Modelling Source- and Target-Language Syntactic Information as Conditional Context in Interactive Neural Machine Translation

Kamal Kumar Gupta, Rejwanul Haque,† Asif Ekbal, Pushpak Bhattacharyya and Andy Way†
Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
†ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
kamal.pcs17,asif,pb@iitp.ac.in
†firstname.lastname@adaptcentre.ie

Abstract

In interactive machine translation (MT), human translators correct errors in automatic translations in collaboration with the MT systems, which is seen as an effective way to improve the productivity gain in translation. In this study, we model source-language syntactic constituency parse and target-language syntactic descriptions in the form of supertags as conditional context for interactive prediction in neural MT (NMT). We found that the supertags significantly improve productivity gain in translation in interactive-predictive NMT (INMT), while syntactic parsing somewhat found to be effective in reducing human efforts in translation. Furthermore, when we model this source- and target-language syntactic information together as the conditional context, both types complement each other and our fully syntax-informed INMT model shows statistically significant reduction in human efforts for a French–to–English translation task in a reference-simulated setting, achieving 4.30 points absolute (corresponding to 9.18% relative) improvement in terms of word prediction accuracy (WPA) and 4.84 points absolute (corresponding to 9.01% relative) reduction in terms of word stroke ratio (WSR) over the baseline.

1 Introduction

Interactive MT (IMT) is viewed as an effective mean to increase productivity in the translation industry. In principle, IMT aims to reduce human effort in automatic translation workflows by employing an iterative collaborative strategy with its two most important components, the human agent and the MT engine. Figure 1 represents the interactive protocol.

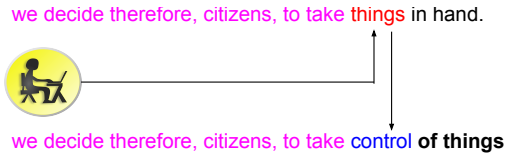


Figure 1: Interactive protocol in collaboration with an MT system and a user. The user wants to translate the French sentence 'Nous décidons donc, citoyens, de prendre les choses en main.' to English. The reference translation is 'we decide therefore, citizens, to take control of things' which is used here to simulate the user. The user corrects the first wrong word (*things*) from the hypothesis. The validated prefix (magenta phrase) and the last modified word (*control*) are fed back to the NMT system which generates a correct suffix (*of things*).

As of today, NMT (Bahdanau et al., 2015; Vaswani et al., 2017) represents the state-of-the-art in MT research. This has led researchers to test interactive-predictive protocol on NMT too, and papers (Knowles and Koehn, 2016; Peris et al., 2017) that pursued this line of research suggest that NMT is superior than phrase-based statistical MT (Koehn et al., 2003) as far as interactive-predictive translation is concerned.

In a different MT research context, Nădejde et al. (2017) have successfully integrated CCG (combinatory categorical grammar) syntactic categories (Steedman, 2000) into the target-side of the then state-of-the-art recurrent neural network (RNN) MT models (Bahdanau et al., 2015). In this work, we investigate the possibility of modelling the target-language syntax in the form of supertags (Bangalore and Joshi, 1999; Steedman, 2000) as a conditional context in an interactive-predictive protocol on Transformer (Vaswani et al.,

2017), the current state-of-the-art NMT model. In a reference-simulated setting, we found that our target-language syntax-informed interactive setup can significantly reduce human effort in a French-to-English translation task.

We also extract syntactic features from constituency-based parse trees of the source French sentences following Akoury et al. (2019), and use them as the conditional context in the interactive-predictive Transformer framework. Experiments show that this contextual information can reduce human efforts in translation to some extent.

In addition, we apply the above strategies together, and model supertags and constituency parse tree-based features collectively as the conditional context for interactive prediction in NMT. Our experimental results indicate that these syntactic feature types are complementary. As a result, this collaborative strategy turns out to be the best-performing in the French-to-English task while significantly outperforming those setups that include either feature type on WPA and WSR. To the best of our knowledge, this is the very first study that investigates the possibility of integrating syntactic knowledge into an interactive MT model.

2 Related Work

Foster et al. (1997) were the first to introduce the idea of interactive-predictive MT as an alternative to pure post-editing MT. There have been a number of papers that explored this strategy in order to minimise human effort in translation and cover many use-cases involving SMT: e.g. applying online (Ortiz-Martínez, 2016) and active (González-Rubio et al., 2012) learning techniques, use of translation memories (Barrachina et al., 2009; Green et al., 2014), predicting the partially typed words and prefix matching (Koehn et al., 2014), word-graphs for reducing response time (Sanchis-Trilles et al., 2014), alignment based post-editing (Simianer et al., 2016), segment-based approaches (Peris et al., 2017), suggesting more than one suffix (Koehn, 2009), and exploring multimodal interaction (Alabau et al., 2014). This use-case has also been moderately tested on NMT, e.g. (Knowles and Koehn, 2016; Wuebker et al., 2016; Peris and Casacuberta, 2018; Lam et al., 2019). To the best of our knowledge, no one has investigated the interactive-predictive protocol on the state-of-the-art Transformer.

