Parallel Web Corpus	1	99.0k	2.6M	2.3M
bible-uedin v1	Multilingual Bible Corpus	2	62.2k	1.8M	1.4M
TedTalks v1	Ted Talks crowd sourced Translation	1	86.3k	1.5M	1.3M
Ubuntu v14.10	Linux Software Localization	293	52.7k	0.5M	0.2M
EUbookshop v2	EU Bookshop corpora	23	6.2k	0.2M	0.2M
	Total	**68702**	**43.1M**	**562.7M**	**351.7M**

Table 1: Opus Project: The open parallel corpus
From < http://opus.nlpl.eu/ >

For this study, we built a Croatian to English neural machine translation model using a corpus built from translations of TED Talks and parallel corpora mined from the web by EU funded ParaCrawl project. These two corpuses were selected due to prior experience with them and are highlighted in yellow. The Ted Talks corpus is of high translation quality. The EU funded ParaCrawl project has an extremely large repository of parallel corpora, being that the data was mined from the entire web. However, the quality varies due to the lack of control of over the translation process.

Next, we selected one grammatical feature, the passive voice, and use it to explore how to develop an approach to implementing machine translation quality feedback. In our hypothetical scenario, we imagine that a post-editor had documented that our machine translation model poorly translates phrases in the passive voice. In Table 2, we show several examples of phrases in the passive voice and their active voice equivalents.

Active Voice	**Passive Voice**
Harry ate six shrimp at dinner.	At dinner, six shrimp were eaten by Harry.
Sue changed the flat tire.	The flat tire was changed by Sue.
We are going to watch a movie tonight.	A movie is going to be watched by us tonight.
I ran the obstacle course in record time.	The obstacle course was run by me in record time.
Mom read the novel in one day.	The novel was read by Mom in one day.
I will clean the house every Saturday.	The house will be cleaned by me every Saturday.
Tom painted the entire house.	The entire house was painted by Tom.
A forest fire destroyed the whole suburb.	The whole suburb was destroyed by a forest fire.
The two kings are signing the treaty.	The treaty is being signed by the two kings.

Table 2: Active vs. Passive Voice Phrases

From <https://examples.yourdictionary.com/examples-of-active-and-passive-voice.html>

2.1 Test Design

In order to create the scenario where a machine translation model poorly translated phrases in the passive voice, we used python’s NLTK Library to develop a script to identify passive voice phrases in our corpus and removed them. We then trained a neural translation model using this corpus and tested how well it translated passive voice phrases using the BLEU[3] metric. Below in Table 3, details are provided on the number of sentences and tokens (words) for our baseline and hypothesis models after preprocessing, deduplicating, and cleaning up the corpora.

Training Corpus	Sentences	Croatian Tokens	English Tokens
Only Active Voice Phrases	1,398,802	31,04,0676	33,558,172
Both Active and Passive Voice Phrases	1,775,644	42,761,266	46,477,826

Table 3: Size and composition of baseline and hypothesis corpora

For our test corpus, we selected only passive phrases (identified by our passive voice identification script) that our models would translate. We created two separate test corpus files. One from the EU funded ParaCrawl corpus and the other from the Ted Talks corpus. The evaluation is done by machine translating the Croatian source sentences and comparing that output with human translation output that was provided in the parallel corpus. Specifically, we use the BLEU metric developed by IBM to evaluate how well (or how close) the machine translated output matches the human translated output.

2.2 Grammatical Feature Identification

Using Python's Natural Language Toolkit (NLTK), we created a script to identify phrases in our corpus that are in the passive voice. This script was originally developed by a Chinese developer who made his code open source and available on GitHub[4]. The goal is to extend NLTK with a library of grammatical feature detection.

Currently with Python's NLTK Library, we can break a corpus into sentences, then tokens, and finally part-of-speech tags (Bird, Klein, & Loper, 2009). These are built-in functions. But the functionality stops there. There is no built-in detection of grammatical features. These grammatical features aren't dependent on a specific corpus or specific use cases, so there is no reason why they couldn't be included. Figure 1 illustrates the workflow of abstracting out grammatical features out of a corpus. The first four stages are built in to the NLTK. This study implemented the fifth stage for

[3] BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics. https://en.wikipedia.org/wiki/BLEU

[4] flycrane01 / nltk-passive-voice-detector-for-English on GitHub
https://github.com/flycrane01/nltk-passive-voice-detector-for-English

"passive voice" detection[5] and proposes that additional grammatical feature detection be added and the functionality be included in the NLTK Library.

