Semantics-aware Attention Improves Neural Machine Translation

Anonymous ACL submission

Abstract

The integration of syntactic structures into Transformer machine translation has shown positive results, but to our knowledge, no work has attempted to do so with semantic structures. In this work we propose two novel parameter-free methods for injecting semantic information into Transformers, both rely on semantics-aware masking of (some of) the attention heads. One such method operates on the encoder, through a Scene-Aware Self-Attention (SASA) head. Another on the decoder, through a Scene-Aware Cross-Attention (SACrA) head. We show a consistent improvement over the vanilla Transformer and syntax-aware models for four language pairs. We further show an additional gain when using both semantic and syntactic structures in some language pairs.

1 Introduction

It has long been argued that semantic representation can benefit machine translation (Weaver, 1955; Bar-Hillel, 1960). Moreover, RNN-based neural machine translation (NMT) has been shown to benefit from the injection of semantic structure (Song et al., 2019; Marcheggiani et al., 2018). Despite these gains, to our knowledge, there have been no attempts to incorporate semantic structure into NMT Transformers (Vaswani et al., 2017). We address this gap, focusing on the main events in the text, as represented by UCCA (Universal Cognitive Conceptual Annotation; Abend and Rappoport, 2013), namely *scenes*.

UCCA is a semantic framework originating from typological and cognitive-linguistic theories (Dixon, 2009, 2010, 2012). Its principal goal is to represent some of the main elements of the semantic structure of the sentence while disregarding its syntax. Formally, a UCCA representation of a text is a directed acyclic graph where leaves correspond to the words of the sentence and nodes correspond to semantic units. The edges are labeled by the role of their endpoint in the relation corresponding to their starting point (see Fig. 1). One of the motivations for using UCCA is its capability to separate the sentence into *"Scenes"*, which are analogous to events (see Fig. 1). Every such Scene consists of one main relation, which can be either a Process (i.e., an action), denoted by P, or a State (i.e., continuous state), denoted by S. Scenes also contain at least one Participant (i.e., entity), denoted by A. For example, the sentence in Fig. 1 comprises two scenes: the first one has the Process "saw" and two Participants – "I" and "the dog"; the second one has the Process "barked" and a single Participant – "dog".

So far, to the best of our knowledge, the only structure-aware work that integrated linguistic knowledge and graph structures into Transformers used syntactic structures (Strubell et al., 2018; Bugliarello and Okazaki, 2020; Akoury et al., 2019; Sundararaman et al., 2019; Choshen and Abend, 2021). The presented method builds on the method proposed by Bugliarello and Okazaki (2020), which utilized a Universal Dependencies graph (UD; Nivre et al., 2016) of the source sentence to focus the encoder's attention on each token's parent, namely the token's immediate ancestor in the UD graph. Similarly, we use the UCCA graph of the source sentence to generate a scene-aware mask for the self-attention heads of the encoder. We call this method *SASA* (see §2.1).

We test our model (§2) on translating English into four languages. Two that are more syntactically similar to English –(Nikolaev et al., 2020; Dryer and Haspelmath, 2013) German (En-De), Russian (En-Ru), and two that are much less so – Turkish (En-Tr) and Finnish (En-Fi). We find consistent improvements across multiple test sets for all four cases. In addition, we create a syntactic variant of our semantic model for better comparability. We observe that on average, our semanti-

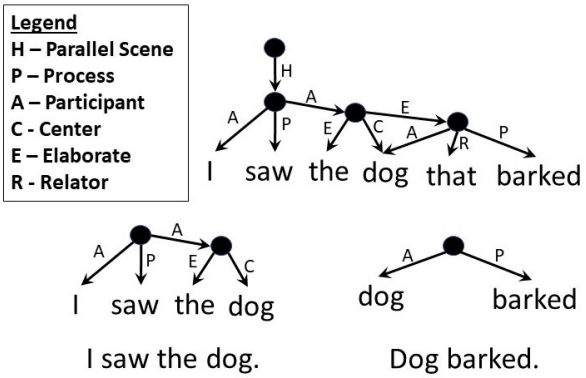

Figure 1: Example of UCCA-parsing of the sentence "I saw the dog that barked" and its separation into two scenes.

cally aware model outperforms the syntactic models. Moreover, for the two languages less similar to English (En-Tr and En-Fi), combining both the semantic and the syntactic data results in a further gain. While improvements are often small, at times the combined version outperforms SASA and UDISCAL (our syntactic variant, see §3) by 0.52 and 0.69 BLEU points (or 0.46 and 0.43 chrF), respectively.

We also propose a novel method for introducing the source graph information during the decoding phase, namely through the cross-attention layer in the decoder (see §2.2). We find that it improves over the baseline and syntactic models, although SASA is generally better. Interestingly, for En-Fi, this model also outperforms SASA, suggesting that some language pairs may benefit more from semantic injection into the decoder.

Overall, through a series of experiments (see §4), we show the potential of semantics as an aid for NMT. We experiment with a large set of variants of our method, to see where and in what incorporation method they best help. Finally, we show that semantic models outperform UD baselines and can be complementary to them in distant languages, showing improvement when combined.

2 Models

Transformers have been shown to struggle when translating some types of long-distance dependencies (Choshen and Abend, 2019; Bisazza et al., 2021a) and when facing atypical word order (Bisazza et al., 2021b). Sulem et al. (2018a) proposed UCCA based preprocessing at inference time, splitting sentences into different scenes. They hypothesized that models need to decompose the input into scenes implicitly, and provide them with such a decomposition, as well as with the original sentence. They show that this may facilitate machine translation (Sulem et al., 2020) and sentence simplification (Sulem et al., 2018b) in some cases.

Motivated by these advances, we integrate UCCA to split the source into scenes. However, unlike Sulem et al., we do not alter the sentence length in pre-processing, as this method allows less flexibility in the way information is passed, and also reimplementing it yielded inferior results (see §A.5). Instead, we investigate ways to integrate the split into the attention architecture.

We follow previous work (Bugliarello and Okazaki, 2020) in the way we incorporate our semantic information. In their paper, Bugliarello and Okazaki (2020) introduced syntax in the form of a parent-aware mask, which was applied before the softmax layer in the encoder's self-attention. We mask in a similar method to introduce semantics. However, *parent* in the UCCA framework is an elusive concept. Hence, we use a different way to express the semantic information in our mask, i.e., we make it *scene-aware*, rather than *parent-aware*.