The strategy of exploiting syntactic knowledge from the source and/or target languages for improving the translation quality is not new in MT research. It was successfully applied in the era of classical MT (Hassan et al., 2007; Haque et al., 2011), and is continually being applied to improve the current state-of-the-art NMT models, e.g. (Luong et al., 2016; Nădejde et al., 2017).

3 Fully Syntactified Interactive NMT

This section presents our fully syntactified interactive NMT model. In NMT, at time step *i*, the conditional probability of predicting output token y_i given a source sentence x_1^J and the previously generated output token $y_1, ..., y_{i-1}$ is modelled as $\mathrm{p}(y_i|\{y_1, ..., y_{i-1}\}, x_1^J)$.

In the interactive protocol, the user corrects the wrongly translated word (by the MT system) which appears at the left-most side. The feedback is returned back to the MT system in the form of $\hat{y}_1^{i-1}$ which is the validated prefix together with the corrected word $\hat{y}_{i-1}$. Thus, in interactive NMT, the conditional context becomes $\hat{y}_1^{i-1}$, and the conditional probability of predicting output token y_i is modelled as $\mathrm{p}(y_i|\{\hat{y}_1, ..., \hat{y}_{i-1}\}, x_1^J)$. This model serves as our baseline in this work.

In our supertag-based interactive-predictive scenario, we first predict the CCG supertag ($\hat{s}_i$) of the word (y_i) to be predicted next. As a result, the length of the conditional context becomes twice the number of words in context plus one. As far as the target-syntactified interactive NMT is concerned, the conditional probability of predicting the output token y_i is modelled as $\mathrm{p}(y_i|\{\hat{s}_1, \hat{y}_1, ..., \hat{s}_{i-1}, \hat{y}_{i-1}, \hat{s}_i\}, x_1^J)$, where $\hat{s}_1^{i-1}$ is the CCG sequence of the validated prefix $\hat{y}_1^{i-1}$ and $\hat{s}_i$ is the supertag of the word (y_i) to be predicted next.

As for the modelling of source-side syntax, we extract a chunk sequence from the constituency parse tree of a source sentence by setting random a maximum chunk size ($\{1...6\}$) for every sentence (cf. Section 4.2).

Let us define a chunk sequence c_1^M extracted from the input source sentence x_1^J, where M is the number of chunk identifiers (a concatenation of the constituent type and subtree size) of the chunk sequence. This results in an input sequence l_1^{J+M}, where J is the total number of words arbitrarily separated by M number of chunk identifiers. In this model, at time step *i*, the conditional probability of predicting output token y_i given a source sequence (words and chunk identifiers) l_1^{J+M}, and the validated prefix together with the corrected token $\hat{y}_1, ..., \hat{y}_{i-1}$ is modelled

as $\mathrm{p}(y_i|\{\hat{y}_1, ..., \hat{y}_{i-1}\}, l_1^{J+M})$.

In our fully syntactified interactive NMT model, the conditional probability of predicting the output token y_i is modelled as $\mathrm{p}(y_i|\{\hat{s}_1, \hat{y}_1, ..., \hat{s}_{i-1}, \hat{y}_{i-1}, \hat{s}_i\}, l_1^{J+M})$, where $\hat{s}_1^{i-1}$ is the CCG sequence of the validated prefix $\hat{y}_1^{i-1}$, $\hat{s}_i$ is the supertag of the word (y_i) to be predicted next, and l_1^{J+M} is the input sequence constituting J and M numbers of words and chunk identifiers, respectively.

4 Syntactic Context Features

4.1 Modelling CCG Supertags as Target Language Context

This section explains why we consider a rich and complex syntactic feature, supertags, as context in our experiments. Supertags (Bangalore and Joshi, 1999; Steedman, 2000) are known to be context-sensitive tags that preserve the global syntactic information at local lexical level. Having this property, supertags resolve ambiguity in short- and long-distance dependencies by capturing the preceding and succeeding syntactic dependencies of a lexical term. For example, they signify whether a particular lexical term is expecting a preposition as an argument in order to complete the sentence.

The interactive neural MT models predict a new hypothesis primarily based on the validated context (prefix) including the left-most modified word by the user. In the case of our syntax-informed model, prediction of the next word is also conditioned on CCG supertags (Steedman, 2000) of the validated prefix and the word to be predicted next. Our intuition underpinning this is that such kinds of rich syntactic knowledge sources, which inherently capture long-distance word-to-word dependencies in a sentence, may be useful to improve the prediction quality of interactive NMT, especially for the longer sentences.

