Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3079–3083
Marseille, 20-25 June 2022

Using Semantic Role Labeling to Improve Neural Machine Translation

Reinhard Rapp
Athena R.C.
Magdeburg-Stendal University of Applied Sciences
University of Mainz
reinhardrapp@gmx.de

Abstract

Despite impressive progress in machine translation in recent years, it has occasionally been argued that current systems are still mainly based on pattern recognition and that further progress may be possible by using text understanding techniques, thereby e.g. looking at semantics of the type "Who is doing what to whom?". In the current research we aim to take a small step into this direction. Assuming that semantic role labeling (SRL) grasps some of the relevant semantics, we automatically annotate the source language side of a standard parallel corpus, namely Europarl, with semantic roles. We then train a neural machine translation (NMT) system using the annotated corpus on the source language side, and the original unannotated corpus on the target language side. New text to be translated is first annotated by the same SRL system and then fed into the translation system. We compare the results to those of a baseline NMT system trained with unannotated text on both sides and find that the SRL-based system yields small improvements in terms of BLEU scores for each of the four language pairs under investigation, involving English, French, German, Greek and Spanish.

Keywords: neural machine translation, semantic role labeling, semantics-based translation

1. Introduction

After decades of research on rule-based machine translation (MT), data-driven approaches led to substantial advances in translation quality. Apparently, language translation is too sophisticated, ambiguous and irregular to be described by a reasonably sized and manageable set of rules, put together by linguists. However, it can be argued that data-driven approaches, including example-based, statistical and neural, are essentially based on pattern recognition, and to further advance the technology some form of language understanding may be desirable. Although it is not easy to define what understanding means, at the core of it might be questions such as "Who is doing what to whom, how, why, when and where?".

Semantic role labeling is often considered as providing some (though limited) access to the problem of understanding. In SRL it is tried to identify the semantic arguments of a predicate and to label them with their semantic roles. To train automatic systems on this task, annotated corpora are necessary which are large enough that samples of as many predicate/role instances as possible are covered. In projects such as FrameNet (Fillmore, 1982) and Propbank (Palmer et al., 2005) researchers have taken up this challenge. More so than FrameNet, Propbank has its focus on practical corpus processing and was therefore chosen as our semantic framework here.

To give an example of the semantic role annotations provided in such frameworks, let us look at the sentence "John hits the ball" which might be annotated as "John-[agent] hits-[predicate] the ball-[patiens]". Note that such semantic annotations are meant to be largely language independent and can therefore be useful to bridge the gap between languages. In the context of MT, they may even be considered as a form of a coarse-grained interlingua.

The approach pursued in the current work is to annotate the source language part of a parallel corpus with semantic roles, and to train an NMT system with a training set consisting of this annotated corpus on the source language side and the original unannotated corpus on the target language side. The hope is that this will improve NMT training and consequently translation quality.

Of course, it can be argued that NMT implicitly takes semantics into account, and that this may well include the information ascribed to SRL. However, a system specifically dedicated to SRL may help in doing so better, although, on the other hand, a major advantage of NMT is that it considers all levels of language processing, including morphology, syntax and semantics, at the same time, thereby potentially taking into account sophisticated interdependencies between levels. This effect may be compromised if we conduct SRL in a separate pre-processing step. As it is hard to predict whether in practice the pros of explicit SRL will outweigh the cons, we decided to conduct the current study. The work is innovative in so far as we are not aware of many other studies combining NMT and SRL in a fully operational MT system for several language pairs. Problems when conducting this kind of work are that SRL is sophisticated, that a high-quality SRL system is required, and that the original annotations produced by the systems need to be modified for NMT use. Also, as NMT requires substantial training data, relatively large corpora need to be annotated with semantic roles, which is computationally demanding.