Following Sulem et al. (2018b), we divide the source sentence into scenes, using the sentence's UCCA parse. We then define our Scene-Aware mask

$$M[i,j] = \begin{cases} 1, & \text{if i,j in the same scene} \\ 0, & \text{otherwise} \end{cases}$$

Intuitively, an attention head masked this way is allowed to attend to other tokens, as long as they share a scene with the current one.[1]

Our base model is the Transformer (Vaswani et al., 2017), which we enhance by making the attention layers more scene-aware. We force one[2] of the heads to attend to words in the same scene which we assume are more likely to be related than words from different scenes. As we replace regular self-attention heads with our scene-aware ones, we maintain the same number of heads and layers as in the baseline.

2.1 Scene-Aware Self-Attention (SASA)

Figure 2 presents the model's architecture. For a source sentence of length L, we obtain the

[1] In case a token belongs to more than one scene, as is the case with the word "dog" in Fig. 1, we allow it to attend to tokens of all the scenes it belongs to.

[2] Initial trials with more than one head did not show further benefit for UCCA based models.

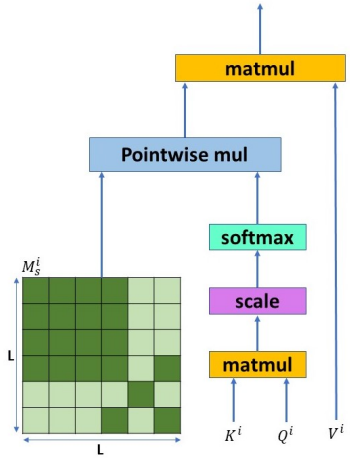

Figure 2: Scene-aware self-attention head for the input sentence "I saw the dog that barked", consisting of two scenes: "I saw the dog" and "dog barked".

keys, queries, and values matrices denoted by $K^i, Q^i, V^i \in \mathbb{R}^{L \times d}$, respectively. We then multiply K^i with Q^i, scale the result and pass it through a softmax layer. The difference between our method and a vanilla Transformer (Vaswani et al., 2017) is that the vanilla model multiplies the output of the softmax by V^i. We, however, first mask it with our pre-generated scene-aware mask $M_S^i \in \mathbb{R}^{L \times L}$, using a pointwise multiplication, and only then do we multiply the result with V^i.

2.2 Scene-Aware Cross-Attention (SACrA)

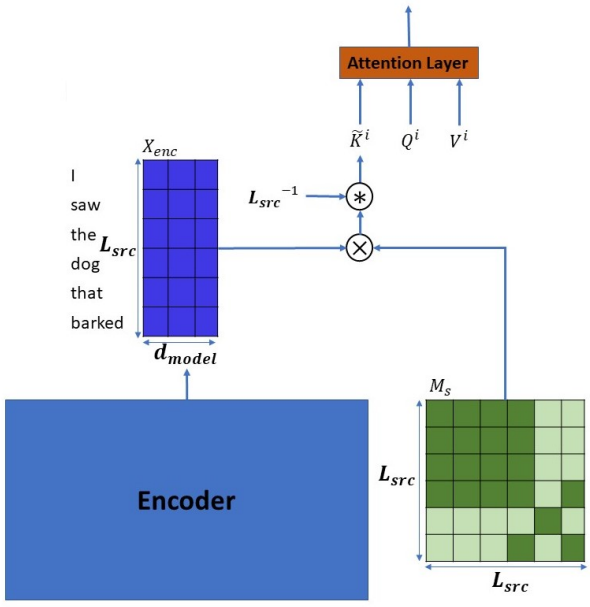

Figure 3: Scene-aware cross-attention head for the source sentence "I saw the dog that barked."

Next, we design a model in which we integrate information about the scene structure through the cross-attention layer in the decoder (see Fig. 3). Thus, instead of affecting the overall encoding of the source, we bring forward the splits to aid in selecting the next token.

Formally, for a source sentence of length L_{src} and target sentence of length L_{trg}, we compute for each head the queries and values matrices, denoted by $Q^i \in \mathbb{R}^{L_{trg} \times d_{model}}$ and $V^i \in \mathbb{R}^{L_{src} \times d}$, accordingly. Regarding key values, denoted by $\tilde{K}^i \in \mathbb{R}^{L_{src} \times L_{trg}}$: we multiply the encoder's output $X_{enc} \in \mathbb{R}^{L_{src} \times d_{model}}$ with our pre-generated mask $M_S \in \{0,1\}^{L_{src} \times L_{src}}$, and then scale it by multiplying the result with $\frac{1}{L_{src}}$. Finally, we pass V^i, Q^i and $\tilde{K}^i$ through a regular attention layer, as with the standard Transformer architecture.

Scene-Aware Key Matrix. The rationale behind the way we compute our scene-aware keys matrix lies in the role of the keys matrix in an attention layer. In the cross-attention layer, the queries come from the decoder. Source-side contextual information is encoded in the keys, which come from the encoder. Therefore, when we assign the same scene masks to all the words that are included in the same set of scenes, the key values for these words will be the same, and they will thus be treated similarly by the query. As a result, the query will give the same weight to source tokens that share the same set of scenes. Therefore, a complete scene (or a few scenes), rather than specific tokens (as with the vanilla Transformer), will influence what the next generated token will be, which will in turn yield a more scene-aware decoding process.

3 Experimental Setting

Data Preparation. First, we unescaped HTML characters and tokenized all our parallel corpora (Koehn et al., 2007). Next, we removed empty sentences, sentences longer than 100 tokens (either on the source or the target side), sentences with a source-target ratio larger than 1.5, sentences that do not match the corpus's language as determined by langid Lui and Baldwin, 2012, and sentences that *fast align* (Dyer et al., 2013) considers unlikely to align (minimum alignment score of -180). Then, for languages with capitalization, we trained true-casing models on the train set (Koehn et al., 2007) and applied them to all inputs to the network. Finally, we trained a BPE model (Sennrich et al., 2016), jointly for language pairs with a similar writing system (e.g., Latin, Cyrillic, etc.)

and separately otherwise, and then applied them accordingly.