4.2 Modelling Syntactic Parse as Source Language Context

Following Akoury et al. (2019) we extract a chunk sequence from the constituency parse tree of a source sentence. Akoury et al. (2019) conducted a series of experiments for getting optimal value (k) for the maximum size of a chunk (subtree). In particular, they tested random and fixed value for (k). The random k ($\{1...6\}$) was found to be best-performing when chunk identifiers were autoregressively predicted in the target using Transformer (Akoury et al., 2019). In our experiments, we adopted their best-up and set the maximum size of a chunk (subtree) random ($\{1...6\}$) for every sentence. Note that a chunk identifier represents a concatenation of the constituent type and subtree size (e.g. VP2). In our case, the chunk identifiers encode additional contextual knowledge on the source side. We adopt the procedure described in Akoury et al. (2019) in order to extract chunk sequences for the source French sentences using the Berkeley Neural Parser.[1] As an example, Table 1 shows a chunk sequence extracted from a French sentence 'si le cliquable doit être à l'état pressé' in row B. The third row of the table (cf. row C) shows the resulting input sequence which is a combination of words and chunk identifiers. As for the chunk identifier, we see from Table 1 that NP3 is a combination of the constituency label NP and the number of terminals of the subtree ('l' état pressé'), i.e. 3. Note that for this example sentence the maximum size of a subtree was 3.

5 Experimental Setups

5.1 Methods of forming conditional syntactic context

In theory, prediction of an output token in the interactive-predictive scenario is conditioned on a user-validated prefix and the input sentence. As discussed above, we model rich syntactic features from the constituency-based parse trees as source context with an expectation to improve the prediction quality in INMT. Hence, in our case, the source-side context is an input sequence of words and chunk identifiers. In interactive mode, if the user makes a correction, the conditional context is modified, i.e. the validated prefix including the last modified word is provided to the MT model for the prediction of the remaining hypothesis. Nonetheless, the source-side context including our syntactic parse features remains unchanged over the course of generation of the target translation.

We model target-side syntactic contexts (CCG supertags) as conditional context in two different ways as follows. In our first setup, we directly use the supertags that are predicted by Transformer as a part of the conditional context for the prediction of the remaining hypothesis. It implies that the setup follows the interleaving technique of Nădejde et al. (2017) in which the CCG tag of a token is kept before its token as shown in Table 1. For example, $word_i$ is produced by the decoder in a hypothesis having ccg_i as its CCG supertag that was predicted in the previous time step. If the

[1] https://github.com/nikitakit/self-attentive-parser

A	à la 4e séance , M Oberthür a rendu compte des résultats des consultations
B	P1 NP3 PONCT1 NC1 PONCT1 VN3 P+D1 NP3
C	P1 à NP3 la 4e séance PONCT1 , NC1 M PONCT1 Oberthür VN3 a rendu compte P+D1 des NP3 résultats des consultations
D	P1 à NP3 la 4e séance PONCT1 , NC1 M PONCT1 Ober@@ PONCT1 th@@ PONCT1 ü@@ PONCT1 r VN3 a rendu compte P+D1 des NP3 résultats des consultations
E	at the 4th meeting , Mr. Oberthür reported on the results of the consultations
F	(S/S)/NP at NP[nb]/N the N/N 4th N meeting N/N , N/N Mr. N Oberthür (S[dcl]\NP)/NP reported PP/NP on NP[nb]/N the N results (NP\NP)/NP of NP[nb]/N the N consultations
G	(S/S)/NP at NP[nb]/N the N/N 4th N meeting N/N , N/N Mr. N Ober@@ N th@@ N ü@@ N r (S[dcl]\NP)/NP reported PP/NP on NP[nb]/N the N results (NP\NP)/NP of NP[nb]/N the N consultations

Table 1: A: a French sentence, B: chunk identifiers, C: input sequence: a combination of the French words and chunk identifiers, D: the segmented version of the French sentence, E: an English sentence, F: the English sentence with CCG supertags, G: the segmented version of the English sentence.

Input sentence	il y a des voitures neuve et chère à tout les coins de rue, exactement comme avant la crise de 2008.
Input sequence with parsing info	VN3 il y a DET1 des NC1 voitures AP3 neuve et chère P1 à ADJ1 tout DET1 les NC1 coins P1 de NC1 rue PONCT1 , ADV1 exactement P1 comme P1 avant DET1 la NC1 crise PP2 de 2008 PONCT1 .
Reference	there are new and expensive cars on every street corner , exactly like before the 2008 crisis .
Initial hypothesis	there (S[dcl]\NP[thr])/NP are N/N new conj and N/N sh@@ N/N ere N cars ((S\NP)\(S\NP))/NP across NP[nb]/N the N/N streets N , ((S\NP)\(S\NP))/((S\NP)\(S\NP)) just ((S\NP)\(S\NP))/((S\NP)\(S\NP)) as ((S\NP)\(S\NP))/PP prior PP/NP to NP[nb]/N the N/N 2008 N/N crisis N .
Hypothesis after several iterations	NP[thr] there (S[dcl]\NP[thr])/NP are N/N new conj and N/N expensive N cars ((S\NP)\(S\NP))/NP on NP[nb]/N every N/N street N corner **((S\NP)\(S\NP))/((S\NP)\(S\NP)) just** ((S\NP)\(S\NP))/((S\NP)\(S\NP)) as ((S\NP)\(S\NP))/PP prior PP/NP to NP[nb]/N the N/N 2008 N/N crisis N .
INMT interface	there are new and expensive cars on every street corner **just** as prior to the 2008 crisis .
Correction by user	there are new and expensive cars on every street corner **,** as prior to the 2008 crisis .
Applying on the fly CCG supertagger	NP[thr] there (S[dcl]\NP[thr])/NP are N/N new conj and N/N expensive N cars ((S\NP)\(S\NP))/NP on NP[nb]/N every N/N street N/N corner **N ,** ((S\NP)\(S\NP))/((S\NP)\(S\NP)) as ((S\NP)\(S\NP))/PP prior PP/NP to NP[nb]/N the N/N 2008 N/N crisis N .
New hypothesis	there are new and expensive cars on every street corner **,** exactly like before the 2008 crisis .

