Syntax-aware Neural Machine Translation Using CCG

Article · February 2017

7 authors, including:

Some of the authors of this publication are also working on these related projects:

Project   https://github.com/emjotde/Marian View project

Project   TraMOOC View project

Syntax-aware Neural Machine Translation Using CCG

Maria Năjedje[1] and **Siva Reddy**[1] and **Rico Sennrich**[1] and **Tomasz Dwojak**[1,2]
Marcin Junczys-Dowmunt[2] and **Philipp Koehn**[3] and **Alexandra Birch**[1]
[1]School of Informatics, University of Edinburgh
[2]Adam Mickiewicz University
[3]Dep. of Computer Science, Johns Hopkins University
m.nadejde@sms.ed.ac.uk,{siva.reddy, a.birch, rico.sennrich}@ed.ac.uk
{t.dwojak,junczys}@amu.edu.pl, phi@jhu.edu

Abstract

Neural machine translation (NMT) models are able to partially learn syntactic information from sequential lexical information. Still, some complex syntactic phenomena such as prepositional phrase attachment are poorly modeled. This work aims to answer two questions: 1) Does explicitly modeling source or target language syntax help NMT? 2) Is tight integration of words and syntax better than multitask training? We introduce syntactic information in the form of CCG supertags either in the source as an extra feature in the embedding, or in the target, by interleaving the target supertags with the word sequence. Our results on WMT data show that explicitly modeling syntax improves machine translation quality for English↔German, a high-resource pair, and for English↔Romanian, a low-resource pair and also several syntactic phenomena including prepositional phrase attachment. Furthermore, a tight coupling of words and syntax improves translation quality more than multitask training.

1 Introduction

Sequence-to-sequence neural machine translation (NMT) models (Sutskever et al., 2014; Cho et al., 2014b; Bahdanau et al., 2015) are state-of-the-art on a multitude of language-pairs (Sennrich et al., 2016a; Junczys-Dowmunt et al., 2016). Part of the appeal of neural models is that they can learn to implicitly model phenomena which underlie high quality output, and some syntax is indeed captured by these models. In a detailed analysis, Bentivogli et al. (2016) show that NMT significantly improves over phrase-based SMT, in par-

ticular with respect to morphology and word order, but that results can still be improved for longer sentences and complex syntactic phenomena such as prepositional phrase (PP) attachment. Another study by Shi et al. (2016) shows that the encoder layer of NMT partially learns syntactic information about the source language, however complex syntactic phenomena such as coordination or PP attachment are poorly modeled.

Recent work which incorporates additional linguistic information in NMT models (Luong et al., 2016; Sennrich and Haddow, 2016) show that even though neural models have strong learning capabilities, explicit features can still improve translation quality. In this work we perform a thorough investigation of rich syntactic features in NMT. We examine the benefit of adding syntactic information in the source, as an extra feature in the embedding layer following the approach of Sennrich and Haddow (2016). We also propose a method for generating syntactic information in the target: tightly coupling words and syntax by interleaving target syntactic representation with the word sequence. We compare this to loosely coupling words and syntax using multitask solutions, where the shared parts of the model are trained to produce either a target sequence of words or supertags in a similar fashion to Luong et al. (2016).

We use CCG syntactic categories (Steedman, 2000), also known as *supertags*, to represent syntax explicitly. Supertags provide global syntactic information locally at the lexical level. They encode subcategorization information, capturing short and long range dependencies and attachments, and also tense and morphological aspects of the word in a given context. Consider the sentence in Figure 1. This sentence contains two PP attachments and could lead to several disambiguation possibilities ("*in*" can attach to "*Netanyahu*" or "*receives*", and "*of*" can attach to "*capital*",

"Netanyahu" or *"receives"*). These alternatives may lead to different translations in other languages. However the supertag S\NP/PP/NP of *"receives"* indicates that the preposition *"in"* attaches to the verb, and the supertag NP\NP/NP of *"of"* indicates that it attaches to *"capital"*, thereby resolving the ambiguity.

Our research contributions are as follows:

- We show that both source and target language syntax improves translation quality for English↔German, English↔Romanian as measured by BLEU.

- We present a fine grained analysis of Syntax-aware NMT (SNMT) and show consistent gains when looking at different linguistic phenomena and sentence lengths.

- We propose three novel approaches to integrating target syntax at word level in the decoder, by serializing CCG supertags in the target word sequence and by multitasking with either a shared or distinct attention model and decoder.

- Our results suggest that a tight coupling of target words and syntax (by serializing) improves translation quality more than the decoupled signal from multitask training.