We trained our model on the full WMT16 dataset for the English→German (En-De) task, using the WMT *newstest2013* as development set. We also trained our models on a train set consisting of Yandex Corpus, News Commentary v15, and Wikititles v2 for the English→Russian (En-Ru) task. In addition, we trained our models on the full WMT19 dataset (excluding ParaCrawl, in order to avoid noisiness in the data) for the English→Finnish (En-Fi). Finally, we trained on the full WMT18 dataset for the English→Turkish (En-Tr) task. For the test sets, we used all the newstests available for every language pair since 2012, excluding the one designated for development.

Models. Hyperparameters shared by all models are described in §3. We tune the number of heads that we apply the mask to ($\#heads$) and the layers of the encoder we apply SASA to ($layer$), using the En-De development set. We start with tuning the layers for SASA, which we find is $layer = 4$, and then we tune the $\#heads$ (while fixing $layer = 4$), and get $\#head = 1$. We also use the En-De development set to tune the $\#heads$ and the layers of the SACrA model in a similar fashion, namely first the layers and then the $\#heads$ (with the tuned layers fixed). We find the best hyperparameters are $\#heads = 1$ and $layers = 2\&3$. For both models, we apply the tuned hyperparameters to all other language pairs. Interestingly, while it is common practice to change all the layers of the model, we find it suboptimal. Moreover, the fact that semantic information is more beneficial in higher layers, in contrast to the syntactic information that is most helpful when introduced in lower layers (see §3) may suggest that semantics is relevant for more complex generalization, which is reminiscent of findings by previous work (Tenney et al., 2019a; Belinkov, 2018; Tenney et al., 2019b; Peters et al., 2018; Blevins et al., 2018; Slobodkin et al., 2021).

UCCA parses are extracted using a pretrained BERT-based TUPA model, that was trained on sentences in English, German and French (Hershcovich et al., 2017).

Binary Mask. For the SASA model, we experiment with two types of masks: a binary mask, as described in §2, and scaled masks, i.e.,

$$M_C[i,j] = \begin{cases} 1, & \text{if i,j in the same scene} \\ C, & \text{otherwise} \end{cases} \quad (1)$$

where $C \in (0,1)$. By doing so, we allow some out-of-scene information to pass through, while still emphasizing the in-scene information (by keeping the value of M for same-scene tokens at 1). In order to tune C, we performed a small grid search over $C \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.5\}$.

Additionally, similarly to Bugliarello and Okazaki (2020), we test a normally-distributed mask, according to the following equation:

$$M_{i,j} = f_{norm}(x = C \cdot dist(i,j)) \quad (2)$$

where f_{norm} is the density function of the normal distribution:

$$f_{norm}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \quad (3)$$

We define a scene-graph where nodes are scenes and edges are drawn between scenes with overlapping words. $dist(i,j)$ is the shortest distance between tokens i and j. $\sigma = \frac{1}{\sqrt{2\pi}}$, to ensure the value of M is 1 for words that share a scene ($dist(i,j)$=0), and C is a hyperparameter, which is determined through a grid search over $C \in \{0.1, 0.2, 0.5, \sqrt{0.5}\}$. For each of those two scaled versions of the mask, we choose the mask which has the best performance and compare it to the binary mask (see 1). We find that neither outperforms the binary mask. Therefore, we report the rest of our experiments with the binary mask.

Baselines. We compared our model to a few other models:

- **Transformer.** Standard Transformer-based NMT model, using the standard hyperparameters, as described in §3.
- **PASCAL.** Following Bugliarello and Okazaki (2020), we generate a syntactic mask for the self-attention layer in the encoder. We extract a UD-graph (Nivre et al., 2016) with udpipe (Straka and Straková, 2017). The value of the entries of the masks equal (see equation 3):

$$M_{p_t,j} = f_{norm}(x = (j - p_t)) \quad (4)$$

models	2012	2013	2014	2015	2016	2017	2018	2019	2020	2020B
Transformer	17.60	20.49	20.55	22.17	25.46	19.70	28.01	26.84	17.71	16.94
+ binary mask (#h=1, l= 4)	**17.64**	20.37	**20.84**	**22.48**	25.32	19.76	**28.36**	26.80	17.74	16.98
+ scaled mask (#h=2, l=4, C=0.1)	17.41	20.21	20.53	22.43	24.95	**19.81**	28.25	**27.21**	**18.03**	**17.01**
+ normally distributed mask (#h=2, l=4, C=$\sqrt{0.5}$)	17.39	**20.52**	20.57	22.24	25.44	19.63	28.35	26.6	17.14	16.77

Table 1: BLEU scores for the top versions of our binary mask, scaled mask, and normally-distributed mask methods across all the WMT En-De newstests. Each column contains the BLEU scores over the WMT newstest corresponding to the year the column is labeled with (e.g., the scores under column *2015* are for En-De newstest2015). For newstest2020, there was more than one version on WMT, each translated by a different person. Both versions were included, with the second version denoted with a "B". The best score for each test set is boldfaced, unless none is better than the baseline Transformer.

with $\sigma = 1$ and p_t being the middle position of the t-th token's parent in the UD graph of the sentence.

We use the same general hyperparameters as in the Transformer baseline. In addition, following the tuning of Bugliarello and Okazaki (2020), we apply the PASCAL mask to five heads of the first attention layer of the encoder, but unlike the original paper, we apply it after the layer's softmax, as it yields better results and also resembles our model's course of action.

- **UDISCAL.** In an attempt to improve the PASCAL model, we generate a mask that instead of only being sensitive to the dependency parent, is sensitive to all the UD relations in the sentences. We denote it UD-Distance-Scaled mask (UDISCAL). Namely, in order to compute the mask, we use a similar equation to that of PASCAL, with a minor alteration:

$$M_{i,j} = f_{norm}(x = dist(i,j)) \qquad (5)$$

Where $\sigma = 1$, and $dist(i,j)$ is defined to be the distance between the token i and the token j in the UD graph of the sentence while treating the graph as undirectional. As with the PASCAL layer, we apply the UD-scaled mask after the softmax layer. But, unlike the PASCAL head, we tuned the architecture's hyperparameters to be just one head of the first layer, after performing a small grid search, namely testing with all layers $l \in [1,4]$, and then with $\#head \in [1,5]$.