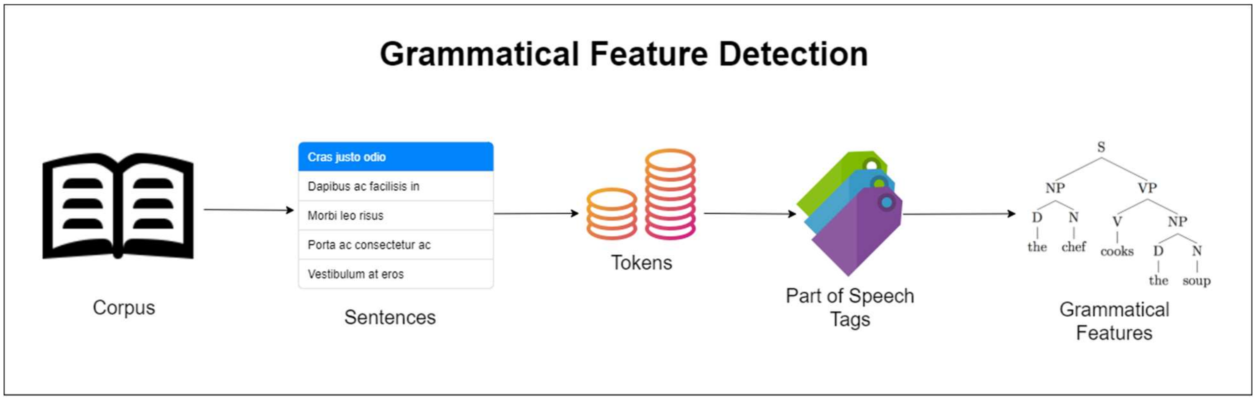

Figure 1: Grammatical Feature Detection Workflow

3. Results

The baseline translation model trained on a corpus without any passive voice phrases achieved a BLEU score of approximately 32 on the ParaCrawl test corpus and a score of 24 on the Ted Talks corpus. To test our hypothesis that we can improve a translation model given post-editor feedback, we now create a second neural translation model with the identified passive voice phrases added back in.

Neural Translation Models Trained with Corpus	ParaCrawl "Passive Voice" BLEU Tokenized	TedTalks "Passive Voice" BLEU Tokenized
Only Active Voice Phrases	31.71	24.09
Both Active and Passive Voice Phrases	36.81	25.12

Table 4: Comparison of translation quality of passive voice phrases

We see an improvement of over 5 BLEU points tested on 2,000 sentences from the ParaCrawl corpus. We saw only an improvement of 1 BLEU point on the Ted Talks test corpus, but its sample size was too small (400 samples) to show a significant improvement. We added it because it the TED Talks are of

[5] Steve3p0 / LING576 GitHub Repository for LING 576: Corpus Linguistics https://github.com/steve3p0/LING576

high translation quality and suitable for human review. Let's see if the BLEU metric is supported by some human evaluation of the results. Table 3 below displays the translation output.

Croatian Source Sentence	English - HUMAN Translation	Active Model – MT Output	Active & Passive Model – MT Output
Trgovina je porasla.	Trade has increased.	Trade rose.	The trade ***has*** increased.
Ovaj dečko se zove Sadik.	This boy is called Sadik.	This guy called Sadik.	This guy ***is*** called Sadik.
Ovo su slike koje se automatski stvaraju.	These are the images that are automatically constructed.	These are images that automatically create.	These are images that ***are*** automatically creat**ed**.
Potom je otkriveno da se svemir širi .	Then it was discovered that the universe was expanding.	He then discovered that the universe is expanding.	Then ***it was discovered*** that the universe is expanding.
Ovaj model možemo primjeniti na svaki skup bolesti .	This model can be adapted to every disease process .	This model we can apply to every set of disease .	This model ***can be applied*** to each set of diseases.
Rečeno mi je da ćeš me ostaviti ovdje .	I am told you will leave me here .	He told me that you will leave me here .	I ***was told*** you will leave me here.
A 2.000 novih domova se gradi odmah pokraj te elektrane .	And there are 2,000 new homes being built next to this power station .	And 2,000 new homes are building right next to these power plants .	And 2,000 new homes ***are being built*** right next to these power plants.
PDV je uračunat u cijenu.	VAT is calculated into the price.	VAT included in price.	VAT ***is included*** in the price.
Jedino naselje nalazi se na sjevernoj obali otoka.	The only village is located on the northern coast of the island.	The only village is on the northern coast of the island.	The only settlement **is located** on the northern coast of the island.
Stranice ovog korisnika su posjećene 9857 puta od 4/5/2011.	These user pages have been visited 9872 times since 4/5/2011.	The pages of this user are **posjećene** puta times from 4/5/2011.	These user pages ***have been visited*** 9857 times since 4/5/2011.
Sliku Opis teksta je preveden sa wikipedia i licencirano pod GFDL.	The text description is extracted from wikipedia and licensed under the GFDL.	Sliku Description of the text is translated from wikipedia and **licencirano** under GFDL.	The text description ***is extracted*** from wikipedia and licensed under the GFDL.
Trajanje ture je oko 3 sata i moguće ju je prilagoditi vašim željama.	The tour takes about 3 hours and it can be adapted to your wishes.	The tour duration is about 3 hours and it is possible to adjust it to your wishes.	The duration of the tour is about 3 hours and ***can be adjusted*** to your wishes.