2. Related work

There is not a lot of previous work combining NMT and SRL. Such work includes Nguyen et al. (2021) who use semantic graphs from abstract meaning representation (AMR) to improve NMT. The graphs are generated using the NeuralAmr toolkit[1] (Konstas et al., 2017) which implements sequence-to-sequence models for AMR parsing and generation. In this work, the training is based on about 130,000 English-Vietnamese sentence pairs, using byte pair encoding with only 8000 operations. Both numbers mean that it is a relatively small system which is dedicated to the domain of TED talks, where sentences can be expected to be relatively short on average, making the translation problem easier.

[1] https://github.com/sinantie/NeuralAmr

Another related paper is Marcheggiani (2018) who combine SRL with NMT for the language pair en-de. Using a training corpus of 4.5 million sentence pairs they achieve a BLEU score of 23.3 without SRL and of 24.5 with SRL. This compares to BLEU scores of 30.2 and 31.6 for the same language pair in our system, whose training is based on a corpus of about 2 million sentences.
Wu et al. (2021) annotate sentences with predicate-argument structures at the word level and use these to train an NMT system. They are able to improve the performance of NMT in a low-resource setting involving Chinese, Mongolian, Uyghur, and Tibetan by an average of 1.18 BLEU points.

3. Semantic role labeling

Implementing a system for semantic role labeling is a major task (Gildea & Jurafsky, 2000). Therefore, we investigated whether we could utilize existing systems such as the AllenNLP SRL system[2] or the Illinois SRL system[3]. As only the AllenNLP SRL system uses the latest neural technology and claims to achieve state-of-the-art results (86.49 test F1 score on the Ontonotes 5.0 dataset), we decided to use this system. It implements a variant of the BERT-based model described by Shi & Lin (2019).
AllenNLP SRL uses PropBank (Palmer et al., 2005) annotations, which is the most widely used standard in SRL. To give an example, using these annotations the sentence "He would not accept green apples from the fruit market." would be annotated as follows:

[He A0] [would AM-MOD] [not AM-NEG] [accept V] [green apples A1] [from the fruit market A2] .

The set of roles for the predicate *accept* is defined in the PropBank frames scheme as follows:

V: verb
A0: acceptor
A1: thing accepted
A2: accepted-from
A3: attribute
AM-MOD: modal
AM-NEG: negation

To obtain such annotations, we installed the AllenNLP SRL system on two computers, both running Ubuntu 20.04 (LTS). One was a laptop with i5 CPU, another a high end PC with an i9 CPU. We wrote a Python script for applying the system sentence by sentence to the English parts of the de-en, el-en, es-en, and fr-en[4] Europarl v7 corpora. The runtime per corpus was about four days on the laptop and about three days on the PC for the English parts of de-en, es-en, and fr-en corpora whose size was in the order of 2 million sentences or 50 million words each. In the case of el-en runtime was roughly a day less on both systems because this corpus is smaller.
Although the English parts of the four parallel corpora have a lot of overlap, it was easier for us to conduct the annotations four times separately rather than to identify which sentences were in common.

The AllenNLP SRL system reproducibly crashed with a runtime error when applied to sentences longer than about 1500 characters (which tended to be lists of items), so we cut off the few sentences where this was the case. This caused no problems as the data was meant to be used for NMT training, and for this purpose very long sentences are usually not taken into account anyway.
Like other readily available SRL systems, AllenNLP SRL can only deal with English. The reason is that its internal neural system needs to be trained with a large annotated corpus, which in the form of Ontonotes 5.0 (Weischedel et al, 2011) is available for English only. In principle, for parallel corpora it would be possible to project the SRL annotations from English to other languages via word alignments (e.g. using Giza++ or fastalign). However, the quality of the resulting annotations suffers from a number of problems, including the following: 1) Sentence translations can be free. 2) Word alignment is difficult and error-prone. 3) The English verbs and their target language translations can well have different frames and require different annotations.[5]
For such reasons, annotation quality via cross-lingual projection is lower than in the case of direct SRL. We therefore do not further pursue this in the current study and are thus limited to language pairs with English as the source language.
AllenNLP SRL produces output which uses a lot of bracketing, tends to be rather lengthy and contains quite a bit of redundancy. As it therefore appears not to be well suited for direct use in NMT training, we wrote a converter program to parse the AllenNLP SRL output, to extract the essential labeling information from it, and to annotate each word with the appropriate labels.
Let us exemplify the procedure by looking at a sample sentence from the Europarl de-en corpus and by providing the respective AllenNLP annotations as well as the transformed annotations as generated by our converter.
Given the English sentence "Today's decision not to renew the embargo is extremely dangerous considering the situation there." the AllenNLP SRL system comes up with the following description:

{'verbs': [{**'verb': 'renew'**, 'description': "Today 's decision [ARGM-NEG: not] to [V: renew] [ARG1: the embargo] is extremely dangerous considering the situation there .", 'tags': ['O', 'O', 'O', 'B-ARGM-NEG', 'O', 'B-V', 'B-ARG1', 'I-ARG1', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']}, {**'verb': 'is'**, 'description': "[ARG1: Today 's decision not to renew the embargo] [V: is] [ARG2: extremely dangerous] [ARGM-ADV: considering the situation there] .", 'tags': ['B-ARG1', 'I-ARG1', 'I-ARG1', 'I-ARG1', 'I-ARG1', 'I-ARG1', 'I-ARG1', 'I-ARG1', 'B-V', 'B-ARG2', 'I-ARG2', 'B-ARGM-ADV', 'I-ARGM-ADV', 'I-ARGM-ADV', 'I-ARGM-ADV', 'O']}, {**'verb': 'considering'**, 'description': "Today 's decision not to renew the embargo is extremely dangerous [V: considering] [ARG1: the situation there] .", 'tags': ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-V', 'B-ARG1', 'I-ARG1', 'I-ARG1', 'O']}], **'words'**: ['Today', "'s", 'decision', 'not', 'to', 'renew', 'the', 'embargo', 'is', 'extremely', 'dangerous', 'considering', 'the', 'situation', 'there', '.']}

[2] https://demo.allennlp.org/semantic-role-labeling
[3] https://cogcomp.seas.upenn.edu/page/software_view/SRL
[4] el = Greek, en = English, es = Spanish, fr = French, de = German.
[5] Fei et al. (2020) cite recent literature discussing such issues.

For an explanation of the arguments and further information, see the release notes of the OntoNotes project.[6] To emphasize the structure of the SRL output, we have highlighted the strings that start new frames. These are the three verbs (*renew*, *is*, and *considering*) occurring in the sentence. Their frames show considerable overlap, leading to redundancy. Towards the end of the description comes the "words" tag which shows how the sentence was tokenized.

As can be seen, the tokenizer integrated in the AllenNLP SRL tool has split "Today's" into two words, which may be controversial. Although the above sentence is not very long, for the reason that it contains three verbs its description is nevertheless somewhat sophisticated. Our specially developed parsing and conversion process, which tries to simplify the annotation (for use in NMT training) whilst keeping much of the essential information, converts the above to the following:

```
Today          B-ARG1
's             I-ARG1
decision       I-ARG1
not            B-ARGM-NEG I-ARG1
to             I-ARG1
renew          B-V I-ARG1
the            B-ARG1 I-ARG1
embargo        I-ARG1 I-ARG1
is             B-V
extremely      B-ARG2
dangerous      I-ARG2
considering    B-ARGM-ADV B-V
the            I-ARGM-ADV B-ARG1
situation      I-ARGM-ADV I-ARG1
there          I-ARGM-ADV I-ARG1
.
```

This conversion is based on the 'words' and 'tags' sections of the AllenNLP SRL output, but eliminating the many 'O' labels which are only placeholders for the respective positions and are not required in our format. Some words in the sentence have two labels as they relate to two of the three verbs in the sentence. In this particular example, none of the words relates to all three verbs, although in general this would be possible.