Training Details. All our models are based on the standard Transformer-based NMT model (Vaswani et al., 2017), with 4000 warmup steps. In addition, we use an internal token representation of size 256, per-token cross-entropy loss function, label smoothing with $\epsilon_{l_s} = 0.1$ (Szegedy et al., 2016), Adam optimizer, Adam coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and Adam $\epsilon = e^{-1}$. Furthermore, we incorporate 4 layers in the encoder and 4 in the decoder, and we employ a beam-search during inference, with beam size 4 and normalization coefficient $\alpha = 0.6$. In addition, we use a batch size of 128 sentences for the training. We use *chrF++.py* with 1 word and beta of 3 to obtain chrF+ (Popovic, 2017) score as in WMT19 (Ma et al., 2019) and detokenized BLEU (Papineni et al., 2002) as implemented in Moses. We use the Nematus toolkit (Sennrich et al., 2017), and we train all our models on 4 NVIDIA GPUs for 150K steps. The average training time for the vanilla Transformer is 21.8 hours, and the average training time for the SASA model is 26.5 hours.

4 Experiments

We hypothesize that NMT models may benefit from the introduction of semantic structure, and present a set of experiments that support this hypothesis using the above-presented methods.

4.1 Scene-Aware Self-Attention

We find that on average, SASA outperforms the Transformer for all four language pairs (see 3), at times having gains larger than 1 BLEU point. Moreover, we assess the consistency of SASA's gains, using the sign-test, and get a p-value smaller than

	Source sentences and Translations	Literal Translations into English
SRC	I promised a show ?	
BASE	Я обещал показать?	I promised to show?
SASA	Я обещал шоу?	I promised a show?
SRC	students said they looked forward to his class .	
BASE	Студенты сказали, что они смотрят на свой класс.	Students said, that they look at one's classroom.
SASA	Студенты сказали, что они с нетерпением ждали своего класса.	Students said, that they impatiently waited one's classroom.
SRC	I remember those kids I used to play with in the yard who never got out .	
BASE	Я помню тех детей, которые я играл с двором, который никогда не выходил.	I remember those kids, that I played with yard, that never got out ("that" and "got out" refer to yard).
SASA	Я помню тех детей, с которыми я играл на дворе, которые никогда не вышли.	I remember those kids, with which I played in yard, that never got out ("that" and "got out" refer to kids).

Table 2: Examples of correct translations generated by SASA, compared to the baseline Transformer.

0.01, thus exhibiting a statistically significant improvement (see §A.4). We see a similar trend when evaluating the performance using the chrF metric (see §A.2), which further highlights our model's consistent gains.

We also evaluate our model's performance on sentences with long dependencies (see A.3), which were found to pose a challenge for Transformers (Choshen and Abend, 2019). We assume that such cases could benefit greatly from the semantic introduction. In contrast to our hypothesis, we find the gain to be only slightly larger than in the general case, which leads us to conclude the improvements we see do not specifically originate from the syntactic challenge. Nevertheless, we still observe a consistent improvement, with gains of up to 1.41 BLEU points, which further underscores our model's superiority over the baseline model.

Qualitative Analysis. Table 2 presents a few examples in which the baseline Transformer errs, whereas our model translates correctly. In the first example, the Transformer translates the word "show" as a verb, i.e. *to show*, rather than as a noun. In the second example, the baseline model makes two errors: it misinterprets the word "look forward to" as "look at", and it also translates it as a present-tense verb rather than past-tense. The third example is particularly interesting, as it highlights our model's strength. In this example, the Transformer makes two mistakes: first, it translates the part "play with (someone) in the yard" as "play with the yard". Next, it attributes the descriptive clause "which never got out" to the yard, rather than the children. It seems then that introducing information about the *scene* structure into the model facilitates the translation, since it both groups the word "kids" with the phrase "I used to play with in the yard", and it also separates "never got out" from the word "yard". Instead, it clusters the latter with "kids", thus highlighting the relations between words in the sentence. In general, all these examples are cases where the network succeeds in disambiguating a word in its context.

4.2 Comparison to Syntactic Masks

Next, we wish to compare our model to other baselines. Given that this is the first work to incorporate semantic information into the Transformer-based NMT model, we compare our work to syntactically-infused models (as described in §3): one is the PASCAL model (Bugliarello and Okazaki, 2020), and the other is our adaptation of PASCAL, the UD-Distance-Scaled (UDISCAL) model, which resembles better our SASA mask. We find (Table 3) that on average, SASA outperforms both PASCAL and UDISCAL. We also compare SASA with each of the syntactic models, finding that it is significantly (sign-test $p < 0.01$; see §A.4) better. This suggests that semantics might be more beneficial for Transformers than syntax.

En-De

models	2012	2014	2015	2016	2017	2018	2019	2020	2020B	average
Transformer	17.6	20.55	22.17	**25.46**	19.7	28.01	26.84	17.71	16.94	21.66
PASCAL	17.34	20.59	**22.62**	25.1	19.92	28.09	26.61	17.5	16.81	21.62
UDISCAL	17.42	20.86	22.53	25.23	**19.95**	27.87	26.8	17.06	16.39	21.57
SASA	**17.64**$^{\uparrow}$	20.84	22.48	25.32	19.76	**28.36**$^{\uparrow}$	26.8	**17.74**$^{\uparrow}$	**16.98**$^{\uparrow}$	**21.77**$^{\uparrow}$
SASA + UDISCAL	17.51	20.42	22.1	24.9	19.72	28.35	**27.14***	17.59	16.68	21.60
SACrA	17.11	20.9$^{\uparrow}$	22.59	24.64	19.79	27.88	26.28	16.8	16.25	21.36
SACrA + UDISCAL	17.07	**21.09***	22.26	24.85	19.56	28.1*	26.49	16.66	15.93	21.33