2 Related work

Syntax has helped in statistical machine translation (SMT) to capture dependencies between distant words that impact morphological agreement, subcategorisation and word order (Galley et al., 2004; Menezes and Quirk, 2007; Williams and Koehn, 2012; Sennrich, 2015; Chiang, 2007). There has been some work in NMT on modeling source-side syntax implicitly or explicitly. Kalchbrenner and Blunsom (2013); Cho et al. (2014a) capture the hierarchical aspects of language implicitly by using convolutional neural networks. Sennrich and Haddow (2016) generalize the embedding layer of NMT to include explicit linguistic features such as dependency relations and part-of-speech tags and we extend their work to using CCG supertags. Other architectures have been proposed to integrate source-side syntax such as Eriguchi et al. (2016) who use the phrase structure of the source sentence to guide the recurrence and attention model in a tree-to-sequence model. Luong et al. (2016) co-train a translation model and a syntactic parser which share the encoder. Our

multi-task models extend their work to attention-based NMT models and we implement a novel multi-task architecture where the tasks share not just the encoder, but also the attention model and the decoder, keeping only the softmax layer separate.

Applying more tightly coupled linguistic factors on the target for NMT has been previously investigated. Niehues et al. (2016) proposed a factored RNN-based language model for re-scoring an n-best list produced by a phrase-based MT system. In recent work Martínez et al. (2016) implemented an NMT model which first generated lemmas and morphology, and then used these to generate the word form. Unfortunately no real gain was reported for these experiments. In our work we do not focus on model architectures, and instead we explore the more general problem of including syntax in NMT: comparing source and target-side syntax and comparing tightly and loosely coupled syntactic information.

Previous work on integrating CCG supertags in factored phrase-based models (Birch et al., 2007) showed promising results. However the factored models originally proposed for statistical machine translation (Koehn and Hoang, 2007) suffered from data sparsity and did not consider longer sequences as context. In this work we take advantage of the expressive power of neural networks to learn representations that generate both words and CCG supertags.

3 Modeling Syntax in NMT

CCG is a lexicalised formalism in which words are assigned with syntactic categories, i.e., supertags, that indicate context-sensitive morpho-syntactic properties of a word in a sentence. The combinators of CCG allow the supertags to capture global syntactic constraints locally. Though NMT captures long range dependencies using long-term memory, short-term memory is cheap and reliable. Supertags can help by allowing the model to rely more on local information (short-term) and not having to rely heavily on long-term memory.

Consider a decoder that has to generate the following sentences:

1. What$_{(S[wq]/(S[q]/NP))/N}$ city is$_{(S[q]/PP)/NP}$ the Taj Mahal in?

2. Where$_{S[wq]/(S[q]/NP)}$ is$_{(S[q]/NP)/NP}$ the Taj Mahal?

Source-side

BPE:	Obama	receives	Net+	an+	yahu	in	the	capital	of	USA
IOB:	O	O	B	I	E	O	O	O	O	O
CCG:	NP	S\NP/PP/NP	NP	NP	NP	PP/NP	NP/N	N	NP\NP/NP	NP

Target-side

NP Obama S\NP/PP/NP receives NP Net+ an+ yahu PP/NP in NP/N the N capital NP\NP/NP of NP USA

Figure 1: Source and target representation of syntactic information in syntax-aware NMT.

If the decoding starts with predicting "*What*", it is ungrammatical to omit the preposition "*in*", and if the decoding starts with predicting "*Where*", it is ungrammatical to predict the preposition. Here the decision to predict "*in*" depends on the first word, a long range dependency. However if we rely on CCG supertags, the supertags of both these sequences look very different. The supertag (S[q]/PP)/NP for the verb "*is*" in the first sentence indicates that a preposition is expected in future context. Furthermore it is likely to see this particular supertag of the verb in the context of (S[wq]/(S[q]/NP))/N but it is unlikely in the context of S[wq]/(S[q]/NP). Therefore a succession of local decisions based on CCG supertags will result in the correct prediction of the preposition in the first sentence, and omitting the preposition in the second sentence. Since the vocabulary of CCG supertags is much smaller than that of possible words, the NMT model will do a better job at generalizing over and predicting the correct CCG supertags sequence.

CCG supertags also help during encoding if they are given in the input, as we saw with the case of PP attachment in Figure 1. Translation of the correct verb form and agreement can be improved with CCG since supertags also encode tense, morphology and agreements. For example, in the sentence "*It is going to rain*", the supertag (S[ng]\NP[expl])/(S[to]\NP) of "*going*" indicates the current word is a verb in continuous form looking for an infinitive construction on the right, and an expletive on the left.

We explore the effect of syntax by using CCG supertags either on the source-side (encoder) or target-side (decoder) as follows.

Source-side syntax When modeling the source-side syntactic information, we include the CCG supertags as extra features in the NMT encoder using the framework of (Sennrich and Haddow, 2016). The model of (Bahdanau et al., 2015) is extended by learning a separate embedding for several source-side features such as the word itself or its part-of-speech. All feature embeddings are concatenated into one embedding vector which is used in all parts of the encoder model instead of the word embedding.

The baseline features are the subword units obtained using *byte-pair-encoding* (BPE, (Sennrich et al., 2016b)) together with the annotation of the subword structure using IOB format by marking if a symbol in the text forms the beginning (B), inside (I), or end (E) of a word. A separate tag (O) is used if a symbol corresponds to the full word. The word level supertag is replicated for each BPE unit.