Table 2: An example showing applying *On the fly CCG supertagger* on hypothesis.

user sees that $word_i$ is not appropriate in the context (i.e. it is incorrectly predicted by the system), the user edits/removes $word_i$ and replaces it with a new token $word_{new}$. Now, when the modified context (i.e. validated prefix) is fed back to the NMT model, $word_{new}$ will have the tag of $word_i$, i.e. ccg_i. In other words, the final two tokens of the conditional context would be $ccg_i\ word_{new}$. We carried out an analysis to see how closely these supertags are related to the new words added by the user (cf. Section 6.4). In this regard, we applied BPE segmentation on the training sentences. The sub-word units of a word inherit the CCG category of the word. As an example, we show an English sentence with supertags in Table 1. We see from row E of Table 1 that CCG 'N' of a word 'Oberthür' is distributed over its sub-words (i.e. Ober@@ th@@ ü@@ and r). Our first experimental setup is referred to as PredCCG.

Akoury et al. (2019) showed that integrating target-side ground-truth syntactic information into Transformer at decoding time significantly improved translation quality, and their syntax-based model outperformed the baseline Transformer model by a large margin in terms of BLEU (Papineni et al., 2002). In reality, there is no way of obtaining the target-side ground-truth syntactic information at decoding time. However, in interactive-predictive mode, we found a way to obtain a slightly better CCG sequence for the partial translation (i.e. validated prefix) and inject them into the model at run-time, which we believe can positively impact the model's subsequent predictions. In other words, in our second setup, we integrate a CCG supertagger into our INMT framework, and apply that on the validated prefix and unchecked suffix on the fly. The tagger is invoked when the user makes a correction. As an example, when the user inserts a new token $word_{new}$ in place of an incorrectly predicted token ($word_i$), the CCG supertagger is invoked and applied to the validated prefix and unchecked suffix on the fly. In Table 2, we show how *On the fly CCG supertagger* is applied in our interactive interface. We see from rows 6 and 7 of Table 2 that the user replaces the wrongly predicted token *just* with a correct token ','. The CCG supertag ((S\NP)\(S\NP))/((S\NP)\(S\NP)) of the incorrect token 'just' is assigned to the new token ',' which is incorrect in this context. When the user commits this change, *On the fly CCG supertagger* is invoked and applied to the corrected hypothesis

(a combination of validated prefix and unchecked suffix). As can be seen from row 8 of Table 2, a new CCG tag sequence is generated for the hypothesis, and we see that CCG (N) of the newly added token ',' is correct. Finally, INMT predicts another suggestion (row 9 of Table 2) where we see the remaining predictions are correct in the context. We call this experimental setup OnflyCCG. Note that the model is trained at sub-word level and generates sub-words at output; however, word level tokens are presented to the user. Naturally, *On the fly CCG supertagger* is applied to a hypothesis of word level.

5.2 MT systems

We carry out experiments with French-to-English with the UN corpus[2] (Ziemski et al., 2016). The training and development sets contain 12,238,995 and 1,500 sentences, respectively. We use 1,500 sentences from the WMT15 news test set *newstest2015* as our test set. In order to build our MT systems, we use the Sockeye[3] (Hieber et al., 2018) toolkit. Our training setups are as follows. The tokens of the training, evaluation and validation sets are segmented into sub-word units using BPE. We performed 30,000 join operations. We use 6 layers in the encoder and decoder sides, an 8-head attention, hidden layer of size 512, embedding vector of size 512, learning rate 0.0002, and minimum batch size of 1,800 tokens. EasyCCG[4] (Lewis and Steedman, 2014), a CCG supertagger, is used for generating the CCG sequence for the English sentences.

Transformer (Baseline)	26.90
Source Syntactified (SS)	26.96
Target Syntactified (TS)	27.10
Fully Syntactified (FS)	27.36 (p-value: 0.059)

Table 3: The BLEU scores of baseline and syntactified NMT systems.