En-Ru

models	2012	2013	2014	2015	2016	2017	2018	2019	2020	2020B	average
Transformer	24.32	18.11	25.35	21.1	19.77	22.34	19	20.14	15.64	22.33	20.81
PASCAL	23.78	18.37	24.87	20.97	19.81	21.83	18.81	19.93	15.42	21.48	20.53
UDISCAL	23.88	18.31	25.23	20.82	**20.31**	22.15	19.27	20.32	15.7	22.19	20.82
SASA	24.17	**18.43**$^{\uparrow}$	**25.53**$^{\uparrow}$	**21.59**$^{\uparrow}$	20.11	**22.69**$^{\uparrow}$	**19.53**$^{\uparrow}$	20.2	15.76$^{\uparrow}$	**23.36**$^{\uparrow}$	**21.14**$^{\uparrow}$
SASA + UDISCAL	**24.36***	18.29	25.43	21.01	19.79	22.49	19.25	**20.4***	**15.97***	22.42	20.94
SACrA	24.12	18.24	25.43$^{\uparrow}$	21	20.07	22.49$^{\uparrow}$	19.3$^{\uparrow}$	20.18	15.79$^{\uparrow}$	22.15	20.88$^{\uparrow}$
SACrA + UDISCAL	23.54	17.99	24.91	20.62	19.67	21.55	18.63	19.89	15.64	20.79	20.32

En-Fi

models	2015	2016	2016B	2017	2017B	2018	2019	average
Transformer	11.22	12.76	10.2	13.35	11.37	9.32	12.21	11.49
PASCAL	11.2	12.67	10.13	13.54	11.24	9.62	12.23	11.52
UDISCAL	10.87	12.78	10.23	13.51	11.43	9.2	11.99	11.43
SASA	11.37$^{\uparrow}$	**12.88**$^{\uparrow}$	**10.52**$^{\uparrow}$	13.74$^{\uparrow}$	11.5$^{\uparrow}$	9.56	12.12	11.67$^{\uparrow}$
SASA + UDISCAL	**11.56***	12.8	10.28	**13.91***	**11.52***	**9.75***	**12.64***	**11.78***
SACrA	11.48$^{\uparrow}$	12.86$^{\uparrow}$	10.41$^{\uparrow}$	13.66$^{\uparrow}$	11.49$^{\uparrow}$	9.62	12.51$^{\uparrow}$	11.72$^{\uparrow}$
SACrA + UDISCAL	11.06	12.6	10.13	13.43	11.26	9.23	12.05	11.39

En-Tr

models	2016	2017	2018	average
Transformer	8.43	8.55	8.1	8.36
PASCAL	8.5	8.76	7.98	8.41
UDISCAL	8.33	8.66	8.03	8.34
SASA	8.59$^{\uparrow}$	8.86$^{\uparrow}$	8.16$^{\uparrow}$	8.54$^{\uparrow}$
SASA + UDISCAL	**8.64***	**8.87***	**8.2***	**8.57***
SACrA	8.64$^{\uparrow}$	8.81$^{\uparrow}$	7.96	8.47$^{\uparrow}$
SACrA + UDISCAL	8.23	8.54	7.95	8.24

Table 3: BLEU scores for the baseline Transformer model, previous work that used syntactically infused models – PASCAL and UDISCAL, our SASA and SACrA models, and models incorporating UDISCAL with SASA or SACrA, across all WMT's newstests. For every language pair, each column contains the BLEU scores over the WMT newstest corresponding to the year the column is labeled with (e.g., for En-Ru, the scores under column *2015* are for En-Ru newstest2015). For some newstests, there was more than one version on WMT, each translated by a different person. For those test sets, we included both versions, denoting the second one with a "B". In addition, for every language pair, the right-most column represents the average BLEU scores over all the pair's reported newstests. For every test set (and for the average score), the best score is boldfaced. For each of the semantic models (i.e., SASA and SACrA), improvements over all the baselines (syntactic and Transformer) are marked with an arrow facing upwards. For models with both syntactic and semantic masks, improvements over each mask individually are marked with an asterisk.

4.3 Combining Syntax and Semantics

Naturally, our next question is whether combining both semantic and syntactic heads will further improve the model's performance. Therefore, we test the combination of SASA with either PASCAL or UDISCAL, retaining the hyperparameters used for the separate models. We find that combining with UDISCAL outperforms the former, and so we continue with it. Interestingly, En-De and En-Ru hardly benefit from the combination compared just to the SASA model. We hypothesize that this might be due to the fact that the syntax of each language pair is already quite similar, and therefore the model mainly relies on it to separate the sentence that UCCA gives it as well. On the other hand, En-Fi and En-Tr do benefit from the combination, both on average and in most of the test sets. Evaluating the performance using the chrF metric (see A.2) yields a similar behavior, which further confirms its validity. It leads us to hypothesize that language pairs that are more typologically distant from one another can benefit more from both semantics and syntax; we defer a more complete discussion of this point to future work. In order to confirm that the combined version persistently outperforms each of the separate versions for typologically distant languages, we compare each of the pairs using the sign-test (only on the test sets of En-Fi and En-Tr). We get a p-value of 0.02 for the comparison with SASA and 0.0008 for the comparison with UDISCAL. This suggests that for these language pairs, there is indeed a significant benefit, albeit small, from the infusion of both semantics and syntax.

4.4 Scene-Aware Cross-Attention

Following the analysis on the scene-aware *self*-attention, we wish to examine whether Transformers could also benefit from injecting source-side semantics into the decoder. For that, we develop the Scene-Aware Cross-Attention (SACrA) model, as described in §2.2. Table 3 presents the results of SACrA, compared to the Transformer baseline and SASA. We find that in general SASA outperforms SACrA, suggesting that semantics is more beneficial during encoding. With that said, for three out of the four language pairs, SACrA does yield gains over the Transformer, albeit small, and for one language pair (En-Fi) it even outperforms SASA on average. Moreover, comparing SACrA to the Transformer using the sign-test (see §A.4) shows significant improvement ($p = 0.047$).

Surprisingly, unlike its self-attention counterpart, combining the SACrA model with UDISCAL does not seem to be beneficial at all, and in most cases is even outperformed by the baseline Transformer. We hypothesize that this occurs because appointing too many heads for our linguistic injection is inefficient when those heads cannot interact with each other directly, as the information from the UDISCAL head reaches the SACrA head only after the encoding is done. One possible direction for future work would be to find ways to syntactically enrich the decoder, and then to combine it with our SACrA model.