4. Experiments

As in previous work (Rapp, 2021), for running our experiments we used the Marian NMT toolkit (Junczys-Dowmunt et al., 2018). It was installed on a high end PC with an i9 CPU and an Nvidia RTX 3090 GPU with 24 GB of dedicated memory, running under the Ubuntu 20.04 LTS operating system.

For training the NMT system we used the en-de, en-el, en-es and en-fr portions of the Eurparl v7 corpus (Koehn, 2005). For each language pair, 2000 randomly selected sentence pairs were held out as our development set and another 2000 as our test set.

To get our baseline results (without SRL), we tokenized and true-cased the corpus using Moses tools (Koehn et al., 2007) and then applied byte-pair-encoding (Sennrich et al., 2016). For post-processing of the translations, the tokenization and true-casing was reversed.

To obtain the SRL-based results, we processed the English language parts of the above Europarl portions using the AllenNLP SRL tool. As this tool does its own pre-processing, we did not tokenize and true-case here. However, we converted the SRL output into the tabular format as described in the previous section and also performed byte-pair encoding on both the source and the target language side.

The training was conducted using 100,000 merge operations for byte pair encoding and fairly standard Marian NMT parameters. As our system architecture we used Google's transformer (Vaswani et al., 2017). Further details on this, including scripts, are provided in the paper describing our contribution to the WMT 2021 shared task on similar language translation (Rapp, 2021) and in our tutorial on NMT (Rapp, 2022) and its accompanying website.[7]

During training, BLEU scores are computed periodically on the development set and training is stopped if the best score cannot be improved within ten iterations. For computing the BLEU scores, the multi-bleu-detok.perl script from the Moses toolkit is used.

5. Results

Table 1 shows the resulting BLEU scores for systems trained on the Europarl corpus for the five language pairs with and without semantic role labeling on the source language side of the test sets.

Language pair	BLEU score (w/o SRL)	BLEU score (with SRL)
en – de	30.17	31.57
en – el	36.47	36.96
en – es	43.32	43.88
en – fr	38.84	39.02

Table 1: BLEU scores for test sets.

Language pair	BLEU score (w/o SRL)	BLEU score (with SRL)
en – de	29.86	31.40
en – el	36.98	37.63
en – es	42.83	42.86
en – fr	39.38	39.66

Table 2: BLEU scores for development sets.

To give a better idea of the variation of the BLEU scores, Table 2 shows the same results for the development sets. Note that the development sets had only be used as a criterion for ending training sessions.

As can be seen, for all language pairs a small improvement could be achieved when using SRL. As this improvement is consistent over all eight runs, it is unlikely that it is just a random effect. Therefore, it may be worthwhile to further investigate the use of SRL in NMT. This could be done by varying the NMT parameters and by modifying the experimental setting.

To give a qualitative impression of the translation quality, Table 3 shows the translations of the first five sentences from the test set with and without SRL. Note that these are

[6] https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf

[7] https://sites.google.com/view/mttutorial

random sentences extracted from Europarl that do not form a consecutive text.

To confirm that the conversion step from the original SRL annotation to the simplified version (as described in Section 3) is necessary, we also did a run where we trained the NMT system on the original AllenNLP SRL annotations. As these tend to be too long for Marian NMT training, we had to restrict sentence length to 100 words and received a BLEU score of only 16.3 for the language pair en-de, indicating that the NMT system cannot deal well with the original annotations.

6. Summary, discussion and prospects

In this work, we compared NMT results for two scenarios: One being standard NMT, the other using pre-processing on the source language side via SRL annotations. For all four language pairs tested, we achieved a small but consistent improvement in BLEU scores when using SRL. This finding provides some evidence that the sophisticated semantic analysis provided by a dedicated SRL system might be somewhat superior compared to what NMT is doing implicitly. The improvement is in agreement with the findings of the related work mentioned in Section 2.

In our system we used standard Marian NMT parameters. Parameter optimization is very time consuming due to the necessity of running expensive training sessions for every new parameter setting.