Table 3 shows the performance of our baseline and syntax-sensitive NMT systems in terms of BLEU. The second and third rows represent the NMT models that incorporate source- and target-language syntactic contexts, respectively, which we call source- (SS) and target-syntactified (TS) NMT systems, respectively. We see from Table 3 that the BLEU scores of these two MT systems and Transformer are very similar. Additionally, we performed statistical significance test using bootstrap resampling methods (Koehn, 2004). We found that the differences of the BLEU scores of these MT systems are not statistically significant.

The fourth row shows the BLEU score of the NMT system that integrates both the source- and target-language syntactic contexts (i.e. supertags and syntactic parse, respectively) together. We call this model our fully syntactified (FS) NMT system. The FS NMT system produces a 0.46 BLEU point (corresponding to 1.7% relative) gain on the test set over the baseline. The differences of the BLEU scores of the FS and baseline Transformer models are not statistically significant either. When we integrate the source- and target-language syntactic contexts individually into Transformer, they do not positively impact the system's performance. However, when we integrate them collectively into the model, we see that they bring a moderate gain in terms of BLEU over the baseline, and the gain is very close to the significance level (p-value: 0.059) too. It seems that both contextual features complement each other and bring about an (moderate) improvement. Although the primary objective of this work is to observe the prediction of Transformer in an interactive-predictive platform while modelling different syntactic constraints as conditional context, this can also be seen as an important finding to MT research.

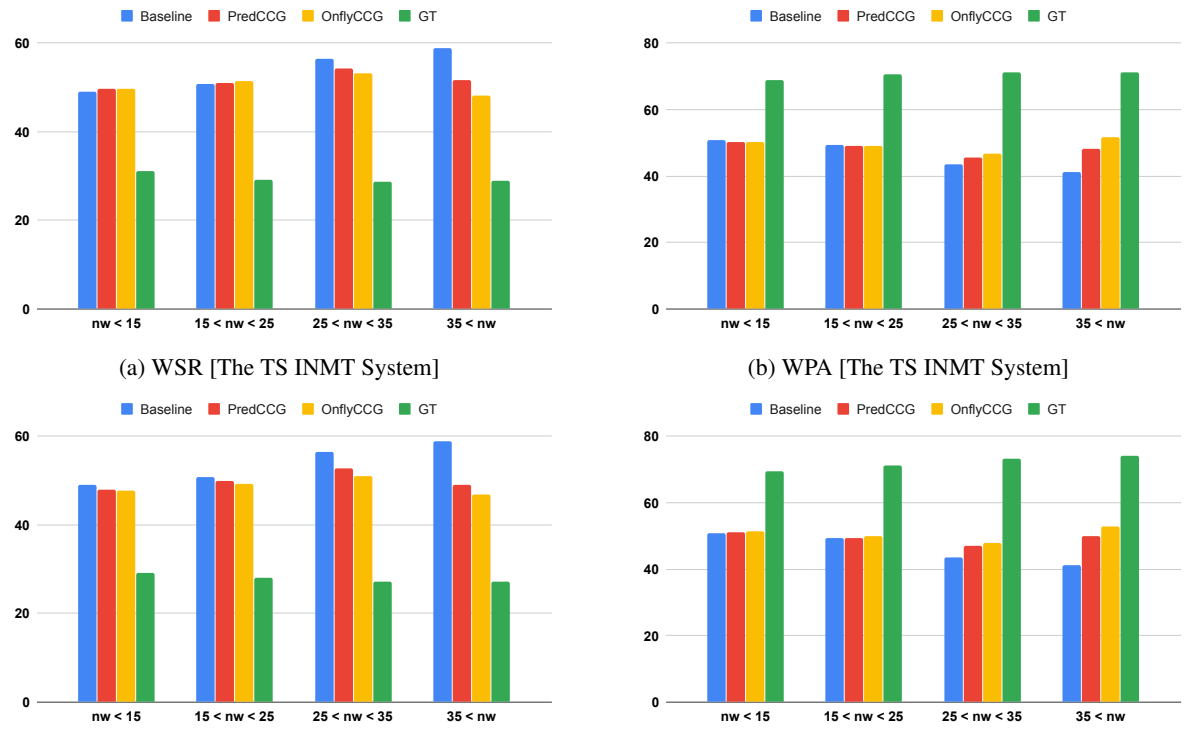
6 Results and Discussion

In this section, first we explain the strategy that we adopted for evaluating the interactive-predictive MT systems. Then, we present our evaluation results along with some discussions and analysis.