5 Conclusion

In this work, we suggest two novel methods for injecting semantic information into an NMT Transformer model – one through the encoder (i.e. SASA) and one through the decoder (i.e. SACrA). The strength of both methods is that they both do not introduce more parameters to the model, and only rely on UCCA-parses of the source sentences, which are generated in advance using an off-the-shelf parser, and thus do not increase the complexity of the model. We compare our methods to previously developed methods of syntax injection, and to our adaptation to these methods, and find that semantic information tends to be significantly more beneficial than syntactic information, mostly when injected into the encoder (SASA), but at times also during decoding (SACrA). Moreover, we find that for distinct languages, adding both syntax and semantics further improves the performance of the translation models. Future work will further investigate the benefits of semantic structure in Transformers, alone and in unison with syntactic structure.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proc. of ACL*, pages 228–238.

Nader Akoury, Kalpesh Krishna, and Mohit Iyyer. 2019. Syntactically supervised transformers for faster neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1269–1281, Florence, Italy. Association for Computational Linguistics.

Y. Bar-Hillel. 1960. The present status of automatic translation of languages. *Adv. Comput.*, 1:91–163.

Yonatan Belinkov. 2018. On internal language representations in deep learning: an analysis of machine translation and speech recognition.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021a. On the difficulty of translating free-order case-marking languages. *CoRR*, abs/2107.06055.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021b. On the difficulty of translating free-order case-marking languages. *arXiv preprint arXiv:2107.06055*.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

Emanuele Bugliarello and Naoaki Okazaki. 2020. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2019. Automatically extracting challenge sets for non-local phenomena in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 291–303, Hong Kong, China. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2021. Transition based graph decoder for neural machine translation. *arXiv preprint arXiv:2101.12640*.

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Lang. Resour. Evaluation*, 49(2):375–395.

R.M.W. Dixon. 2009. *Basic Linguistic Theory Volume 1: Methodology*. Basic Linguistic Theory. OUP Oxford.

R.M.W. Dixon. 2010. *Basic Linguistic Theory Volume 2: Grammatical Topics*. Basic Linguistic Theory. OUP Oxford.

R.M.W. Dixon. 2012. *Basic Linguistic Theory Volume 3: Further Grammatical Topics*. Basic Linguistic Theory. OUP Oxford.

Matthew S Dryer and Martin Haspelmath. 2013. The world atlas of language structures online.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. In *Proc. of ACL*, pages 1127–1138.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proc. of LREC*.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Maja Popovic. 2017. chrf++: words helping character n-grams. In *WMT*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Aviv Slobodkin, Leshem Choshen, and Omri Abend. 2021. Mediators in determining what processing BERT performs first. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–93, Online. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2020. Semantic structural decomposition for neural machine translation. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 50–57, Barcelona, Spain (Online). Association for Computational Linguistics.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2019. Syntax-infused transformer and bert models for machine translation and natural language understanding.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Warren Weaver. 1955. Translation. *Machine translation of languages*, 14:15–23.

Krzysztof Wołk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, 18:126–132. International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014, Warsaw, Poland.

A Appendix

A.1 Layer Hyperparameter-tuning for SASA

In order to optimize the contribution of the SASA model, we tuned the hyperparameter of the best layers in the encoder to incorporate our model, using the En-De newstest2013 as our development set. Table 4 presents the results.

Layers	Bleu
1	20.3
2	20.33
3	20.1
4	20.37
1,2	20.2
2,3	20.17
3,4	20.3

Table 4: Validation Bleu as a function of layers incorporating SASA (for En-De).

A.2 ChrF Results

In order to reaffirm our results, we also evaluate the performance of all the models using the chrF metric (see 7). Indeed, all the different behaviors and trends we observed when evaluating using the Bleu metric (see §4) seem to be preserved when under the chrF metric. This further validates our results.

A.3 Challenge Sets

In addition to testing on the full newstests sets, we also experiment with sentences characterized by long dependencies, which were shown to present a challenge for Transformers (Choshen and Abend, 2019). In order to acquire those challenge sets, we use the methodology described by Choshen and Abend (2019), which we apply on each of the newstest sets. In addition, for the En-Tr task, which has a limited number of newstests, we generate additional challenge sets, extracted from corpora downloaded from the Opus Corpus engine (Tiedemann, 2012): the Wikipedia parallel corpus (Wołk and Marasek, 2014), the Mozilla and EUbookshop parallel corpora (Tiedemann, 2012), and the bible parallel corpus (Christodoulopoulos and Steedman, 2015). We observe (see 8) a similar trend to the general case, which reaffirms our results. In fact, there seem to be bigger gains over the Transformer, albeit not drastically, compared to the general case.

BASELINE \ BETTER	PASCAL	UDISCAL	SASA	SASA + UDISCAL	SACrA	SACrA + UDISCAL
Transformer	>0.5	>0.5	<0.01	<0.01	0.047	>0.5
PASCAL		0.17	<0.01	<0.01	0.06	>0.5
UDISCAL			<0.01	<0.01	0.06	>0.5
SASA				0.17	>0.5	>0.5
SASA + UDISCAL					>0.5	>0.5
SACrA						>0.5

Table 5: We perform a significance test over all test sets across all languages for every cell, where the null hypothesis is $H_0 : Bleu(model_{row}) \geq Bleu(model_{column})$

A.4 Sign-Test

In order to assess the consistency of the improvements of our models, we perform the Sign-Test on every two models (see 5). Evidently, SASA persistently outperforms the Transformer baseline and the syntactic models, as does the combined model of SASA and UDISCAL.

A.5 SemSplit

Following Sulem et al. (2020), we implement the SemSplit pipeline. First, we train a Transformer-based Neural Machine Translation model. Then, during inference time, we use the Direct Semantic Splitting algorithm (DSS; Sulem et al., 2018b) to split the sentences, and then translate each separated sentence separately. Finally, we concatenate the translation, using a period (".") as a delimiter. Table 6 presents the results, using the Bleu and chrF metrics. We find that the architecture does not have gains over the baseline Transformer. These results can be accounted for by the fact that in their work, Sulem et al. (2020) assessed the pipeline's performance using Human Evaluation and manual analysis, rather than the Bleu and chrF metrics, which punish for sentence separation in translation. In addition, they tested their pipeline in a pseudo-low resource scenario, and not in normal NMT settings.