Table 5: Samples of human and translation model output

The first column is the Croatian test sentence to be translated. The second column contains human translations into English that we compare our machine translation output to. The third column is our baseline translations. This output is from the translation model that was not trained on passive voice phrases. This output is what our post-editor (translator) might complain about and ask us to improve. And lastly, the fourth column is from the translation model that was augmented with passive voice phrases identified by our script. It represents the improvements requested by the post-editor. It is obvious that the model trained on passive voice phrases can translate sentences into the passive voice. The model trained only on active voice phrases is unable to do so. Some of active voice translation are simply grammatical incorrect. We can see a few that even fail to translate some of the verbs directly involved in the passive voice (they are left untranslated – in bold) A few are surprisingly grammatically correct, but again, they are not in the passive voice.

4. Conclusions

It has been demonstrated that grammatical features can be automatically detected in corpora. The example of the passive voice grammatical feature was detected in our corpus. As a proof of concept, these phrases were removed from a corpus used to train a baseline translation model. This model struggled to translate phrases that were in the passive voice. We then added those phrases back into the training data and trained a model[6] that could translate the passive voice.

On the surface, it might seem obvious that a machine can not learn from data that is purposely excluded, as we did with our first baseline model. However, the author has been presented with scenarios professionally where post-editor feedback detailing poor translation quality of a particular grammar feature was documented and was unable to find a solution. Industry experts were consulted but were unable to help. Instead of artificially creating a deficiency in our study, that is, excluding the passive voice corpora so it doesn't translate the passive voice well, a better study might be to identify a particular grammatical feature that an existing MT model does not do well, and then augment its corpus with phrases containing that feature.

We could extend this study to developing scripts to identify more grammatical features. Armed with this ability, machine translation engineers will be in a better position to respond to feedback by post-editors (translators) to make improvements to translation quality. Additionally, if we had a large enough library of grammatical feature detection, we could run metrics to see how well our training corpus provides coverage of a language, and thus reduce the need for post-editor feedback.

[6] A demo of the Croatian to English and English to Croatian machine translation models is available at http://mt.autom8tr8n.com.

References

Beregovaya, O. (2014, January). *Evaluating Post-Editor Performance Guidelines*. Retrieved from TAUS: The Language Data Network: https://www.taus.net/academy/best-practices/postedit-best-practices/evaluating-post-editor-performance-guidelines

Dipper, S. (2008). Theory-driven and Corpus-driven Computational Linguistics and the Use of Corpora. *Corpus Linguistics: An International Handbook, volume 1* (pp. 68-96). Berlin: Mouton de Gruyter.

Klein, G. K. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations* (pp. 67-72). Vancouver, Canada: Association for Compututational Linguistics.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge : Cambridge University Press.

Natural Language Toolkit. (2020, January). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Natural_Language_Toolkit

NLTK 3.4.5 documentation. (2019, August 20). Retrieved from Natural Language Toolkit: http://www.nltk.org/

Slocum, J. (1985). A Survey of Machine Translation: it's History, Current Status, and Future Prospects. *Computational Linguistics, volume 11*, 1-17.

Somers, H. (2008). Corpora and Machine Translation. *Corpus Linguistics: An International Hanbook, volume 2* (pp. 1175-1196). Berlin: Mouton de Gruyter.

Steven Bird, E. K. (2009). *Natural Language Processing with Python.* O'Reilly Media. Retrieved from Natural Language Processing with Python - Analying Text with the Natural Language Toolkit: http://www.nltk.org/book_1ed/