Figure 1 gives an example of the source-side feature representation. It shows the segmentation of the noun *Netanyahu* into three BPE sub-units *Net+ an+ yahu* as well as the duplication of the supertag *NP* for each of these sub-units.

Target-side syntax When modeling the target-side syntactic information we consider different strategies of coupling the CCG supertags with the translated words in the decoder: serializing, multi-tasking with shared encoder and multitasking with distinct softmax layer.

- **Serializing** In this paper we propose a tight integration in the decoder of the syntactic representation and the surface forms. For each target word we include its supertag as an extra token before the first BPE sub-unit of the word. With this arhitecture a single decoder learns to predict both the target supertags and the target words conditioned on previous syntactic and lexical context. Figure 1 gives an example of the target-side representation in the case of serializing. The supertag *NP* corresponding to the word *Netanyahu* is included only once before the three BPE subunits *Net+ an+ yahu*.

- **Multitasking (1) – shared encoder** A loose coupling of the syntactic representation and the surface forms can be achieved by co-training a

translation model with a secondary prediction task, in our case CCG supertagging. In the multitask framework (Luong et al., 2016) the encoder part is shared while the decoder is different for each of the prediction tasks: translation and tagging. Different from Luong et al. we also train a separate attention model for each task. Another difference is that we experiment with both source and target syntax as a secondary task, while Luong et al. only used source syntax. We use EasySRL to label the parallel corpus with CCG supertags instead of using a corpus with gold annotations.

- **Multitasking (2) – distinct softmax** In this approach only the softmax layer is distinct for each task while the encoder, attention model and decoder are shared. The input to the two softmax layers is indentical (the context vector, the previous hidden state and the previous predicted word) but the second softmax layer predicts CCG supertags. The cost of predicting the wrong supertag is added to the cost of predicting the wrong target word. We do not add a connection between the CCG supertag at time step $t - 1$ and the word predicted at time step t as that would require a modification of the beam search.

Each of the three models has its disadvantages. The *serializing* approach increases the length of the target sequence which might lead to loss of information learned at lexical level. For the *multitasking (1)* approach there is no explicit way to constrain the number of predicted words and tags to match. The *multitasking (2)* approach does not condition the prediction of target words on the syntactic context.

4 Experimental Setup and Evaluation

4.1 Data and methods

We train the neural MT systems on all the parallel data available at WMT16 (Bojar et al., 2016) for the German↔English and Romanian↔English language pairs. The English side of the parallel data is annotated with CCG lexical tags using EasySRL (Lewis et al., 2015). Some longer sentences cannot be processed by the parser and therefore we eliminate them from our training and test data. We report the sentence counts for the filtered data sets in Table 1. During training we validate our models with BLEU on devel-

	train	dev	test
DE-EN	4,468,314	2,986	2,994
RO-EN	605,885	1,984	1,984

Table 1: Number of sentences in the training, development and test sets.

opment sets: newstest2013 for German↔English and newsdev2016 for Romanian↔English. We evaluate the systems using BLEU (Papineni et al., 2002) and report results on newstest2016 for both language pairs.

All the neural MT systems are attentional encoder-decoder networks (Bahdanau et al., 2015) as implemented in the Nematus toolkit.[1] We use similar hyper-parameters to those reported by (Sennrich et al., 2016a) with minor modifications: we used mini-batches of size 60 and Adam optimizer (Kingma and Ba, 2014).

Words are segmented into sub-units that are learned jointly for source and target using BPE (Sennrich et al., 2016b), resulting in a vocabulary size of 85,000. The vocabulary size for CCG supertags was 500.

We select the best single models according to BLEU on the development set and use the four best single models for the ensembles. Due to small differences in pre-processing with CCG supertags, our results for the baseline systems are similar but do not match the results reported by (Sennrich et al., 2016a).

For the experiments with source-side features we use the BPE sub-units and the IOB tags as baseline features. We keep the total word embedding size fixed to 500 dimensions. When adding the CCG feature we allocate 135 dimensions for CCG supertags, 360 for BPE sub-units and 5 for IOB tags.

For the experiments where we add CCG supertags in the target sequence we increase the maximum length of sentences from 50 to 100. On average the length of English sentences for newstest2013 in BPE representation is 22.7, while the average length when adding the CCG supertags is 44. Increasing the length of the target recurrence results in larger memory consumption and slower training.[2]

---

[1]https://github.com/rsennrich/nematus
[2]Roughly 10h30 per 100,000 sentences (20,000 batches) for SNMT compared to 6h for NMT.

| model | single | | ensemble |
	dev	test	test
NMT	22.28	27.26	28.25
*SNMT	22.48	27.70	**28.95**
*Multitask(1)	22.03	27.69	28.73

Table 2: BLEU scores for baseline NMT, syntax-aware NMT (SNMT) and multiasking with shared encoder for English→German. * indicates source-side CCG supertags.

| model | single | | ensemble |
	dev	test	test
NMT	22.32	21.39	21.51
*SNMT	22.86	21.83	**21.87**

Table 3: BLEU scores for baseline NMT and syntax-aware NMT (SNMT) with source-side CCG supertags for English→Romanian.