6.1 Evaluation Plan for INMT

We evaluate the performance of the INMT systems using two evaluation metrics, WSR and WPA. WSR denotes the total number of token replacements required to obtain the desired hypothesis (Peris et al., 2017). WPA is the percentage of words that the INMT system predicts correctly, given a prefix of all the previous translator-produced words (Knowles and Koehn, 2016). WSR and WPA are calculated on word level. The process of evaluating translations in interactive scenarios is expensive as it requires human evaluators. As an alternative, we adopted a reference-simulated evaluation strategy as in Peris et al. (2017), where instead of taking feedback from the real user, the reference sentence is used as the feedback. In other words, each time an interactive MT model generates a hypothesis it is com-

[2]https://www.statmt.org/wmt13/training-parallel-un.tgz
[3]https://github.com/awslabs/sockeye
[4]https://github.com/mikelewis0/easyccg



(a) WSR [The TS INMT System]

(b) WPA [The TS INMT System]

(c) WSR [The FS INMT System]

(d) WPA [The FS INMT System]

Figure 2: WSR and WPA scores of the syntax-informed and baseline INMT systems with respect to sentence lengths.

pared with the reference sentence from left to right.

6.2 Evaluation Results

6.2.1 The SS INMT System

In this section, we present the evaluation results that we obtain using the source-language syntactic constituency parse as conditional context in the interactive-predictive Transformer model. The WSR and WPA scores of the baseline and SS Transformer models are shown in Table 4. Note that WSR is an error metric, which means that lower scores are better. We see from the table that integrating this context into the model brought about a 0.56 point absolute (corresponding to 1.04% relative) reduction and a 0.31 point absolute (corresponding to 0.66% relative) gain in terms of WSR and WPA, respectively, over the baseline. We use approximate randomization (Yeh, 2000) to test the statistical significance of the difference between the two systems. We found that these differences are not statistically significant. These results indicate that using the syntactic constituency parse as context in interactive neural MT models has only a minor impact on reducing human effort in translation.

	Baseline	SS INMT
WSR	53.77	53.21
WPA	46.82	47.13

Table 4: Performance of the SS INMT System.

6.2.2 The TS INMT System

In this section, we obtain experimental results to evaluate the interactive-predictive Transformer model that uses target-language supertags as conditional context on the test set. We report the results in Table 5. The third and fourth columns of Table 5 represent two setups (PredCCG and OnflyCCG) that we describe in Section 5.1. The first column of the table represents the baseline Transformer system. The gains in WSR and WPA over the baseline are found to be the highest when *On the fly CCG supertagger* is applied on the user modified hypothesis (cf. Section 5.1). With this, we achieve a 3.16 point absolute (corresponding to 5.87% relative) reduction and a 2.65 point absolute (corresponding to 5.65% relative) improvement in terms of WSR and WPA, respectively, on the test set over the baseline. These differences are statistically significant. When we compare PredCCG and OnflyCCG setups, we see that OnflyCCG brings a 1.09 WSR point absolute (corresponding to 2.10% relative) reduction and a 1.18 WPA point absolute (corresponding to 2.44% relative) improvement over the PredCCG setup, which

	Baseline	PredCCG	OnflyCCG	GT
WSR	53.77	**51.70**	**50.61**	**29.44**
WPA	46.82	**48.29**	**49.47**	**70.53**

Table 5: Performance of the TS INMT System

are statistically significant too. This indicates that especially with the OnflyCCG setup supertags as target-language context can have significant impact on reducing human effort in translation.

For comparison we also report the WPA and WSR scores of our TS INMT system on an ideal setup, i.e. when we feed Transformer with ground-truth CCG supertags instead of those predicted by the Transformer or generated by *On the fly CCG supertagger*. As expected, this setup surpasses the baseline and context-based setups by a large margins in terms of WSR and WPA.

6.2.3 The FS INMT System

As discussed above, we use both source and target syntax as the conditional context in interactive prediction in NMT. The first two rows of Table 6 represent the evaluation results obtained by integrating both as a collective feature into the INMT model. This feature brings about a statistically significant improvements in terms of WPA and WSR, respectively, over the baseline across two setups: PredCCG and OnflyCCG. We see from Table 6 that OnflyCCG is the best-performing setup that produces a 4.84 point absolute (corresponding to 9.01% relative) reduction and a 4.30 point absolute (corresponding to 9.18% relative) improvement in terms of WSR and WPA, respectively over the baseline.

	Baseline	PredCCG	OnflyCCG	GT
WSR	53.77	**50.03**	**48.93**	**28.24**
WPA	46.82	**49.67**	**51.12**	**71.69**
WSR		-1.67	-1.68	-1.20
WPA		+1.38	+1.65	+1.16

Table 6: Performance of the FS INMT System.

As for PredCCG and OnflyCCG, the FS INMT model with OnflyCCG statistically significantly surpassed the one with PredCCG as far as reduction of human effort is concerned. As above, we see that the ideal setup (GT) again surpasses the baseline and context-based setups by large margins. We make a comparison of Table 5 and 6 for the three setups (PredCCG, OnflyCCG, and GT), and differences in WSR and WPA scores are presented in the last rows of Table 6. We see consistent reductions in WSR and improvements in WPA across the three setups with the combined contextual features, which are statistically significant.