However, in future work optimizing parameters would be desirable e.g. in the following ways:

- Trying out the big transformer architecture.
- Comparing the effects of varying the numbers of merge operations in byte pair encoding.
- Comparing other ways of how to convert the AllenNLP SRL output to input appropriate for the NMT system. In particular, not only the semantic roles but also the respective verbs could be included.

Furthermore, other training corpora, further language pairs and additional evaluation metrics could be considered.

Acknowledgements

This work was funded by the Marie Curie Individual Fellowship SEBAMAT (grant agreement number 844951) within the European Commission's Horizon 2020 Framework Programme. Many thanks to Athena R.C. and to the creators of the tools and resources used in this work.

References

Hao Fei, Meishan Zhang, Donghong Ji (2020). Cross-lingual semantic role labeling with high-quality translated training corpus. *Proceedings of the 58th Annual Meeting of the ACL*, 7014–7026.

Charles Fillmore (1982). Frame semantics. In: *Linguistics in the Morning Calm.* Seoul, Hanshin Publishing Co., 111–137.

Daniel Gildea; Daniel Jurafsky (2000). Automatic Labeling of Semantic Roles. *Computational Linguistics* 28 (3).

Junczys-Dowmunt, Marcin; Grundkiewicz, Roman; Dwojak, Tomasz; Hoang, Hieu; Heafield, Kenneth; Neckermann, Tom; Seide, Frank; Germann, Ulrich; Aji, Alham; Bogoychev, Nikolay; Martins, André; Birch, Alexandra (2018). Marian: Fast Neural Machine Translation in C++. *Proc. of ACL, System Demonstrations*, 116–121.

English Original	**German translation without SRL**	**German translation with SRL**
Mr President, we see the reality of the situation all around us. If 23% of Europe' s population is over 65 years old today, this figure will have reached 40% by 2025, and the proportion of four working people to every retired person will have become two working people to every retired person by 2020	Herr Präsident, wir sehen die Realität um uns herum: Wenn heute 23% der Bevölkerung in Europa älter als 65 Jahre sind, wird diese Zahl bis 2025 40% erreicht haben, und der Anteil von vier Arbeitnehmern an jedem Rentner wird bis 2020 zu zwei Erwerbstätigen für jeden Rentner geworden sein.	Herr Präsident, wir sehen die Realität überall: Wenn heute 23% der Bevölkerung Europas älter als 65 Jahre sind, dann wird diese Zahl bis 2025 40% erreicht haben, und der Anteil von vier Arbeitnehmern an jedem Rentner wird bis 2020 zu zwei Arbeitnehmern für jeden Rentner geworden sein.
Mr President, perhaps I might say, before I say anything else, that the behaviour indulged in by some of the dockers yesterday here in Strasbourg goes far beyond what is tolerable or permissible in political life..	(EN) Herr Präsident! Bevor ich etwas anderes sage, möchte ich sagen, dass das Verhalten, das gestern hier in Straßburg von einigen Abgeordneten hier im Parlament begangen wurde, weit über das hinausgeht, was im politischen Leben hinnehmbar oder zulässig ist.	Herr Präsident, gestatten Sie mir zunächst die Bemerkung, daß das Verhalten, das einige der Hafenarbeiter gestern hier in Straßburg an den Tag gelegt haben, weit über das hinausgeht, was im politischen Leben akzeptabel und zulässig ist.
Mr President, I would like to point out that competition is an important area of policy, but that it is not part of the Built-In Agenda..	Herr Präsident! Ich möchte darauf hinweisen, dass Wettbewerb ein wichtiger Politikbereich ist, dass er aber nicht Teil der Agenda ist.	(EN) Herr Präsident! Ich möchte darauf hinweisen, dass der Wettbewerb ein wichtiger Politikbereich ist, dass er jedoch nicht Teil der "In-in" Agenda ist.
If it supposedly poses no danger, then it is more than suspicious that liability is so vehemently rejected.	Wenn es angeblich keine Gefahr darstellt, dann ist es mehr als zweifelhaft, dass die Haftung so vehement abgelehnt wird.	Wenn sie angeblich keine Gefahr darstellt, dann ist es mehr als verdächtig, dass die Haftung so vehement abgelehnt wird.
You took the initiative, helped find a solution to the protracted internal conflict in Ireland and helped normalise the situation there.	Sie haben die Initiative ergriffen, dazu beigetragen, eine Lösung für den langwierigen internen Konflikt in Irland zu finden, und dazu beigetragen, die Situation dort zu verbessern.	Sie haben die Initiative ergriffen, dazu beigetragen, eine Lösung für den langwierigen internen Konflikt in Irland zu finden, und dazu beigetragen, die Lage dort zu normalisieren.