En-De

Metric	Models	2012	2014	2015	2016	2017	2018	2019	2020	2020B	average
Bleu	Transformer	17.6	20.55	22.17	25.46	19.7	28.01	26.84	17.71	16.94	21.66
	SemSplit	12.16	14.25	14.46	17.53	13.18	19.39	18.46	15.12	14.93	15.50
chrF	Transformer	47.37	51.85	52.52	55.06	50.87	57.81	55.48	45.19	44.18	51.15
	SemSplit	43.42	47.19	47.05	49.86	45.87	51.50	50.24	47.71	46.93	47.75

En-Ru

Metric	Models	2012	2013	2014	2015	2016	2017	2018	2019	2020	2020B	average
Bleu	Transformer	24.32	18.11	25.35	21.1	19.77	22.34	19	20.14	15.64	22.33	20.81
	SemSplit	15.29	10.9	16.43	13.28	12.79	14.61	11.95	12.56	9.92	15.25	13.30
chrF	Transformer	51.39	45.69	53.31	50.16	48.10	50.54	48.01	45.78	42.51	53.07	48.86
	SemSplit	46.10	40.50	47.66	44.58	43.16	45.34	43.38	40.97	38.93	47.84	43.85

En-Fi

Metric	Models	2015	2016	2016B	2017	2017B	2018	2019	average
Bleu	Transformer	11.22	12.76	10.2	13.35	11.37	9.32	12.21	11.49
	SemSplit	6.97	7.72	6.55	8.75	7.54	6.18	7.73	7.35
chrF	Transformer	43.79	45.48	43.43	46.39	43.96	42.06	43.10	44.03
	SemSplit	40.18	41.42	39.94	42.18	40.20	38.76	40.12	40.40

En-Tr

Metric	Models	2016	2017	2018	average
Bleu	Transformer	8.43	8.55	8.1	8.36
	SemSplit	6.15	6.07	5.37	5.86
chrF	Transformer	40.24	40.37	39.75	40.12
	SemSplit	39.04	39.00	38.85	38.97

Table 6: Bleu and ChrF scores of the baseline Transformer and the SemSplit model.

En-De

models	2012	2014	2015	2016	2017	2018	2019	2020	2020B	average
Transformer	47.37	51.85	52.52	**55.06**	50.87	57.81	55.48	**45.19**	**44.18**	51.15
PASCAL	47.27	51.87	**52.82**	54.73	50.83	57.65	55.28	44.80	43.78	51.00
UDISCAL	47.26	51.95	52.45	54.99	50.78	57.40	55.30	44.48	43.43	50.89
SASA	**47.48**$^{\uparrow}$	**52.03**$^{\uparrow}$	52.74	54.99	**51.23**$^{\uparrow}$	**57.88**$^{\uparrow}$	**55.69**$^{\uparrow}$	45.03	43.99	**51.23**$^{\uparrow}$
SASA + UDISCAL	47.42	51.94	52.50	55.00*	50.86	57.74	55.62	44.72	43.62	51.05
SACrA	47.02	51.66	52.48	54.49	50.55	57.16	55.05	44.08	43.15	50.63
SACrA + UDISCAL	46.71	51.63	52.18	54.37	50.22	57.20	54.96	43.42	42.40	50.34

En-Ru

models	2012	2013	2014	2015	2016	2017	2018	2019	2020	2020B	average
Transformer	51.39	45.69	53.31	50.16	48.10	50.54	48.01	45.78	42.51	53.07	48.86
PASCAL	51.03	45.66	53.04	49.87	48.05	50.32	47.98	45.86	42.35	52.42	48.66
UDISCAL	51.26	45.73	53.45	50.01	48.57	50.50	48.27	46.03	42.60	52.89	48.93
SASA	51.34	**45.81**$^{\uparrow}$	53.49$^{\uparrow}$	**50.32**$^{\uparrow}$	**48.60**$^{\uparrow}$	50.67$^{\uparrow}$	**48.45**$^{\uparrow}$	45.81	42.76$^{\uparrow}$	**53.62**$^{\uparrow}$	**49.09**$^{\uparrow}$
SASA + UDISCAL	**51.43***	45.67	**53.56***	50.03	48.29	50.67	48.25	**46.08***	**42.81***	53.14	48.99
SACrA	51.28	45.57	53.50$^{\uparrow}$	49.81	48.42	**50.82**$^{\uparrow}$	48.28$^{\uparrow}$	45.92	42.68$^{\uparrow}$	52.76	48.90
SACrA + UDISCAL	50.58	45.31	52.90	49.40	47.77	50.03	47.49	45.26	42.33	51.93	48.30

En-Fi

models	2015	2016	2016B	2017	2017B	2018	2019	average
Transformer	43.79	45.48	43.43	46.39	43.96	42.06	43.10	44.03
PASCAL	**43.91**	44.93	42.99	46.02	43.57	41.88	42.60	43.70
UDISCAL	43.42	45.37	43.42	46.51	44.07	42.03	43.03	43.98
SASA	43.76	45.33	43.38	46.40	43.89	42.10$^{\uparrow}$	43.02	43.98
SASA + UDISCAL	43.77*	45.20	43.17	**46.74***	44.15*	**42.34***	43.08*	44.07*
SACrA	43.88	45.20	43.15	46.62$^{\uparrow}$	44.02$^{\uparrow}$	42.25$^{\uparrow}$	43.23$^{\uparrow}$	44.05$^{\uparrow}$
SACrA + UDISCAL	43.80	**45.53***	**43.52***	46.71*	**44.19***	42.16	**43.28***	**44.17***

En-Tr

models	2016	2017	2018	average
Transformer	40.24	40.37	39.75	40.12
PASCAL	40.59	40.64	39.89	40.37
UDISCAL	40.27	40.49	40.01	40.26
SASA	40.27	40.46	39.98	40.24
SASA + UDISCAL	**40.61***	**40.92***	**40.12***	**40.55***
SACrA	40.44	40.68$^{\uparrow}$	39.85	40.33
SACrA + UDISCAL	40.23	40.48	39.96	40.22

Table 7: ChrF scores for the baseline Transformer model, the baseline Syntactically infused models PASCAL and UDISCAL, our SASA and SACrA models, and models incorporating UDISCAL with each of SASA and SACrA, across all WMT's newstests. For every language pair, each column contains the Bleu scores over the WMT newstest equivalent to the column's year (e.g., for En-Ru, the scores under column *2015* are for En-Ru newstest2015). For some newstests, there was more than one version on WMT, each translated by a different person. For those test sets, we included both versions, denoting the second one with a "B". In addition, for every language pair, the right-most column represents the average Bleu scores over all the pair's reported newstests. For every test set (and for the average score), the best score is boldfaced. For each of the semantic models (i.e., SASA and SACrA), improvements over all the baselines (syntactic and Transformer) are marked by an arrow facing upwards. For models with both syntactic and semantic masks, improvements over each mask individually are marked by an asterisk.

En-De

models	2012	2014	2015	2016	2017	2018	2019	2020	2020B	average
Transformer	15.08	16.94	17.36	21.11	14.84	23.43	22.42	16.79	15.75	18.19
PASCAL	14.96	17.45	17.85	20.22	14.66	23.76	21.28	**16.9**	**16.22**	18.14
UDISCAL	14.46	**17.84**	17.7	**21.26**	**15.48**	23.75	22.36	16.37	15.37	18.29
SASA	14.67	17.68	**18.04**$^{\uparrow}$	20.89	15.09	**24.8**$^{\uparrow}$	22.86$^{\uparrow}$	16.85	15.76	**18.52**$^{\uparrow}$
SASA + UDISCAL	**15.39***	17.07	17.38	20.42	15.35	23.53	**22.87***	16.79	15.98*	18.31
SACrA	14.67	17.03	16.89	19.69	14.45	22.21	22.08	16.64	15.6	17.70
SACrA + UDISCAL	15.07*	17.23	16.52	20.82	14.6	22.38	22.61*	16.53	15.81*	17.95

En-Ru

models	2012	2013	2014	2015	2016	2017	2018	2019	2020	2020B	average
Transformer	23.4	14.67	**24**	16.82	17.52	19.74	17.78	17.12	13.39	19.47	18.39
PASCAL	22.6	**15.67**	23.56	17.08	17.79	19.46	17.9	16.13	13.7	19.44	18.33
UDISCAL	23.19	14.75	23.46	17.06	18.17	19.67	18.32	15.7	13.44	**21.14**	18.49
SASA	23.53$^{\uparrow}$	15.38	23.9	17.77$^{\uparrow}$	**18.37**$^{\uparrow}$	**20.12**$^{\uparrow}$	18.33$^{\uparrow}$	16.55	13.37	20.88	**18.82**$^{\uparrow}$
SASA + UDISCAL	23.77*	14.67	23.65	16.96	18.21	19.8	18.06	**17.15***	13.57*	20.02	18.59
SACrA	**23.83**$^{\uparrow}$	15.15	22.86	**18.09**$^{\uparrow}$	18.13	19.98$^{\uparrow}$	**18.7**$^{\uparrow}$	17.1	**13.83**$^{\uparrow}$	19.41	18.71$^{\uparrow}$
SACrA + UDISCAL	22.98	14.58	23.16	16.76	17.37	18.89	17.4	16.07	13.18	18.53	17.89

En-Fi

models	2015	2016	2016B	2017	2017B	2018	2019	average
Transformer	9.57	**11.05**	8.8	11.45	9.99	7.78	10.22	9.84
PASCAL	9.75	10.77	8.72	11.43	10.11	8.06	10.24	9.87
UDISCAL	9.04	10.85	8.63	11.46	10.1	7.7	9.85	9.66
SASA	9.65	10.87	**9.03**$^{\uparrow}$	11.62$^{\uparrow}$	10.1	7.99	10.53$^{\uparrow}$	9.97$^{\uparrow}$
SASA + UDISCAL	9.45	10.96*	8.91	**11.88***	**10.33***	**8.42***	10.62*	10.08*
SACrA	**10.26**$^{\uparrow}$	10.95	8.89$^{\uparrow}$	11.57$^{\uparrow}$	10.13$^{\uparrow}$	8.17$^{\uparrow}$	**10.76**$^{\uparrow}$	**10.10**$^{\uparrow}$
SACrA + UDISCAL	9.42	10.84	8.83	11.51	9.9	7.71	10.7	9.84

En-Tr

models	2016	2017	2018	wikipedia	Eubookshop	mozilla	bible	average
Transformer	7.99	8.15	8.06	7.55	4.87	3.34	0.36	5.76
PASCAL	7.81	7.83	7.69	7.52	5.04	3.41	**0.54**	5.69
UDISCAL	7.68	7.83	7.4	7.63	4.92	3.34	0.49	5.61
SASA	8.2$^{\uparrow}$	8.31$^{\uparrow}$	**8.12**$^{\uparrow}$	7.63	5.21$^{\uparrow}$	3.09	0.52	5.87$^{\uparrow}$
SASA + UDISCAL	7.81	7.92	8.1	7.58	**5.28***	3.36*	0.35	5.77
SACrA	7.75	8.33$^{\uparrow}$	7.51	**7.68**$^{\uparrow}$	5.11$^{\uparrow}$	**3.59**$^{\uparrow}$	0.5	5.78$^{\uparrow}$
SACrA + UDISCAL	**8.23***	**8.54***	7.95*	7.51	5.22*	3.45	0.52*	**5.92***

Table 8: Bleu scores of challenge sentences for the baseline Transformer model, the baseline Syntactically infused models PASCAL and UDISCAL, our SASA and SACrA models, and models incorporating UDISCAL with each of SASA and SACrA, across all WMT's newstests. For every language pair, each column contains the Bleu scores over the WMT newstest equivalent to the column's year (e.g., for En-Ru, the scores under column *2015* are for En-Ru newstest2015). For some newstests, there was more than one version on WMT, each translated by a different person. For those test sets, we included both versions, denoting the second one with a "B". In addition, for every language pair, the right-most column represents the average Bleu scores over all the pair's reported newstests. For every test set (and for the average score), the best score is boldfaced. For each of the semantic models (i.e., SASA and SACrA), improvements over all the baselines (syntactic and Transformer) are marked by an arrow facing upwards. For models with both syntactic and semantic masks, improvements over each mask individually are marked by an asterisk.