4.2 Results

In this section we evaluate whether our syntax-aware NMT model (SNMT) with source-side and target-side CCG supertags improves translation quality as compared to a baseline NMT model (Bahdanau et al., 2015; Sennrich et al., 2016a). In tables and figures we use the symbol "*" to indicate that syntactic information is used on the source (eg. *de-en), the target (eg. de-en*) or both (eg. *de-en*).

CCG on the source-side We first evaluate the impact of source-side CCG features on over-all translation quality. We report results for English→German, a high-resource language pair, and for English→Romanian, a low-resource language pair. Both the target languages are morphologically richer than the source language, with German exhibiting more word order flexibility than Romanian.

We report BLEU scores in Table 2 and Table 3 for both the best single models and ensemble models. However, we will only refer to the results with ensemble models since these are generally better.

Adding the source-side CCG feature improves translation quality for both language pairs as well as for both single and ensemble models. For English→German BLEU increases by up to 0.7 points and for English→Romanian by up to 0.4 points.

Although the training data for English→German is large, the CCG supertags still improve translation quality. This suggests

that syntax provides complementary information that the baseline NMT model is not able to learn from the source word sequence alone.

Next we make a finer grained analysis of the impact of source-side syntactic features by look-ing at a breakdown of BLEU scores with respect to different linguistic constructions. We classify sentences into different linguistic constructions based on the CCG supertags that appear in them, e.g., the presence of category (NP\NP)/(S/NP) indicates subordinate construction. Figure 2 shows the breakdown of BLEU scores for the following linguistic constructions: coordination *(conj)*, control and raising *(control)*, prepositional phrase attachment *(pp)*, questions and subordinate clauses *(subordinate)*. We report the number of sentences for each category in Table 4.
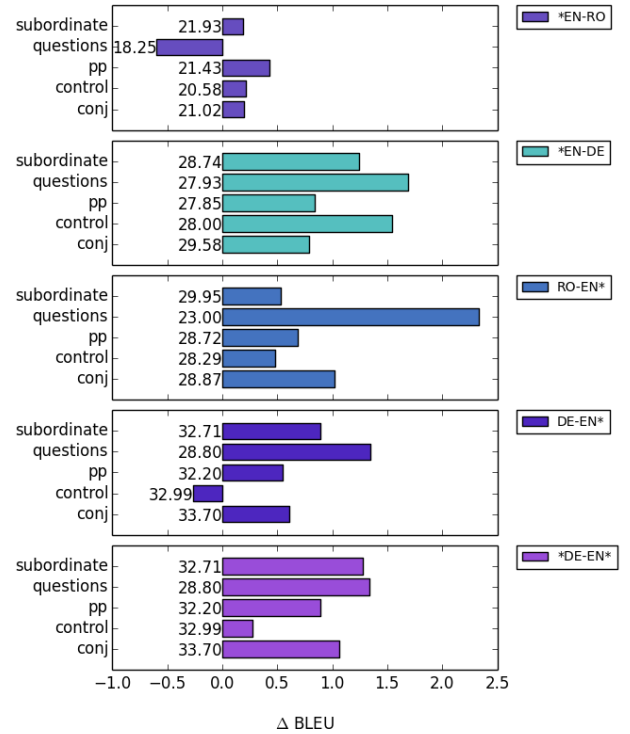

Figure 2: Difference in BLEU points between SNMT and NMT wrt different linguistic constructs. Baseline NMT scores are shown.

	sub.	qu.	pp	contr.	conj
RO↔EN	742	90	1,572	415	845
DE↔EN	936	114	2,321	546	1,129

Table 4: Sentence counts for different linguistic constructions.

We see a consistent improvement for English→German for all linguistic constructions when using the CCG supertags, especially

for questions which involve a lot of movement, and control constructions which involve long range verb agreements. In the case of English→Romanian, where word order is closer between languages, we still see improvements across most constructions when using the CCG supertags.

CCG on the target-side Here we evaluate the impact of using the target-side CCG supertags for the Romanian→English and German→English translation directions. In Table 5 and Table 6 we report results with BLEU.

The target-side CCG feature improves BLEU scores by 0.9 for Romanian→English. For German→English, the ensemble model with target-side CCG supertags improves BLEU scores by 0.6. These results suggest that the baseline NMT decoder benefits from modeling the global syntactic information locally via supertags.

model	single		ensemble
	dev	test	test
NMT	29.68	28.13	28.43
SNMT*	30.44	29.16	**29.32**

Table 5: BLEU scores for baseline NMT and syntax-aware NMT (SNMT) with target-side CCG supertags for Romanian→English.

model	single		ensemble
	dev	test	test
NMT	26.40	31.03	32.09
SNMT*	27.73	31.96	**32.65**
Multitask(1)*	26.68	31.38	32.03
Multitask(2)*	26.63	31.01	31.99

Table 6: BLEU scores for baseline NMT, syntax-aware NMT (SNMT), multiasking with shared encoder (1) and multitasking with distinct softmax (2) for German→English. * indicates target-side CCG supertags.