Table 3: Sample translations with and without SRL.

Koehn, Philipp (2005). Europarl: a parallel corpus for statistical machine translation. Proceedings of the Tenth Machine Translation Summit, Phuket, Thai-land, 79–86.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Mo-ran, Richard Zens, Chris Dyer, Ondrej Bojar, Alex-andra Constantin, Evan Herbst. (2007). Moses: Open Source Toolkit for Statistical Machine Trans-lation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, 177–180.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, Luke Zettlemoyer (2017). Neural AMR: sequence-to-sequence models for parsing and generation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol 1: Long Papers*, Vancouver, Canada, https://aclanthology.org/P17-1014, 146-157.Diego Marcheggiani, Jasmijn Bastings, Ivan Titov (2018). Exploiting semantics in neural machine translation with graph convolutional networks. *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana,* 486-492.

Long H. B. Nguyen, Viet H. Pham, Dien Dinh (2021). Improving neural machine translation with AMR semantic graphs. Mathematical Problems in Engineering. https://doi.org/10.1155/2021/9939389.

Martha Palmer, Dan Gildea, Paul Kingsbury (2005). The proposition bank: a corpus annotated with semantic roles. *Computational Linguistics*, 31:1, 71–106.

Reinhard Rapp (2021). Similar language translation for Catalan, Portuguese and Spanish Using Marian NMT. *Proceedings of the Sixth Conference on Machine Translation; Anthology of the Association for Computational Linguistics*, https://aclanthology.org/2021.wmt-1.31, 292–298.

Reinhard Rapp (2022). Neural machine translation explained. How to build your own neural machine translation system using Marian NMT. *tcworld magazine* January 2022, 12–18. https://tcworld.info/e-magazine/translation-and-localization/neural-machine-translation-explained-1167/

Sennrich, Rico; Barry Haddow, Alexandra Birch (2016). Neural machine translation of rare words with subword units. *Proc. of the 54th Annual Meeting of the ACL (Vol. 1: Long Papers)*, 1715–1725.

Peng Shi and Jimmy Lin (2019). Simple BERT models for relation extraction and semantic role labeling. *ArXiv*, 1904.05255.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jacob; Jones, Llion; Gomze, Aidan N. G.; Kaiser, Lukasz; Polosukhin, Illia (2017). Attention is all you need. *Proceedings of the 31st Confer-ence on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 6000–6010.

Ralph Weischedel, Eduard Hovy, Martha Palmer, Mitchell Marcus, Robert Belvin, Sameer Pradhan, Lance Ramshaw, Nianwen Xue (2011). OntoNotes: A large training corpus for enhanced processing. In J. Olive, C. Christianson, and J. McCary (editors): *Handbook of Natural Language Processing and Machine Translation.* Springer.

Nier Wu, Hongxu Hou, Haoran Li, Xin Chang, Xiaoning Jia (2021). Semantic perception-oriented low-resource neural machine translation. In: *Jinsong Su, Rico Sennrich: Machine Translation. 17th China Conference, CCMT 2021, Xining, China, October 8–10, 2021, Revised Selected Papers*. Springer Singapore. http://sc.cipsc.org.cn/mt/conference/2021/papers/T21-1009.pdf