CCG as target context and, to a certain extent, syntactic parse as source context were found to be effective in reducing human effort when applied individually. Nevertheless, CCG (target) and syntactic parse (source) together as a context turn out to be the best-performing setup with statistically significant gains over either feature type. In this sense, we can say that source and target-side syntactic contextual features complement each other as far as neural interactive prediction is concerned. We conjecture that since the conditional context includes source-language syntactic constituency parse and target-language syntactic constructs in the form of CCG supertags together, it provides the NMT model with better syntactic agreement between the source and target sentences, which, in turn, helps the model generate better predictions.

6.3 Impact on Sentence Lengths

For further analysis, we place the sentences of our test set into four sets (c.f. Figure 2) as per the sentence length measures, i.e. number of words nw$<$15, 15$<$nw$\leq$25, 25$<$nw$\leq$35 and 35$<$nw. This division was made based on the lengths of reference sentences. In Figure 2, we plot the distributions of WPA and WSR scores over the sentence length-based sets. As can be seen from the figure, both the TS and FS INMT systems produce increasingly better WSR and WPA scores as the length of the reference sentences increases. As discussed above, supertags encode wider context of a sentence, which could help the decoder to capture long-range word-to-word dependencies at generation time. In other words, as the length of the validated prefix increases, the corresponding CCG supertag sequences help better predict the subsequent tokens correctly.

6.4 CCG supertags of the Words of User Choice

	Fr–>En (TS)		Fr–>En (FS)	
	PredCCG	OnflyCCG	PredCCG	OnflyCCG
Whole testset	41.07	23.95	39.58	22.52
nw $<$15	40.64	23.88	40.25	22.02
15 $<$nw $<$25	40.84	23.04	39.44	21.92
25 $<$nw $<$35	42.80	25.28	40.19	23.35
35 $<$nw	39.32	24.33	38.06	22.89

Table 7: % of CCG supertags that becomes incorrect when the user replace the incorrectly predicted token in hypothesis with the token of his choice.

As mentioned in Section 5.1, we came up with two different ways to use the target-language su-

pertags as conditional context for the predictions in INMT. First, in the PredCCG setup, if the user makes a correction, the user's choice of word inherits the CCG supertag of the word that the user has just corrected, which, in fact, is predicted by the INMT system. The new word and the incorrect word that the user has just corrected could be syntactically or semantically different. As a result, the supertag that the new word inherits could be incorrect. We calculate the percentage of CCG supertags that are incorrect for the new words when the predicted words were wrong and edited by the user. We also produce such statistics for the second experimental setup, OnflyCCG. In Table 7, we show the percentage of CCG supertags those were incorrectly assigned to new words on both the experimental setups. We clearly see from the table that the second setup (OnflyCCG) is far better than the first setup (PredCCG) in terms of assigning correct CCG tags to the new words that the user has just corrected, i.e. better by 17.06% to 17.12%. This is seen consistently across the sentence length-based sets too. When we compare this across the TS and FS INMT systems, we see that the percentage of correctly assigned CCG tags to the words of the user's choice in the FS INMT system is higher (by 1.43%) than the TS INMT system on the test set.

6.5 Latency for the CCG supertagger

We calculate the average delay for a correction (i.e. processing time) by the user for baseline, PredCCG, OnflyCCG and GT (ground-truth) setups using the TS INMT system, which are shown in Table 8. We see from the table that the delays are comparable across the systems. As for OnflyCCG, we exclusively calculate the average latency for applying the CCG supertagger, which is found to be 0.12 seconds only. Hence, the supertagger does not bring much computational overhead and impact latency as far as translation time in the interactive-predictive platform is concerned.

Baseline	PredCCG	OnflyCCG	GT
0.28	0.35	0.47	0.28

Table 8: Average Latency (in seconds) for generating modified hypothesis

6.6 Average Number of Partial Hypothesis Processed

In the interactive protocol, when the user makes a correction, the MT system re-translates the source sentence given the validated partial hypothesis. Finally, the new translation is shown to the user. In Table 9, we show the average number of partial hypotheses processed (i.e. how many the MT system has to re-translate) for each sentence in the test set. For this analysis, we consider all the experimental setups (PredCCG, OnflyCCG and GT) and MT system types (SS, TS and FS INMT). We see from Table 9 that the OnflyCCG on FS INMT setup wins out if we omit the ideal setup (GT). In other words, source- and target-language syntactic contexts in combination have more impact in INMT than either type individually.

		PredCCG	OnflyCCG	GT
Baseline	8.91			
SS	8.84			
TS		8.56	8.38	5.72
FS		8.24	8.11	5.36

Table 9: Average number of partial hypothesis processed.

7 Conclusion

In this paper, we have integrated a rich and complex syntactic knowledge in the form of supertags and/or syntactic constituency parse into the current state-of-the-art neural MT model, Transformer. Furthermore, we tested whether integration of such knowledge sources into Transformer could indeed reduce human efforts in translation in an interactive-predictive scenario. We carried out our experiments on French-to-English, a high resource widely-used translation-pair in industry. We compared our syntax-informed and baseline Transformer models on an interactive-predictive platform. The use of syntactic constituency parse as conditional context has minor impact on reducing human effort in translation. We modelled target-language supertags as conditional context in interactive NMT in two different ways, and both of these significantly positively impact productivity in translation.