We also see consistent improvements across most linguistic constructions for both language pairs as reported in Figure 2. In particular, the increase in BLEU scores for the *prepositional phrase* and *subordinate* constructions suggests that target word order is improved. For the *control and raising* constructions we don't see any improvement for German→English, while the improvement for English→German was 1.5 BLEU points. These results suggest that syntactic information is more

model	syntax	single	ensemble
NMT	na	31.03	32.09
*SNMT	dep	31.41	32.19
SNMT*	ccg	31.96	32.65
SNMT	dep & ccg	32.13	**33.03**

Table 7: Comparison of NMT and syntax-aware NMT (SNMT) for German→English.

helpful for the *control and raising* constructions when translating into German.

CCG on both source-side and target-side Next we combine source-side and target-side syntax in a German→English syntax-aware NMT system. We use CCG supertags on the target-side and dependency labels as source-side features. Although the dependency labels do not encode global syntactic information they disambiguate the grammatical function of words. In Table 7 we compare systems that use syntax either on the source-side or target-side with a system that uses syntax on both sides. First we observe that the source-side dependency labels improve translation quality by only 0.1 BLEU points as compared to the baseline NMT system, while the improvement with target-side CCG supertags is more than 0.5 BLEU points. When using both source-side and target-side factors the syntax-aware NMT system performs better than the baseline by roughly 1 BLEU point.

Next we observe from Figure 2 that having both source and target syntax helps improve BLEU scores for all syntactic phenomena. As compared to the SNMT system with only target syntax, BLEU scores improve by 0.5 for *control and raising* constructs, 0.3 for *prepositional phrase attachment*, 0.4 for *subordinate* constructs and 0.4 for sentences involving *coordination*.

Finally we compare the systems with respect to sentence length. Figure 3 shows the difference in BLEU points between the baseline NMT system and the syntax-aware NMT system with respect to the length of the source sentence measured in BPE sub-units. We report the number of sentences for each category in Table 8. For the majority of systems and sentence lengths adding CCG supertags helps. When using target CCG supertags for German→English there is a decrease in BLEU for short sentences. However when using both source and target syntax, BLEU improves in this case as well. Furthermore combining the

source and target syntax seems to help the most for longer sentences. This is an encouraging result since including target CCG supertags increases the target sequence which may lead to information being forgotten.
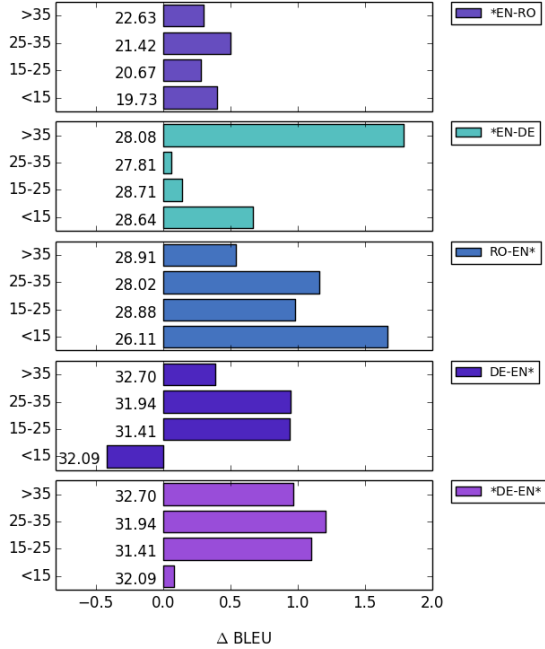

Figure 3: Difference in BLEU points between SNMT and NMT wrt sentence length. Baseline NMT scores are shown.

	<15	15-25	25-35	>35
RO↔EN	491	540	433	520
DE↔EN	918	934	582	560

Table 8: Sentence counts for different sentence lengths.

We conclude that when syntactic information is available for both the source language and the target language it is most beneficial to integrate syntax in both the encoder and the decoder.

Serializing vs multitasking One open question we address in this work is how tightly do we need to integrate syntax in the NMT system. For this purpose we compare different ways of coupling the target word sequence and CCG supertags: *serializing*, *multitasking (1)* with shared encoder and *multitasking (2)* with distinct softmax layer. In our proposed SNMT system with target CCG supertags we take the *serializing* approach. The different approaches are described in Section 3. For the multitasking (1) approach we explore two scenarios: in the first scenario the secondary task is to predict the source-side CCG supertags while in the second scenario it is to predict the target-side CCG supertags.

First we compare the multitask (1) NMT system predicting source CCG supertags and the SNMT system with CCG features on the source. The results in Table 2 show that multitasking also improves BLEU scores, suggesting that the encoder learns a better latent syntactic representation of the source sentence. However the SNMT system performs slightly better, by 0.2 BLEU points.

Next we compare a SNMT system with target CCG supertags and a multitask (1) NMT system which predicts target CCG supertags as a secondary task. The results in Table 6 show that the multitask approach does not improve BLEU scores as compared to the baseline NMT system, suggesting that the shared encoder is not able to learn a good representation of target syntax. In contrast, the SNMT system which integrates syntax in the decoder improves translation by 0.5 BLEU points.

Finally the multitask (2) system which predicts independently the target CCG supertags and words does not perform better than the NMT baseline. The BLEU scores are shown in Table 6. Since neither of the two *multitasking* models improve translation quality as measured by BLEU we conclude that a tight integration of the target syntax and word sequence is important. Conditioning the prediction of words on their corresponding CCG supertag is what sets SNMT apart from the other models and improves the translation quality.

4.3 Discussion

Our preliminary experiments show that source syntax improves translation when translating from a morphologically poor language into a morphologically rich language. It would be interesting to evaluate the impact of target syntax for such a language pair. In the future we plan to use the Hindi CCGBank (Ambati et al., 2016) to run experiments for English↔Hindi.

Although the focus of this paper is not improving CCG tagging, we can measure that SNMT is accurate at predicting CCG supertags. We compare the CCG sequence predicted by the SNMT models with that predicted by EasySRL and obtain the following accuracies: 93.2 for Romanian→English, 95.6 for German→English, 95.8 for German→English with both source and target syntax. This result might be useful for

	DE - EN* Question
Source	Oder wollen Sie herausfinden , **über** was andere reden ?
Ref.	Or do you want to find out what others are talking **about** ?
NMT	Or would you like to find out **about** what others are talking **about** ?
SNMT*	Or do you want to find out what$_{NP/(S[dcl]/NP)}$ others are$_{(S[dcl]\backslash NP)/(S[ng]\backslash NP)}$ talking$_{(S[ng]\backslash NP)/PP}$ **about**$_{PP/NP}$?

	DE - EN* Subordinate
Source	...dass die Polizei jetzt sagt , ..., und dass Lamb in seinem Notruf **Prentiss zwar als seine Frau bezeichnete** ...
Ref.	...that police are now saying ..., and that while Lamb **referred to Prentiss as his wife** in the 911 call ...
NMT	...police are now saying ..., and that in his emergency call **Prentiss he called his wife** ...
SNMT*	...police are now saying ..., and that lamb , in his emergency call , **described**$_{((S[dcl]\backslash NP)/PP)/NP}$ **Prentiss as his wife** ....

	DE - EN Subordinate
Source	Prentiss war eine Krankenschwester , die für verschiedene Unternehmen **online arbeitete** .
Ref.	Prentiss was a nurse who **worked** for various companies online .
NMT	Prentiss was a nurse who **worked** for various companies **online** .
SNMT*	Prentiss was a nurse who **worked**$_{(S[dcl]\backslash NP)/PP}$ for various companies .
SNMT	Prentiss was a nurse who **worked**$_{((S[dcl]\backslash NP)/PP)/NP}$ **online** for several companies .

	RO - EN* Coordination
Source	Sustinerea in randul barbatilor s-a redus considerabil iar Sanders se afla la o distanta de doar 5 puncte .
Reference	Her support among men has dropped considerably and Sanders only trails her by 5 points .
NMT	Support for men has decreased considerably , **and** Sanders **are** at a distance of just 5 points .
SNMT*	Support among men has diminished considerably , **and** Sanders **is** at a distance of just 5 points .

Figure 4: Comparison of baseline NMT and syntax-aware NMT (SNMT) for Romanian→English and German→English. * indicates syntax used on the source or target.

downstream applications using CCG supertagging, such as multi-lingual question answering.

We conclude by giving a few examples in Figure 4 for which the syntax-aware NMT system produced more grammatical translations than the baseline NMT system.

In the example **DE-EN* Question** the baseline NMT system translates the preposition *"über"* twice as *"about"*. The SNMT predicts the correct CCG supertag for *"what"* which expects to be followed by a sentence and not a preposition: NP/(S[dcl]/NP). Therefore the SNMT correctly re-orders the preposition *"about"* at the end of the question.

In the example **DE-EN* Subordinate** the baseline NMT system fails to correctly attach *"Prentiss"* as an object and *"his wife"* as a modifier to the verb *"called (bezeichnete)"* in the subordinate clause. In contrast the SNMT system predicts the correct sub-categorization frame of the verb *"described"* and correctly translates the entire predicate-argument structure.

In the example ***DE - EN* Subordinate** the baseline NMT system translates the sentence correctly while the SNMT drops the argument *"online"*. The SNMT* predicts the wrong CCG supertag (S[dcl]\NP)/PP for the verb *"worked (arbeitete)"*, which guides the system to only translate the prepositional modifier and drop the sec-

ond argument. However the *SNMT* system which also has source syntactic factors predicts the correct subcategorization frame for the verb ((S[dcl]\NP)/PP)/NP and subsequently translates the second argument.

In the example **RO-EN* Coordination** the baseline NMT system makes an agreement mistake between the subject *"Sanders"* and the verb *"are"*. In contrast the SNMT system correctly identifies that the coordination is between sentences, and outputs the correct verb form *"is"*.

5 Conclusions

Our results suggest that having the notion of explicit syntax, here in the form of CCG supertags, in the encoder or the decoder improves machine translation for both English↔German and English↔Romanian language pairs. Earlier work on syntax-aware NMT mainly modeled syntax in the encoder while our results suggest modeling syntax in the decoder is also useful. Moreover by modeling syntax in both encoder and decoder we obtain the most improvement over the baseline NMT system, in particular for longer sentences and syntactic phenomena such as prepositional attachment and coordination. Finally our results show that a tight integration of syntax in the decoder improves translation quality while decoupling of target words and syntax does not.

References

Bharat Ram Ambati, Tejaswini Deoskar, and Mark Steedman. 2016. Hindi CCGbank: CCG Treebank from the Hindi Dependency Treebank. In *Language Resources and Evaluation*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)..*

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016.* pages 257–267.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. Ccg supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, StatMT '07, pages 9–16.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.* 33(2):201–228.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, pages 103–111.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 823–833.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '04.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the IWSLT 2016*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 868–876.

Mike Lewis, Luheng He, and Luke Zettlemoyer. 2015. Joint a* ccg parsing and semantic role labelling. In *Empirical Methods in Natural Language Processing*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of International Conference on Learning Representations (ICLR 2016)*.

Mercedes García Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored Neural Machine Translation Architectures. In *International Workshop on Spoken Language Translation (IWSLT'16)*.

Arul Menezes and Chris Quirk. 2007. Using dependency order templates to improve generality in translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. pages 1–8.

Jan Niehues, Thanh-Le Ha, Eunah Cho, and Alex Waibel. 2016. Using factored word representation in neural network language models. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational

Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318.

Rico Sennrich. 2015. Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation. *Transactions of the Association for Computational Linguistics* 3:169–182.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 83–91.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1526–1534.

Mark Steedman. 2000. *The syntactic process*, volume 24. MIT Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. NIPS'14, pages 3104–3112.

Philip Williams and Philipp Koehn. 2012. Ghkm rule extraction and scope-3 parsing in moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. pages 388–394.

A Appendix

We give a few more examples in Figure 5 and Figure 6.

In the example **RO-EN Preposition** the baseline NMT system drops the preopositional modifier of the verb *"consider"*. The SNMT predicts the correct subcategorization frame for the verb *see* and correctly translates the prepositional modifier *"as a problem"*.

In the second **DE-EN Subordinate** example the baseline NMT inverts the role of the arguments for the verb *"support (unterstützen)"*. The SNMT system produces the correct syntactic structure of the subordinate clause and translates *"sie"* as *"they"*, the subject.

In the example **EN-DE Raising** the baseline NMT system translates the raising construct *"wants ... to be seen"* with the incorrect infinitive verb form *"zu sehen"*. In contrast the SNMT system produces the correct translation for a subordinate sentence *"gesehen werden"*. Furthermore the SNMT system produces the correct nominative inflection for the coordinated subject of the raising construct *"seine Mitgliedschaft im Schachclub ... und sein freundlicher Kontakt"*, while the NMT system inflects the second part as accusative *"seinen freundlichen Kontakt"*.

In the example **EN-DE Suboordination** the baseline NMT system mistranslates the subordinate clause *"which lists 17 faculty members"* as *"die 17 Fakultäten Mitglieder"* which drops the verb *"lists"* and the relative pronoun *"which"*. In contrast the SNMT correctly translates the verb at the end of the clause as well as the relative pronoun *"in denen 17 Fakultätsmitglieder aufgeführt sind"*. A mistake that both system make is the incorrect disambiguation of the verb *"took"* which is translated as *"nahmen"* instead of *"besuchten"*.

In the second **EN-DE Suboordination** example the baseline NMT system mistranslates the subordinate clause *"who say the same of Trump"*, as it fails to correctly order the target verb at the end of the clause *"die sagen , das Gleiche von Trump"*. In contrast the SNMT system translates the verb at the end of the subordinate clause *"die das Gleiche von Trump sagen"*.

In the example **EN-DE Question and Coordination** the baseline NMT system does not predict the correct target order of the verb *"waste (vergeuden)"* and its direct object *"political capital (politisches Kapital)"*. In contrast the SNMT system correctly reorders the target verb *"verschwenden"* at the end of the clause. Moreover the SNMT system correctly identifies the coordinated subject *"Paris or Berlin"* and correctly inflects the auxiliary verb *"should"* to the plural form in German *"sollten"*.

	RO - EN* Preposition
Source	Majoritatea republicanilor nu **consideră** temperamentul lui Trump **o problem**.
Reference	A majority of Republicans don't **see** Trump's temperament **as a problem**.
NMT	The Republican majority do not **consider** Trump's temperament.
SNMT	Most republicans do not **see**$_{((S[b]\backslash NP)/PP)/NP}$ Trump's temperate **as a problem**.

	DE - EN* Subordinate
Source	Mehr als fünf Monate vor Beginn der Vorwahlen , sagen die meisten demokratischen Wähler , dass es zu früh ist , um zu sagen , dass ihre Meinung feststeht , **welchen Kandidaten sie unterstützen werden** .
Reference	More than five months before the start of the primary contests , most Democratic voters say it is too early to say that their minds are made up about **which candidate they will support** .
NMT	More than five months before the start of the pre-elections , most democratic voters say that it is too early to say that their opinion is defined , **which candidates will support them** .
SNMT	More than five months before the start of the preliminary elections , most democratic voters say that it is too early to say that their opinion is determined , **which candidates they will support** .

Figure 5: Comparison of baseline NMT and syntax-aware NMT (SNMT) with target-side CCG supertags for Romanian→English and German→English.

	EN - DE Raising
Source	Gauselmann wants **his membership** of the chess club as well as **his friendly contact** with the " Red-white " tennis club **to be seen** as an expression of his ties with the spa town .
Reference	Gauselmann wünscht sich , dass **die Mitgliedschaft** im Schachclub und auch **freundschaftliche Kontakt** zum Tennisclub " Rot-Weiss " als Ausdruck seiner Verbundenheit mit der Kurstadt **gesehen wird** .
NMT	Gauselmann will **seine Mitgliedschaft** im Schachclub und **seinen freundlichen Kontakt** mit dem " Red-white " Tennisclub als Ausdruck seiner Bande mit der Kurstadt **zu sehen** .
SNMT	Gauselmann wünscht sich , dass **seine Mitgliedschaft** im Schachclub und **sein freundlicher Kontakt** mit dem " Red-Weiss " Tennisclub als Ausdruck seiner Beziehungen zum Kurort **gesehen werden** .

	EN - DE Subordinate
Source	Both taught in the Division of Social Sciences and History , **which lists 17 faculty members** , and many students took courses from both .
Reference	Beide unterrichteten in der Abteilung für Sozialwissenschaften und Geschichte , **deren Lehrkörper 17 Mitglieder umfasst** , und viele Studenten besuchten Kurse von beiden .
NMT	Beide unterrichten in der Abteilung der Sozialwissenschaften und der Geschichte , **die 17 Fakultäten Mitglieder** , und viele Studenten nahmen Kurse aus beiden
SNMT	Beide unterrichten in der Abteilung der Sozialwissenschaften und der Geschichte , **in denen 17 Fakultätsmitglieder aufgeführt sind** , und viele Studenten nahmen an .

	EN - DE Subordinate
Source	Only Ben Carson generates roughly the same level of enthusiasm as Trump (43 percent say they would be " enthusiastic " vs. 40 percent **who say the same of Trump**) .
Reference	Nur Ben Carson schafft ungefähr die gleiche Begeisterung wie Trump (43 Prozent sagen , sie wären " begeistert " vs. 40 Prozent , **die das gleiche über Trump sagen**) .
NMT	Nur Ben Carson erzeugt ungefähr das gleiche Mass an Begeisterung wie Trump (43 Prozent sagen , sie wären " enthusiastisch " vs. 40 Prozent , **die sagen , das Gleiche von Trump**) .
SNMT	Nur Ben Carson erzeugt ungefähr das gleiche Mass an Enthusiasmus , wie Trump (43 Prozent sagen , sie wären " enthusiastisch " gegen 40 Prozent , **die das Gleiche von Trump sagen**) .

	EN - DE Question and Coordination
Source	Why **should** Paris or Berlin **waste** political capital - they have suspicious voters too - on concessions to Britain when it may all be pointless ?
Reference	Warum **sollten** Paris oder Berlin politisches Kapital **verschwenden** , - sie haben auch misstrauische Wähler - um Zugeständnisse an die Briten zu machen , wenn dies alles sinnlos sein kann ?
NMT	Warum **sollte** man Paris oder Berlin **vergeuden** politisches Kapital - sie haben auch verdächtige Wähler - bei Konzessionen gegenüber Grossbritannien , wenn das alles sinnlos sein könnte ?
SNMT	Warum **sollten** Paris oder Berlin das politische Kapital - auch sie haben verdächtige Wähler - mit Zugeständnissen an Grossbritannien **verschwenden** , wenn es überhaupt sinnlos wäre ?

Figure 6: Comparison of baseline NMT and syntax-aware NMT (SNMT) with source-side CCG supertags for English→German and English→Romanian.