Interestingly, supertags (target) and constituency parse (source) together as a context turns out to be the best-performing setup with significant gains over either feature type. In this sense, we can say that source and target-side syntactic contextual features complement as far as neural interactive prediction is concerned. In fact, the conditional context in this setup includes both source-language constituency parse and target-language CCG, which essentially provides the INMT model with better syntactic agreement between the source and target sentences.

We conjecture that this could be the reason why this collaborative strategy turned out to be best-performing.

Our analysis shows that the OnflyCCG setup (where CCG assigned by *On the fly CCG supertagger*) significantly outperformed PredCCG (where CCG predicted by Transformer) in terms of assigning correct CCG to the words of user's choice by large margins (17.06% to 17.12%). In fact, our proposed setup (OnflyCCG), to a certain extent, provides a way to inject correct context into the interactive model. This could be the reason why OnflyCCG turned out to be best-performing.

Our analysis unraveled many sides of our syntax-aware models in an interactive-predictive environment. For an example, we particularly found that our syntax-informed interactive-predictive models have positively impacted more for the translation of longer sentences. Given the importance of interactive MT in translation industry, the findings of this work can be crucial for their production as our methods can positively impact their productivity gain in translation.

Given the fact that linguistic tools such as supertaggers and constituency parsers are only readily available for a handful of languages, in future, we will continue to pursue this line of investigation with exploring integration of language-independent contextual knowledge in interactive-predictive NMT. In future, we plan to evaluate our interactive MT systems with human agents.

8 Acknowledgement

Authors gratefully acknowledge the generous grant of TDIL, MeiTY, Govt. of India for the project "Hindi to English Machine Translation for Judicial Domain [11(3)/2015-HCC(TDIL]" to carry out this research. Asif Ekbal acknowledges Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly MediaLab Asia). Rejwanul Haque and Andy Way acknowledge the ADAPT Centre for Digital Content Technology, funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Akoury, N., Krishna, K., and Iyyer, M. (2019). Syntactically supervised transformers for faster neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 1269–1281, Florence, Italy.

Alabau, V., Sanchis, A., and Casacuberta, F. (2014). Improving on-line handwritten recognition in interactive machine translation. *Pattern Recognition*, 47(3):1217–1228.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

Bangalore, S. and Joshi, A. K. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., et al. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Foster, G., Isabelle, P., and Plamondon, P. (1997). Target-text mediated interactive machine translation. *Machine Translation*, 12:175–194.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2012). Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254, Avignon, France.

Green, S., Chuang, J., Heer, J., and Manning, C. D. (2014). Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 177–187.

Haque, R., Naskar, S. K., van den Bosch, A., and Way, A. (2011). Integrating source-language context into phrase-based statistical machine translation. *Machine Translation*, 25(3):239–285.

Hassan, H., Sima'an, K., and Way, A. (2007). Supertagged phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 288–295, Prague, Czech Republic.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2018). The sockeye neural machine translation toolkit

at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA.

Knowles, R. and Koehn, P. (2016). Neural interactive translation prediction. In *Proceedings of the Association for Machine Translation in the Americas*, pages 107–120, Austin, TX.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.

Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48–54, Edmonton, AB.

Koehn, P., Tsoukala, C., and Saint-Amand, H. (2014). Refinements to interactive translation prediction based on search graphs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 574–578, Baltimore, MD.

Lam, T. K., Schamoni, S., and Riezler, S. (2019). Interactive-predictive neural machine translation through reinforcement and imitation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 96–106, Dublin, Ireland.

Lewis, M. and Steedman, M. (2014). A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-task sequence to sequence learning. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico.

Nădejde, M., Reddy, S., Sennrich, R., Dwojak, T., Junczys-Dowmunt, M., Koehn, P., and Birch, A. (2017). Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark.

Ortiz-Martínez, D. (2016). Online learning for statistical machine translation. *Computational Linguistics*, 42(1):121–161.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Peris, Á. and Casacuberta, F. (2018). Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium.

Peris, Á., Domingo, M., and Casacuberta, F. (2017). Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Sanchis-Trilles, G., Ortiz-Martínez, D., and Casacuberta, F. (2014). Efficient wordgraph pruning for interactive translation prediction. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 27–34, Prague, Czech Republic.

Simianer, P., Karimova, S., and Riezler, S. (2016). A post-editing interface for immediate adaptation in statistical machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 16–20, Osaka, Japan.

Steedman, M. (2000). The syntactic process. *MIT Press*, 24.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wuebker, J., Green, S., DeNero, J., Hasan, S., and Luong, M.-T. (2016). Models and inference for prefix-constrained machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Berlin, Germany.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING 2000*, pages 947–953, Saarbrücken, Germany.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia.