# Target-side CCG Supertag Prediction Improves Machine Translation

*Seth Aycock*


Word count: 7979

Master of Science

Speech & Language Processing

University of Edinburgh

2021

# Abstract

Neural machine translation (NMT) models learn some syntactic information about sequential source and target language text without explicit instruction. This knowledge is incomplete, meaning NMT systems poorly model complex syntactic phenomena including agreement and attachment ambiguities. This study's central thesis is that incorporating target-side syntax in training Transformer-based NMT models improves translation quality. I propose a novel method to directly model target language syntax in Transformer-based NMT models, by training the model to predict an attention distribution over lexical syntactic tags prior to predicting output words, with a composite, decaying loss function. I employ Combinatory Categorial Grammar (CCG) supertags to represent syntactic constraints on a lexical level. I evaluate the method against a baseline NMT model, finding small, consistent improvements of 0.4-0.7 BLEU on WMT17 data from Turkish to English. Improvements are independent of adding source syntax and monolingual data, and complementary to the latter; and ablation experiments show improvements stem specifically from the utility of CCG supertags and the proposed tag attention method. The improved Turkish→English translation quality results in part from better agreement and word order handling for complex constructions.

*Keywords* — Neural machine translation, Syntax, Combinatory Categorial Grammar

# Acknowledgements

First, I would like to thank my supervisor Miloš Stanojević for agreeing to supervise me, and for his unwavering help and guidance throughout the process – I learned a huge amount. I would also like to thank Mark Steedman for his expert high-level assistance, especially with CCG, and who was a pleasure to work with. The MSc SLP and PPLS staff deserve a mention, for getting us all through this most difficult of years. I thank Tolúlọpẹ́ too, for her willingness to lend a hand.

Special thanks must go to my supervisors and lecturers at the University of Cambridge, who forged my love for linguistics. I must specifically thank Ian Roberts for his teaching of Minimalist syntax and our many hours of discussion over my undergraduate dissertation. I also thank Andrew Caines, who was instrumental in instigating my first foray into NLP in 2018; our first meeting was pivotal, and without his encouragement and enthusiasm for NLP I may not have taken this computational route.

I would finally like to thank my family and friends, without whom this academic journey would scarcely have been possible. It's been a wild ride.

Any errors are, of course, the author's own.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Seth Aycock*)

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Problem Statement

State-of-the-art neural machine translation (NMT) models implicitly learn to model syntactic phenomena such as agreement dependencies (Linzen et al., 2016; Raganato and Tiedemann, 2018). While NMT systems' modelling of syntax improves on phrase-based statistical machine translation models (Bentivogli et al., 2016), it is not perfect. Past work has attempted to incorporate syntactic information directly into LSTM-based and Transformer-based sequence-to-sequence NMT models, on both source and target sides.

In this work, I explore the consequences of incorporating explicit target-side syntactic information in a Transformer NMT model. The method trains the model to predict attention over lexical syntactic labels, requiring non-trivial syntactic resources for annotating target language data. I perform experiments in a low-resource setting from Turkish→English, exploiting the substantial target-side resources. I use a statistical parser to annotate English data with Combinatory Categorial Grammar (CCG) supertags (Lewis and Steedman, 2014) which indicate words' categories and syntactic dependencies (detailed fully in Section 2.2.1). For example, a transitive verb supertag, $(\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}$, indicates a noun phrase NP is required to the right and left, which gives a declarative sentence, S[dcl]. Figure 1.1 introduces a Turkish-English example; the proposed method's intuition is to translate from the Turkish source sentence, via prediction of CCG supertags, into the English target sentence.

This subject *wh*-question example demonstrates the divergent word orders of Turkish and English, posing a challenge for NMT. CCG supertags can help by efficiently representing syntactic information about distant elements within the sentence. Figure

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | Kim | onun | bunu | kazanmasını | ister | ? | | | | | | |
| Src-Gloss | *who* | *him* | *it* | *win*-INF | *want*-PRES | ? | | | | | | |
| Target | Who | | | wants | him | to | | win | | it | ? | |
| Trg-CCG | S[wq]/(S[dcl]\NP) | ((S[dcl]\NP)/(S[to]\NP))/NP | NP | (S[to]\NP)/(S[b]\NP) | (S[b]\NP)/NP | NP | . | | | | | |

Figure 1.1: A Turkish→English example translation with source gloss and target CCG supertags.

1.1 will form a running example throughout this work.

The proposed method predicts an attention distribution over the CCG supertag vocabulary before predicting each word, with a composite word-tag loss function. I test the method against models with different target-side tags, a multi-task learning model, a model interleaving tags and words, plus models incorporating monolingual data and source syntax. The proposed method improves translation quality over a baseline system on WMT data by approximately 0.4-0.7 BLEU, with improvements independent of and largely complementary to the above model variations.

## 1.2 Motivation and Applications

NMT is a widely used technology in both research and commercial settings, ultimately aiming to produce human-level translations. This is increasingly achievable for short sentences, but longer or syntactically complex sentences are more difficult. The hope is that this method of making syntactic predictions over CCG supertags before output word prediction can improve the general handling of word order and agreement for more complex syntactic constructions such as coordination.

In low-resource settings as I experiment in here, target-side syntax can be especially advantageous since source-side syntactic resources are not always available (Nădejde et al., 2017). Over 94% of languages are poorly resourced (Joshi et al., 2020), lacking corpora (especially with gold standard labels) or linguistic expertise (Besacier et al., 2014) as well as competent translation systems, providing strong motivation for the current focus on target-side syntax. Further, I experiment with source-side syntax using Dependency Grammar supertags, and additional monolingual data, to test whether target-side syntax is complementary to other approaches used to improve low-resource NMT, where we want to use all available resources. Finally, the empirical gap of work incorporating unbracketed linearised target-side syntax in Transformer-based NMT provides further motivation.

The central thesis proposed here is that incorporation of target-side syntax in Transformer NMT models improves translation quality. This forms the principal research question, which I test in Chapter 5. Secondarily, I hypothesise that effects from target-side syntax are complementary to and independent of adding monolingual training data and source-side syntax; and third, that improvements are attributable to the method and CCG supertags rather than e.g. decoder depth. I also hope this practical application of syntactic formalisms including CCG will encourage closer collaboration of NLP and theoretical linguistics researchers on common problems.

## 1.3 Contributions

The primary contributions of this work are as follows:

- A novel approach to incorporating target-side syntax in the decoder at word-level by predicting an attention distribution over CCG supertags prior to word prediction, plus a decaying supertag loss.
- A first attempt at incorporating unbracketed, approximately linearised target-side syntax into Transformer-based NMT.
- An empirical evaluation of the proposed method in Turkish→English translation, finding consistent improvements over the syntax-unaware baseline in this low-resource setting.
- Ablation experiments attributing improvements to the proposed architecture and the informativeness of CCG supertags, and showing improvements are complementary to monolingual data.
- An error analysis illustrating considerable adequacy and fluency improvements for syntactic constructions including questions and coordination.

## 1.4 Structure

The structure of this thesis is as follows: Chapter 2 introduces NMT and syntactic representations including CCG, and reviews previous work in target-side syntax-aware NMT. Chapter 3 formalises the proposed method of target syntax incorporation, alongside other baselines. Chapter 4 describes the experimental setup and implementation of

the proposed models, and Chapter 5 presents the evaluation of experiments with target-side syntax, plus an error analysis and discussion of performance on various syntactic constructions. Chapter 6 summarises the findings and contributions, and concludes by offering possible future research directions.

# Chapter 2

# Background and Related Work

In this chapter I review neural machine translation, formal syntactic representations, and previous attempts to incorporate syntax into NMT.

## 2.1 Neural Machine Translation

Machine translation systems automatically translate sentences from a source language to a target language. Earlier statistical phrase-based methods (Zens et al., 2002; Koehn, 2009) have been usurped by neural network-based systems which learn complex relationships from large text databases.

The first neural MT models invoked a novel sequence-to-sequence (S2S) architecture, with separate encoder and decoder recurrent neural networks (RNNs), typically LSTMs (Cho et al., 2014; Sutskever et al., 2014; Jean et al., 2015). The encoder RNN learns a hidden representation of the input which initiates the decoder RNN; this then predicts the output. S2S models learn to predict the probability of the target conditioned on the source in an end-to-end fashion without explicit linguistic instruction. Learning involves updating parameters via mini-batched gradient descent to minimise the negative log-likelihood of the training corpus.

LSTM-based S2S models' performance diminished with increased sentence length due to the encoder's fixed-size hidden state. Bahdanau et al. (2015) solved this with an attention mechanism which calculates a score between the current decoder state and the encoder states, used in a weighted sum of encoder states to produce a context vector which is input into prediction layers. Attention scores are typically calculated using a multi-layer perceptron (Bahdanau et al., 2015) or dot products (Luong et al., 2015).

LSTM-based NMT's restriction to training set vocabularies causes poor handling

of rare and unseen words. Additionally, low-resource settings may not have sufficient data to learn good representations of even common words. To address this Sennrich et al. (2016b) propose training on a sub-word level using byte-pair encoding (BPE) which involves successively merging the most frequently co-occurring characters/sub-words up to the desired vocabulary size. BPE over joint source and target data allows tied embeddings and more consistent segmentation (Sennrich et al., 2016b). Figure 2.1 shows BPE applied to a modified Example 1.1 with shared sub-words, with '–' indicating sub-word splits. I note Turkish's agglutinative morphology (van Schaaik, 2020) means more morphemes per word and more BPE sub-words.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Source | Kim | program– | c– | ının | bunu | kazan– | masını | ister ? |
| Target | Who | wants | the | program– | mer | to | win | it ? |

Figure 2.1: A Turkish→English source-target pair with BPE applied.

The general S2S architecture is versatile. RNNs were initially popular; more recently, Gehring et al. (2017) use convolutional neural networks, and Vaswani et al.'s (2017) non-recurrent Transformer networks now reliably achieve state-of-the-art results (Lakew et al., 2018; Bojar et al., 2018; Barrault et al., 2019). Transformers employ highly parallelised, stacked layers of self-attention, encoder-decoder attention, and feed-forward neural networks, learning enhanced source and target representations. I formalise the baseline Transformer in full in Section 3.1.

NMT systems achieve high translation quality on several metrics and language pairs (Barrault et al., 2020), and learn linguistic information without overt supervision (Conneau et al., 2018; Mareček and Rosa, 2019). However, issues in modelling syntax remain: RNN NMT systems produce less fluent and adequate translations for sentences with coordination, (Shi et al., 2016), PP attachment ambiguity (Bentivogli et al., 2016), relative clauses (Linzen et al., 2016), and negation (Sennrich, 2017). Raganato and Tiedemann (2018) show Transformers struggle to produce grammatically adequate translations in low-resource settings, and Mareček et al.'s (2020) analysis suggests grammatical structure may have little influence on Transformers' language understanding. Modelling word alignment and sentence structure therefore remains a central challenge for NMT (Koehn and Knowles, 2017).

This work attempts to directly incorporate target-side syntax to address some of these syntactic shortcomings. I first explain how syntax may be represented before reviewing attempts to incorporate it into NMT.

## 2.2 Syntactic Representation

Syntax is often incorporated in NMT at the lexical level. I now introduce two formalisms which are conducive to incorporation in NMT systems: Combinatory Categorial Grammar and Dependency Grammar.

### 2.2.1 Combinatory Categorial Grammar

Combinatory Categorial Grammar (CCG) is a lexicalised syntactic formalism consisting of a lexicon of words and their possible lexical categories, and combinatory rules defining how categories combine (Steedman, 1996, 2012). Primitive categories include S (sentence), NP (noun phrase), N (noun), and PP (prepositional phrase); features on S specify sentence types e.g. S[dcl], S[b] and S[to] indicate declarative, bare-infinitival and to-infinitival sentences respectively. Complex categories, e.g. S[wq]/(S[dcl]\NP), are functors describing required argument types, accepting directions, and the resulting type; they can also be arbitrarily nested. Combinatory rules combine categories in the derivation, resulting in a grammatically complete sentence as in Figure 2.2.

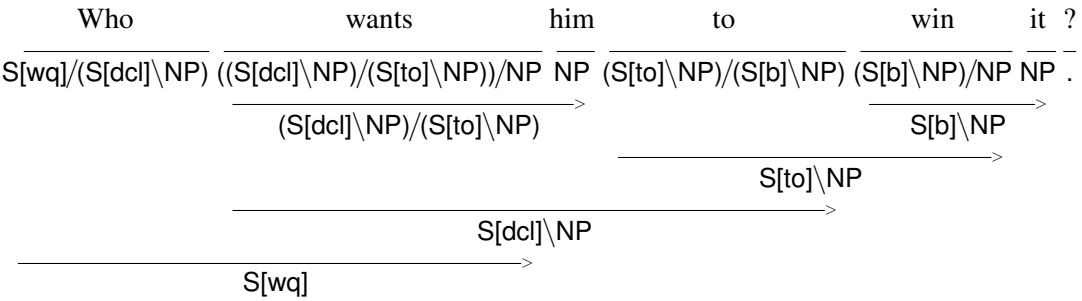


Figure 2.2: CCG derivation for Example 1.1, showing how CCG categories (supertags), combine into complete sentences.

CCG supertags are the terminal categories in the derivation, representing a word's syntactic type, including the presence and order of constraints and dependencies. For example, the supertag for *win*, (S[b]\NP)/NP requires an object NP to the right and a subject NP to the left, resulting in a bare-infinitival clause S[b]. In Figure 2.2, categories combine via forward application (—>); backward application, coordination and type-raising rules also exist. I note CCG's explicit handling of coordination is an advantage over other formalisms including Minimalist Grammar (Chomsky, 1995, 2000).

CCG supertags signal context-sensitive information about local and distant el-

ements in the sentence. Consider Figure 2.2: the tag for declarative verb *wants*, ((S[dcl]\NP)/(S[to]\NP))/NP, signals that the rest of the sentence forms a declarative sentence S[dcl]; it shows an NP argument (*him*) is required immediately rightwards; and it indicates the requirement for a to-infinitival clause S[to]\NP, which itself is dependent on the leftwards *wh*-NP. This illustrates how incremental local supertag prediction facilitates both word prediction (since supertags narrow down possible words) and, crucially, prediction of future arguments and dependencies in the right order.

CCG supertags are so informative that the search problem in supertagging a sentence can be reduced to an exhaustive, deterministic search for the most probable category sequence supporting a CCG derivation, with a simple model then ranking the relatively few possible analyses (Lewis and Steedman, 2014). This is in contrast to part-of-speech tags which, being relatively uninformative, require probabilistic tagging models. The resulting supertags can be easily incorporated into NMT systems either as a full sequence in the encoder or incrementally in the decoder. The compact supertag vocabulary (507) helps NMT models generalise better over supertags than words. In sum, the major advantage of CCG supertags is their high syntactic information density in an efficient lexical representation, without explicit bracketing. Consequently, supertagging is often labelled *almost-parsing* (Bangalore and Joshi, 1999). In this study, I incorporate predictions over CCG supertags in the NMT decoder prior to word prediction.

### 2.2.2 Dependency Grammar

Dependency Grammar (DG) (Tesnière, 1959; Mel'čuk, 1988; Nivre, 2005) is a syntactic formalism representing sentence structure as a labelled graph of binary dependency relations between words. The Universal Dependencies (UD) project (de Marneffe et al., 2014) defines 37 universal head-dependent syntactic relations; Figure 2.3 shows Example 1.1 parsed with dependency relations. Each word has one incoming arc, and the set of arcs and edges implicitly forms a dependency tree. UD delineates three structural types: nominals, clauses, and modifiers. For example, nsubj is a nominal subject dependent, xcomp is a subordinate clausal complement, mark modifies the clausal predicate to indicate the clause type (here, to-infinitival), and root indicates the verbal sentence head.

DG has seen wide usage in NLP (de Marneffe and Nivre, 2019), and it is possible to obtain DG parses for many languages, including otherwise low-resource languages.

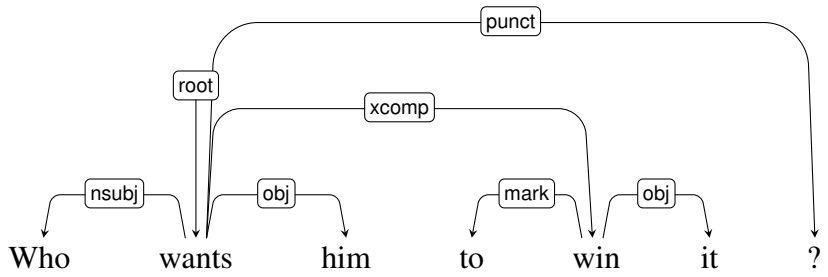


Figure 2.3: Dependency parse of the target-side English sentence from Example 1.1.

I therefore use DG relations in my experiments from the pre-trained statistical Stanza parser (Qi et al., 2020). Unlike CCG supertagging, DG arcs are uninformative thus DG parsing is probabilistic. I use DG parses to construct DG supertags following Ouchi et al. (2014) for comparison with CCG supertags. Their proposed DG supertags include the linear direction of the head and any dependents; I also propose a simple supertag combining the incoming relation and the parent's incoming relation. Table 2.1 summarises the two DG supertag templates; my template DG-A is more fine-grained while Ouchi et al.'s template DG-B has more parallels with CCG supertags.

| Version | Tag form | Example | Vocabulary size | |
|---|---|---|---|---|
| DG-A | label\|parent label | xcomp\|root | TR: 825 | EN: 1632 |
| DG-B | label\|head direction\|dep. direction(s) | xcomp\|L\|L+R | TR: 176 | EN : 362 |

Table 2.1: Dependency supertag templates DG-A (proposed here) and DG-B (Ouchi et al., 2014), with examples for *win* from Figure 2.3, and Turkish and English supertag vocabulary sizes.

While CCG supertags are motivated by their derivational role, DG supertags cannot be used to determine grammaticality and have limited look-ahead capabilities. In Example 1.1, while the CCG supertag for *who*, S[wq]/(S[dcl]/NP), indicates the sentence is a wh-question with a subsequent declarative clause dependency, the equivalent DG supertags, nsubj|root and nsubj|R, only indicate *who* is the subject of the rightwards root verb, with minimal look-ahead information. I test DG supertags on the target side against CCG supertags, and on the source side in combination with target-side CCG supertags, in Chapter 5.

## 2.3 Syntax-aware Neural Machine Translation

NMT's imperfect modelling of syntax motivates approaches to incorporating target-side syntactic information, which I review now. In phrase-based statistical MT, improvements were observed by predicting target-side CCG supertags (Hassan et al., 2007; Birch et al., 2007). Syntax-unaware end-to-end NMT systems soon outperformed these phrase-based models in both fluency and adequacy (Toral and Sánchez-Cartagena, 2017), without any independence assumptions. However, the success of syntax incorporation motivated syntax-aware NMT approaches.

Incremental decoding poses a challenge for target-side syntax incorporation. Ideally we would incorporate a target-side incremental syntactic parser (a syntactic language model), building full syntax trees. We have accurate incremental parsers for constituency grammars (Dyer et al., 2016) and CCG (Stanojević and Steedman, 2020) but incorporation into NMT models is too slow because tree-structures are difficult to mini-batch in GPUs. Akoury et al.'s (2019) Transformer model avoids incremental decoding by predicting a full parse tree then predicting the translation in one-shot conditioned on the tree, speeding up translation. However, all these approaches scale poorly with increased data, motivating alternative target-side syntax approaches, falling into two categories: linearised syntax, either with or without explicit bracketing; and implicit syntax incorporation, which requires architectural modifications for training but uses streamlined models at inference.

Linearisation approaches approximate full syntax incorporation with sequential syntactic information. Previous approaches with explicit bracketing trained recurrent models to translate into linearised constituency parses (Aharoni and Goldberg, 2017) and dependency trees (Le et al., 2017), requiring no decoder modifications and improving over baselines via increased reordering. In Transformer-based NMT, Saunders et al. (2018) observe improvements by ensembling various target-side linearisation strategies, but this requires more complex decoding.

Others incorporate linearised syntax without explicit bracketing, using lexicalised sequential information that tightly couples words and syntax. Nădejde et al. (2017) train a GRU-based S2S model to predict interleaved CCG supertags and words without decoder modifications, observing improvements for different syntactic constructions and outperforming baseline and multi-tasking models. However, Kondratyuk et al.'s (2019) results using interleaved random tags suggest most of this improvement arises from the regularising effects of predicting supertags, plus a deeper decoder, rather than

from useful syntactic generalisation. To my knowledge, no previous work has incorporated unbracketed linearised target-side syntax in Transformer-based NMT. The novel work here with CCG supertags intends to fill this empirical gap, reporting full experiments, including with random tag and interleaved CCG supertag models, in Chapter 5. This method of incorporation lets models freely discern the important syntactic information without rigid constraints, which hurt statistical MT systems (Chiang, 2010) and we expect this to hold for NMT systems.

Finally, implicit target-side syntax incorporation injects a syntactic inductive bias during training. Some propose multi-tasking RNN decoders learning dependency parsing (Kiperwasser and Ballesteros, 2018) or part-of-speech tagging (Niehues and Cho, 2017) alongside translation, both improving on syntax-unaware baselines. The proposed method is somewhat implicit since it does not output linearised parses but implicitly predicts and incorporates a CCG supertag distribution before word prediction. I also experiment with a multi-task learning model to test whether the proposed method outperforms fully implicit methods. Kuncoro et al.'s (2019) implicit approach improves over baselines using knowledge distillation to transfer syntactic knowledge from a syntactic language model (Dyer et al., 2016) to a larger LSTM language model; this addresses syntactic language models' scaling issues, and is potentially applicable to syntax-aware NMT decoders.

Source-side syntax incorporation approaches are similarly varied: some encode full source-side dependency trees using multi-tasking NMT-RNNG systems (Eriguchi et al., 2017) or Graph Convolutional Networks (Bastings et al., 2017); others incorporate linearised syntax with multi-tasking systems learning to encode (Currey and Heafield, 2018) or predict (Luong et al., 2016) source-side linearised parses; and finally, some incorporate source-side syntactic features as embeddings (Sennrich, 2017; Duan et al., 2019), or by interleaving CCG supertags and semantic supersense labels (Vanmassenhove and Way, 2018). I test my method's compatibility with source syntax using interleaved source-side Dependency Grammar supertags.

In addition to incorporating syntax, using backtranslated monolingual data (Sennrich et al., 2016a) can improve NMT, especially in low-resource settings (Bojar et al., 2017). Similarly, Currey et al. (2017) use copied monolingual data, i.e. appending English-English data to Turkish-English parallel data, and observe considerable improvements in translation quality for high and low-resource pairs. I experiment with copied monolingual data in Chapter 5.

Finally, although improvements for incorporating syntax have been observed in RNN-based NMT, tests in Transformer-based NMT are motivated by the different expressive power of Transformer and LSTM S2S models on an empirical (Bhattamishra et al., 2020) and theoretical level (Hahn, 2020).

## 2.4 Summary

In sum, the method proposed here provides a first attempt at incorporating unbracketed, approximately linearised target-side syntax in Transformer-based NMT with an informed architectural design. The principal research question follows naturally from this review, due to the empirical gap. The secondary avenues are also highly motivated, testing several methods, data modifications and syntactic formalisms discussed here.

# Chapter 3

# Target-side Syntax in Transformer-based NMT

This study's central thesis is that incorporating target-side syntax into Transformer-based NMT improves translation quality. I now propose a novel method to integrate target-side CCG supertags into the Transformer decoder. Figure 3.1 summarises the proposed architectural modifications, discussed in detail in Section 3.2.1.

## 3.1 Baseline model

I first formalise the baseline S2S Transformer model (Vaswani et al., 2017). The Transformer is trained on parallel *i.e.* human translated source-target data, learning to translate from source sentence $x$ to target sentence $y$ of length $\mathcal{I}$ by computing the conditional probability $p(y|x)$ as below, where $y_{<i} = y_1...y_{i-1}$, and $\theta$ is the set of model parameters:

$$p(y|x;\theta) = \prod_{i=1}^{\mathcal{I}} p(y_i|y_{<i},x;\theta) \tag{3.1}$$

The encoder contains 6 stacked layers, with multi-headed self-attention and fully connected feed-forward neural network (FFNN) sub-layers, each with residual connections and layer-normalisation. These layers output a hidden state encoding of the source sentence, $H_{enc}^{out}$. Self-attention is position-invariant so input embeddings are augmented with sinusoidal positional encodings, giving embedding matrix $X \in \mathbb{R}^{\mathcal{I} \times d_{model}}$, scaled by $\sqrt{d_{model}}$. Self-attention is computed for $k$ heads, where $Q_{enc}^{k}$, $K_{enc}^{k}$, and $V_{enc}^{k}$ are parametrised linear transformations of $X$ for head $k$. Heads are concatenated, normalised and passed through the FFNN.

$$H^k_{enc} = \text{softmax}(Q^k_{enc} K^{k\top}_{enc}) V^k_{enc} \tag{3.2}$$

$$H_{enc} = [H^1_{enc}; ...; H^k_{enc}] \tag{3.3}$$

$$H'_{enc} = \text{norm}(H_{enc} + X) \tag{3.4}$$

$$H^{out}_{enc} = \text{norm}(\text{FFNN}(H'_{enc}) + H'_{enc}) \tag{3.5}$$

The decoder structure is identical except: it operates incrementally at test time; during training, a mask is applied to multi-head attention to avoid attending to future target embeddings, allowing parallel computation; and a source-target attention sub-layer $Z$ lies between the self-attention and FFNN sub-layers (with transformation parameter matrices $A$, $B$ & $C$):

$$Z_{dec} = \text{norm}([\text{softmax}(H'_{dec} A H^{out\top}_{enc} B^\top) H^{out}_{enc} C]^{[1:k]} + H_{dec}) \tag{3.6}$$

A linear transformation of the decoder hidden state $H^{out}_{dec}$ gives a vocabulary probability distribution $y_{word_i}$ for the $i^{\text{th}}$ word.

$$y_{word_i} = \text{softmax}(H^{out}_{dec} W^{out}) \tag{3.7}$$

The training objective is to minimise the negative log likelihood (i.e. cross-entropy loss) of the generated target sentence $y$ given the source $x$, where $(x, y) \in G$, the group of source-target translations, and $y_i$ is the $i^{\text{th}}$ word of $y$:

$$\mathcal{L}_{word} = -\sum_{i=1}^{I} \log p(y_i | y_{<i}, x; \theta) \tag{3.8}$$

$$\mathcal{L}^G_{word} = \sum_{g=1}^{G} \mathcal{L}^g_{word} \tag{3.9}$$

At inference, output words are predicted using auto-regressive beam search decoding, where $y'$ represents a candidate output.

$$\hat{y} = \underset{y'}{\text{argmax}}\, p(y' | x; \theta) \tag{3.10}$$

This Transformer model forms the baseline against which the proposed method is tested in Chapter 5.

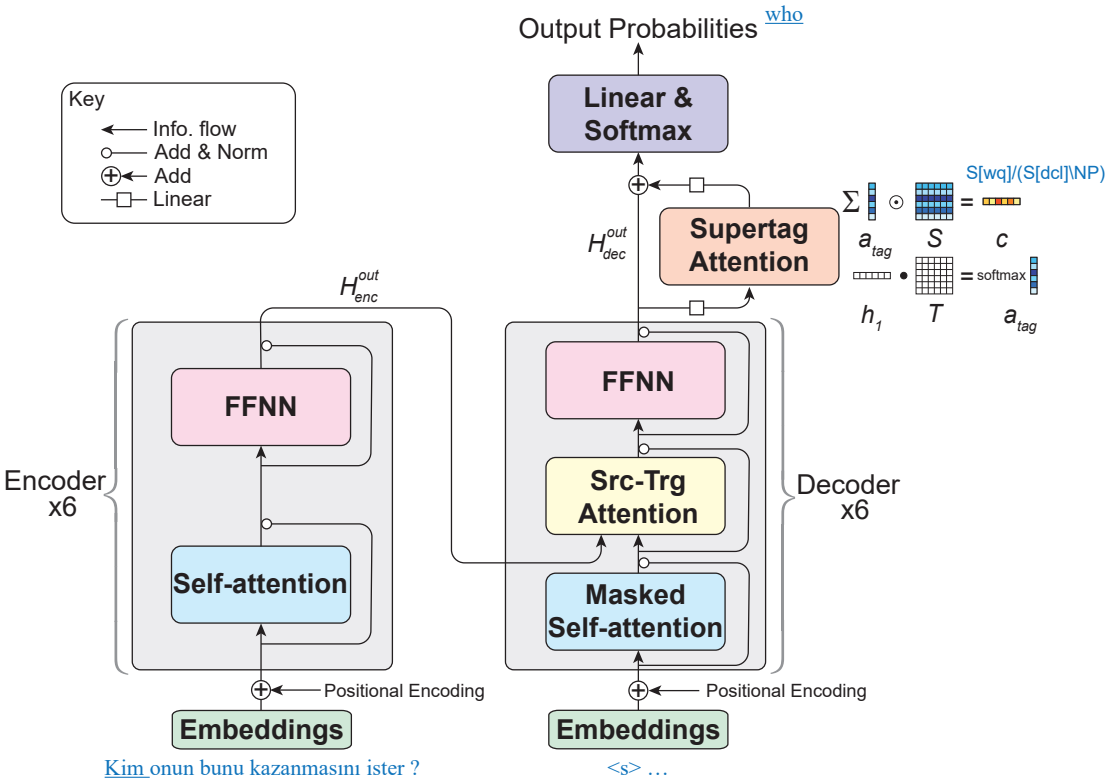


Figure 3.1: Schematic of model integrating target syntax in Transformer decoder; terms are defined in Eq.s 3.11 to 3.17, and Example 1.1 provides the input & output. Adapted from Vaswani et al. (2017). Decoder self-attention is only masked during training.

## 3.2 Implementation

Syntax can be incorporated into Transformer-based NMT in various ways. Here I propose a novel method predicting attention over CCG supertags in the Transformer decoder as a way to tightly and somewhat implicitly incorporate syntax, summarised in Figure 3.1. I also detail a multi-task learning model predicting CCG supertags and translations, and Nădejde et al.'s interleaved tag-word model, both of which form baselines for the proposed method.

### 3.2.1 Target-side Tag Attention

The Tag Attention model requires a small modification to the Transformer decoder. After the 6 stacked layers, the final hidden state representation of the target sentence $H^{out}_{dec}$ is used to predict target words. The key intuition of this method is to use $H^{out}_{dec}$ to predict target-side syntax, by using it to predict an attention distribution over a CCG supertag embedding matrix, then computing a supertag-context vector which is combined with $H^{out}_{dec}$ for word prediction.

First, I transform the decoder hidden state into the tag embedding dimension (512 to 128-dimensions), where $W_1 \in \mathbb{R}^{d_{model} \times d_{tag}}$. Attention scores $\alpha_{tag}$ are computed between $h_1 \in \mathbb{R}^{1 \times d_{tag}}$ and tag embedding matrix $T \in \mathbb{R}^{v_{tag} \times d_{tag}}$ using dot product attention (with $v_{tag}$ the supertag vocabulary size), followed by a softmax to give attention weights $a_{tag} \in \mathbb{R}^{v_{tag} \times 1}$. These weights scale each embedding vector in a second learnable tag embedding matrix $S \in \mathbb{R}^{v_{tag} \times d_{tag}}$. The resulting weighted embeddings are summed element-wise along the embedding dimension, producing a supertag-context vector $c \in \mathbb{R}^{1 \times d_{tag}}$, capturing information about relevant CCG supertags:

$$h_1 = W_1 H^{out}_{dec} \tag{3.11}$$

$$\alpha_{tag} = h_1 \cdot T^{\top} \tag{3.12}$$

$$a_{tag} = \text{softmax}(\alpha_{tag}) \tag{3.13}$$

$$c = \sum a_{tag} \odot S \tag{3.14}$$

The supertag-context vector is then transformed by $W_2 \in \mathbb{R}^{d_{tag} \times d_{model}}$ back to dimension $d_{model}$ and summed to the original $H^{out}_{dec}$ hidden state for word prediction:

$$h_{tag} = W_2 c \tag{3.15}$$

$$H^{out'}_{dec} = H^{out}_{dec} + h_{tag} \tag{3.16}$$

$$y_{word_i} = \text{softmax}(H^{out'}_{dec} W^{out}) \tag{3.17}$$

Crucially, this model has a composite loss function; in addition to word loss, the model minimises the loss of the predicted attention weights (i.e. output probabilities) over the tag vocabulary. In essence, the transformation and dot product in Eq.s 3.11 and 3.12 predict a distribution over tags $a_{tag}$ for the current word (used as an attention score to create a supertag-context vector), with loss backpropagated through both matrices $T$ and $W_1$. The loss for a sequence of tags $t$ is calculated as in Eq. 3.18 for $(x, y, t) \in G$, the set of source, target and tag sequences:

$$\mathcal{L}_{tag} = -\sum_{i=1}^{I} \log p(t_i | t_{<i}, y_{<i}, x; \theta) \tag{3.18}$$

$$\mathcal{L}^G_{tag} = \sum_{g=1}^{G} \mathcal{L}^g_{tag} \tag{3.19}$$

$$\mathcal{L}^G = \mathcal{L}^G_{word} + \mathcal{L}^G_{tag} \tag{3.20}$$

Consider Example 1.1 and Figure 3.1. First, the 6 encoder layers take the complete Turkish source sentence as input, returning $H^{out}_{enc}$; then the start-of-sentence token

(<s>) embedding is passed into the 6 decoder layers, which also attend to $H_{enc}^{out}$. The decoder output hidden state $H_{dec}^{out}$ is used to predict a distribution for the first tag, summarised as $h_{tag}$ and summed to $H_{dec}^{out'}$, which is then used to predict a distribution for the first English word. Correctly predicting the tag S[wq]/(S[dcl]\NP) will help to predict both the *wh*-word *who* and the following declarative verb *wants*. Loss for the first step is the sum of the predicted negative log probabilities of the first reference tag S[wq]/(S[dcl]\NP) and word *who*. In training, this process occurs in parallel, while at test time it repeats incrementally until the model predicts an end-of-sentence token (</s>) or reaches its maximum output length.

While training initial models, I observed that tag attention initially improves validation scores but becomes less advantageous and even inhibitory towards convergence. I therefore adapted the original loss function to include a weighted, decaying tag loss. The decay factor $\mathcal{D}$ was determined empirically for the specific models built (i.e. by validation BLEU scores), and is defined as:

$$\mathcal{D}_{\mathcal{E}} = 0.65^{\mathcal{E}-1} \tag{3.21}$$

$$\mathcal{D} = \mathcal{D}_{\mathcal{E}} \quad \text{if} \quad \mathcal{D}_{\mathcal{E}} > 0.1, \quad \text{else} \quad \mathcal{D} = 0 \tag{3.22}$$

where $\mathcal{E}$ is the training epoch number such that at epoch 1, the decay factor is $0.65^0 = 1$, at epoch 2, $\mathcal{D} = 0.65^1$, and so on. When the decay factor drops below 0.1 at epoch 8, it is set to 0, meaning tag loss is no longer used in parameter optimisation. The total training corpus loss $\mathcal{L}^G$ is then defined as:

$$\mathcal{L}^G = \mathcal{L}^G_{word} + \mathcal{D}\mathcal{L}^G_{tag} \tag{3.23}$$

This method is tag agnostic; I test this method in Chapter 5 with CCG supertags, plus DG supertags and randomly generated target-side tags, against a baseline Transformer, to investigate this work's hypotheses.

#### 3.2.1.1 Design Choices

The proposed method's design is motivated by three advantages over Nădejde et al.'s method:

- Soft supertag decisions: While Nădejde et al. choose 1 supertag, predicting an attention distribution permits incorporating information about any number of likely supertags in the context vector, without imposing rigid syntactic constraints.

- Decaying loss: This novel modification decreases the importance of tag loss over time, shifting focus to the harder word prediction task after initially injecting an inductive bias.
- Search: In principle, soft decisions mean there is no competition during beam search between different derivations, a potential issue of the interleaving approach.

### 3.2.2 Multi-task model

I implement a multi-task learning (MTL) architecture, to test whether fully implicit supertag incorporation is as effective for translation as direct incorporation into word prediction. This architecture draws inspiration from Nădejde et al.'s (2017) MTL model, except it replaces the GRU with a Transformer, and is more parameter efficient because the decoder layers are shared, using the final hidden state for both tag and word prediction concurrently.

The MTL model uses a simple FFNN to predict CCG supertags from $H_{dec}^{out}$:

$$h_m = \tanh(W_1 H_{dec}^{out} + b_1) \tag{3.24}$$

$$y_{tag_i} = \text{softmax}(W_2 h_m) \tag{3.25}$$

where $W_1 \in \mathbb{R}^{d_{model} \times d_{h_m}}$ ($d_{h_m} = 256$, $d_{model} = 512$), and $W_2 \in \mathbb{R}^{d_{h_m} \times v_{tag}}$. The FFNN projects the decoder output to a hidden layer, applies a non-linearity, then transforms it into a probability distribution over the tag vocabulary, with which loss is computed. The loss for one sentence, with tag loss decay, is calculated as:

$$\mathcal{L}_{word} = -\sum_{i=1}^{\mathcal{I}} \log p(y_i | y_{<i}, t_{<i}, x; \theta) \tag{3.26}$$

$$\mathcal{L}_{tag} = -\sum_{i=1}^{\mathcal{I}} \log p(t_i | y_{<i}, t_{<i}, x; \theta) \tag{3.27}$$

$$\mathcal{L} = \mathcal{L}_{word} + \mathcal{D}\mathcal{L}_{tag} \tag{3.28}$$

### 3.2.3 Interleaved Model

Finally, I replicate the target-side interleaving model from Nădejde et al. (2017), with interleaved CCG supertags and target words. I use the baseline Transformer, with a shared word and supertag embedding space and vocabulary, and postprocessing to

remove tags after computing loss and before calculating BLEU score. The Transformer decoder is unchanged, and the probability of the target tag-word sequence $y'$ is now:

$$y' = y_{tag_1}, y_{word_1}, ..., y_{tag_I}, y_{word_I} \tag{3.29}$$

$$p_{y'} = \prod_{i=1}^{2I} p(y'_i | y'_{<i}, x; \theta) \tag{3.30}$$

In later experiments, I also interleave DG supertags (detailed in Section 2.2.2) with source words as a simple method for incorporating source-side syntactic information. Source-side interleaving leaves the encoder unchanged but doubles the source length.

# Chapter 4

# Experimental Setup

I now describe the experimental conditions and analysis strategy for investigating the principal research question.

## 4.1 Data and Pre-processing

I train all models on publicly available WMT parallel (and monolingual) training data (Bojar et al., 2017). For Turkish→English, I use `newsdev2016` and `newstest2017` for validation and test sets, and for monolingual data, a lack of available backtranslated data led to using copied English data (English on both source and target sides), randomly sampled from `newscrawl2016`. As per Currey et al. (2017), I use a ratio of 1:2 parallel:copied data, appending the 413,250 copied sentences to the source and target training sets for later experiments. I tokenise and process the data with the standard Moses scripts (Koehn et al., 2007), then shuffled before training.

Turkish is a morphologically rich language with context-dependent word order subject to local scrambling, most commonly having SOV order (Hoffman, 1995), and English is morphologically sparse with SVO word order (Greenbaum, 1996). Turkish to English translation therefore offers considerable syntactic challenges which the current method attempts to address.

I use the pre-trained EasyCCG model (Lewis and Steedman, 2014) to annotate the target-side English training data with lexical CCG supertags, detailed in Section 2.2.1. I removed sentences from the training set which EasyCCG failed to parse, either due to being longer than 70 words or being partially formed sentences. This filtering was minimal, reducing the datasets by less than 1%. Table 4.1 shows sentence counts for the finalised datasets. The final training setup for tag attention used parallel Turkish,

English and CCG supertag datasets (as in Example 1.1), with one tag per English word and duplicated tags for BPE subwords.

| | Train | Dev | Test |
|---|---|---|---|
| TR-EN | 206,625 | 3,000 | 3,007 |
| EN-mono | 413,250 | - | - |

Table 4.1: Sentence counts for the various different subsets of the parallel Turkish→English and monolingual English data.

After filtering, I used the Stanza toolkit (Qi et al., 2020) to generate DG parses for source and target-side data. In both cases, I use two DG supertag templates as in Table 2.1, interleaving supertags for source-side experiments. I also generate target-side random tags with the same vocabulary size as CCG supertags, for a baseline which removes any syntactic advantage.

## 4.2 Implementation and Training

The NMT systems used are Transformer networks (Vaswani et al., 2017) from the JoeyNMT toolkit (Kreutzer et al., 2019). The models' hyper-parameters resemble those of Hieber et al. (2018) for their optimal WMT17 English→German Transformer model; I report the full parameter configuration in Appendix A. Modifications to the Transformer decoder, as described in Section 3.2.1, were implemented directly in JoeyNMT's Transformer decoder. Due to time constraints, I use the first single models trained and do not use ensembles.

Models are evaluated with BLEU score (Papineni et al., 2002), an automatic metric calculating document-level n-gram precision for NMT hypotheses against reference translations, where higher scores are a proxy for improved translation fluency and adequacy; it is computed using `multi-bleu.perl` (Koehn et al., 2007) over the tokenised validation set during training, and on the test set for the final evaluation. BLEU score also forms the early stopping metric and controls learning rate decreases. The MTL model validation only uses translations rather than tag outputs. I used greedy 1-best search for validation, and beam search ($n = 5$, with length penalty $\alpha = 1$) at test time.

I segment words into BPE sub-word units (Sennrich et al., 2016b) learned over both the source and target languages, with 85,000 merge operations (as per Nădejde et al.). The CCG supertag vocabulary size was 511 (including <s> and </s>, a padding token

and an UNK token), over which the tag attention distribution is predicted.

I use 512-dimensional tied source-target word embeddings and output representations ($\frac{1}{4}$ of the FFNN layer dimension). Supertag embeddings are 128-dimensional, with embedding matrix $T$ of size $511 \times 128$. Lower dimensional supertag embeddings are suitable due to their smaller vocabulary, implying less complex relationships that need to be modelled. I use a maximum input/output length of 70 tokens, and 140 tokens for interleaving setups.

To analyse the CCG supertag attention model's performance, I evaluate the model via BLEU score of WMT test set predictions, in a low-resource Turkish→English setting, against various baselines: a baseline NMT system, a multi-task learning model, target-side DG supertag and random tag attention models, an interleaving model, models with monolingual data, and source syntax NMT models. These baselines will help determine the precise source of improvements in translation quality.

# Chapter 5

# Results and Analysis

In this section, I evaluate the proposed target-side tag attention model against a baseline Transformer NMT model. Regarding the principal research question, I show that the proposed tag attention method improves Turkish→English translation quality. I eliminate other explanations for these improvements, and show improvements are independent of source syntax and complementary to monolingual data.

## 5.1 Target-side Syntax Experiments

To reiterate, this study's central thesis is that target-side supertag attention improves Transformer-based NMT. To test this, I experimented on Turkish→English WMT data with the tag attention NMT model (TA-NMT) against a baseline model. Table 5.1 presents the main results.

The test set results show a small improvement of around 0.1 BLEU for the tag attention model's output over the baseline model's output; this is consistent with the central hypothesis, and suggests prediction of CCG supertags helps the model produce more fluent translations. Later experiments show how this improvement can be increased.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| Tag Attention | 15.77 | 8.80 |
| + Loss Decay | 15.90 | **9.11** |

Table 5.1: Target-side syntax experiments for Turkish→English translation, reporting BLEU scores for baseline NMT, TA-NMT, and TA-NMT with loss decay. Dev is newsdev2016, Test is newstest2017; highest score in bold.

**Loss decay —** During training I noticed the tag attention model's validation BLEU scores are initially higher than the baseline, but improvements tail off. I therefore implemented a decaying tag attention loss to maintain the initial benefits without restricting later performance. This resulted in a further 0.3 BLEU score improvement, shown in Table 5.1. Decaying tag loss likely improves performance because learning CCG supertags is more constrained (507 tags vs 85,000 sub-words), so the model rapidly learns to generalise well over supertags then once proficient it can focus on the harder word prediction task.

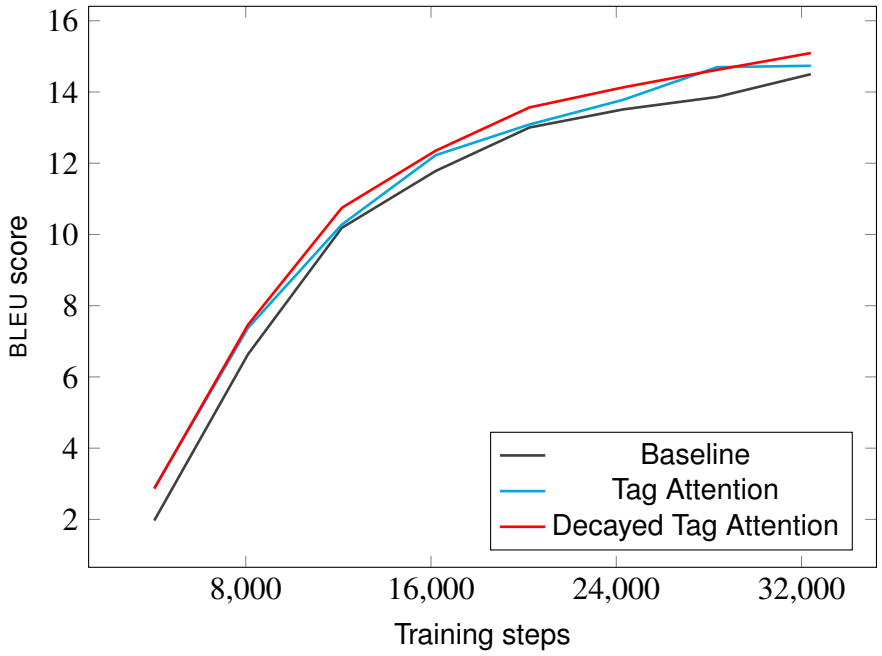


Figure 5.1: Validation BLEU score curves for a baseline NMT system, TA-NMT and Decayed TA-NMT systems.

Figure 5.1 demonstrates the decaying supertag loss' advantage. While the TA-NMT model outperforms the baseline for most of training, the baseline nearly catches up by continuing to improve past 28,000 steps, likely because the tag loss becomes restrictive and detracts from word prediction. The decaying loss TA-NMT model combines both models' advantages: it initially learns faster using the supertag loss, then later uses the word loss to maintain this advantage and continue learning consistently like the baseline, unrestricted by the tag loss. To my knowledge, this is a novel finding.

**Random Tags —** This ablation experiment tests whether the improvements stem from CCG supertags' utility or from a deeper decoder with more parameters, as Kondratyuk et al. (2019) suggest about target-side syntax. I test a decaying TA-NMT model

with randomly generated tags against a baseline system. The results in Table 5.2 show random tags marginally improve over the baseline, but do not improve on the CCG TA-NMT model. This suggests the overwhelming majority of the observed improvements from tag attention do not result simply from a deeper decoder or regularising effects of predicting tags, running *contra* Kondratyuk et al.'s findings, and supporting the tertiary hypothesis. The improvements must therefore be attributed to the dense syntactic information conveyed by CCG supertags.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| Tag Attention + Decay | 15.90 | **9.11** |
| + Random Tags | 15.13 | 8.79 |

Table 5.2: Turkish→English experiments with random tags, reporting BLEU scores for baseline NMT, TA-NMT with CCG supertags and with random tags.

**DG Supertags —** I also test whether improvements stem from the syntactic utility of CCG supertags specifically, using Dependency Grammar supertags (described in Section 2.2.2) as the target-side syntactic information. The results, shown in Table 5.3, show a decrease of approximately 0.2 BLEU for target-side DG supertags against the baseline. This suggests the improvements in Table 5.1 stem specifically from CCG supertags' dense syntactic information, cementing them as the syntactic labels of choice in the tag attention architecture, and supporting the tertiary hypothesis.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| Tag Attention + Decay | 15.90 | **9.11** |
| + DG-A supertags | 15.21 | 8.51 |
| + DG-B supertags | 15.17 | 8.59 |

Table 5.3: Target-side syntax experiments for Turkish→English translation, reporting BLEU scores for baseline NMT, and TA-NMT with CCG, DG-A and DG-B supertags.

It is likely DG supertags are not sufficiently informative about linear word order to improve target-side prediction; additionally, DG-A tags' increased vocabulary (1632

vs 507 CCG supertags) is likely to be a distractor, explaining their marginally worse performance. As an example, DG-A main verb supertags are, uninformatively, always root, while verbal CCG supertags are dense, indicating the sentence type plus the type and order of dependencies. CCG supertags' derivational role demonstrates their inherent utility; conversely, DG supertags are constructed arbitrarily, only signalling hierarchical relationships rather than sentence-level constraints.

**Multi-task learning —** I now test whether improvements are replicated by the MTL model from Section 3.2.2, which has a composite loss but discards tag predictions at test time. The results, shown in Table 5.4, show that the multi-task model without decay performs considerably worse than the baseline, while the model with decay reaches parity. This suggests first that TA-NMT improvements result from tightly incorporating syntax via a supertag-context vector, rather than just predicting CCG supertags, supporting the tertiary hypothesis. This may be because task determination detracts from performance, as Macháček (2018) suggests about MTL NMT. Second, the decayed MTL's 0.7 BLEU improvement further illustrates the restrictiveness of tag loss and the importance of decay.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| Tag Attention + Decay | 15.90 | **9.11** |
| Multi-task | 14.93 | 8.03 |
| + Decay | 15.50 | 8.73 |

Table 5.4: Turkish→English target-side syntax experiments, reporting BLEU scores for multi-task learning NMT model against a baseline and TA-NMT systems.

**Interleaving tags —** I replicate Nădejde et al.'s target-side interleaving model, using modified data with a baseline system. The results in Table 5.5 show improvements from interleaving are substantial (around 0.4 BLEU), but only marginally different to the decaying TA-NMT model's performance. The similarity is likely because both models employ a tight coupling of words and syntax; further tests on larger datasets and other language pairs may help differentiate their performance, and I might expect that the decaying TA-NMT model's tuneable, differential tag and word loss may prove advantageous in other settings. This experiment confirms tight, sequential incorporation of syntax as an effective general strategy to improve NMT.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| Tag Attention + Decay | 15.90 | 9.11 |
| Interleaved tags | 14.71 | 9.16 |

Table 5.5: Target-side syntax experiments for Turkish→English translation, reporting BLEU scores for an interleaved CCG tag NMT system against a baseline and TA-NMT.

**Monolingual data —** Next, I test whether the observed improvements persist through and complement adding monolingual data, a common extension to improve low-resource NMT. I use copied (English→English) data, which is only marginally less effective than backtranslated data due to increased encoder noise (Currey, 2019). A shared source-target BPE vocabulary permits English and Turkish source input without issue. I test the decayed TA-NMT model with 413,250 additional monolingual sentences against the baseline with monolingual data, presenting results in Table 5.6.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| + Monolingual | 16.20 | 9.99 |
| Tag Attention + Decay | 15.90 | 9.11 |
| + Monolingual | 17.01 | **10.68** |

Table 5.6: Experiments with target-side syntax for Turkish→English with monolingual copied data, reporting BLEU scores for baseline and TA-NMT models.

The results indicate around 0.7 BLEU improvement between the two monolingual data models. First, this suggests the TA-NMT method improves translation quality independently of whether monolingual data is added, showing promising signs regarding scalability issues affecting previous syntactic approaches (Kuncoro et al., 2019). Further, the BLEU score improvements are nearly 2× greater than improvements for decayed tag attention alone, suggesting TA-NMT is complementary to additional monolingual data; both findings support the secondary hypothesis. The increased improvements suggest that, in addition to letting the model learn a better target-side language model (outweighing the source-side noise), monolingual data lets the model learn more syntax through CCG supertags, further improving the output quality.

**Source syntax —** In addition, I test whether the same improvements can be achieved by the addition of source syntax. The method was necessarily limited by the available source-side (Turkish) syntactic resources, therefore I used two versions of Dependency Grammar supertags (see Section 2.2.2), interleaved with the source data, for the baseline and decaying tag attention models. Table 5.7 presents the results.

| | TR→EN | |
|---|---|---|
| Model | Dev | Test |
| Baseline | 15.55 | 8.73 |
| + Source Syntax A | 15.07 | 8.54 |
| + Source Syntax B | 14.94 | 8.53 |
| Tag Attention + Decay | 15.90 | **9.11** |
| + Source Syntax A | 15.45 | 9.01 |
| + Source Syntax B | 15.54 | 8.89 |

Table 5.7: Source and target-side syntax experiments for Turkish→English translation, reporting BLEU scores for baseline and TA-NMT models, with DG-A and DG-B supertags.

These results illustrate that no matter how we model the source (with DG-A or DG-B supertags, or no syntax), using target-side syntax consistently improves performance by approximately 0.4 BLEU, supporting part of the secondary hypothesis. Additionally, source-side dependency syntax has no beneficial effect in this setting. While intended to improve translation adequacy through better source understanding, it appears DG supertags are not as informative as CCG supertags (supported by results in Table 5.3), distracting the model and resulting in less useful source representations. The increased interleaved sentence lengths may also degrade translation quality. It therefore remains to be seen what kind of source syntax would help alongside TA-NMT; in future work I intend to test less *ad hoc*, more sophisticated methods such as linearised source parses (Currey and Heafield, 2018) and source-side CCG supertags.

**Overview —** In response to the principal research question, the chief conclusion from these Turkish→English experiments is that the proposed method of predicting attention over CCG supertags with decaying loss improves output translation quality. These improvements from the TA-NMT model are independent of those from decoder depth, DG supertags, multi-tasking, monolingual data or source syntax, and are complementary to adding monolingual data, with source-side syntax results proving incon-

clusive on this point. The results thus strongly support this work's three hypotheses.

Further extensions are possible that would provide more conclusions but were limited by time. One main extension to consider is whether hard supertag decisions, instead of the soft decisions in TA-NMT, produce better predictions. This could be done in a supervised setting by explicitly predicting CCG supertags, or in an unsupervised setting using Gumbel-softmax (Jang et al., 2017).

## 5.2 Error Analysis

I now conduct an analysis of translation performance on different syntactic constructions, in terms of relative document-level BLEU scores, for the decaying TA-NMT model against the baseline, both with monolingual data. Sentences with different syntactic constructions are divided by CCG supertags as per Nădejde (2018): questions contain S[q], S[wq] or S[qem]; conj signals conjunctions; prepositional phrases (PPs) are indicated by PP, ((S\NP)/(S\NP))/NP, or (NP\NP)/NP categories; tags with S[to] dependencies indicate control/raising clauses; and relative/subordinate clauses are indicated by complex categories including (NP\NP)/(S/NP).

| | PP | Conj. | Rel./Sub. | Control | Questions | Total |
|---|---|---|---|---|---|---|
| TR→EN | 2175 | 1221 | 843 | 479 | 112 | 3007 |

Table 5.8: Frequencies of sentences with different syntactic constructions in English *newstest2017* set.

Table 5.8 shows construction frequencies in the test set; sentences can contain multiple construction types, but if multiple instances of one construction occur in one sentence, it is only counted once. Successfully translating these phenomena from Turkish to English requires proficient handling of both long-distance agreement and word order.

The results over the subsets in Figure 5.2 demonstrate a largely consistent improvement over the different sentence types, control sentences notwithstanding. The largest improvements are seen for questions (1.44 BLEU) and conjunctions (1.00 BLEU), which makes sense since CCG makes these constructions explicit via S[q/wq/qem] and conj categories. I note the question subset's small size warrants further tests with a larger question set. Overall, BLEU improvements for questions, conjunctions, prepositional and relative/subordinate clauses therefore indicate predicting informative CCG su-

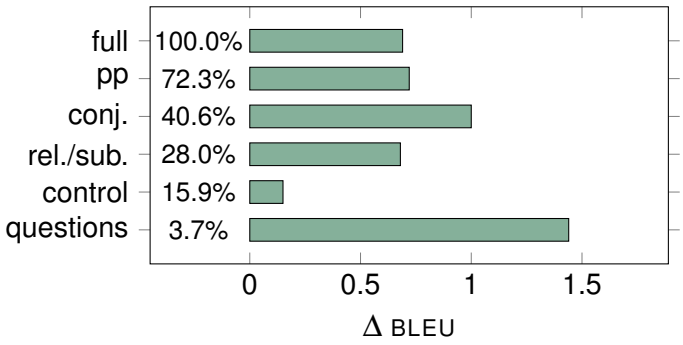


Figure 5.2: Improvements in BLEU score for syntactic construction test sets for the TA-NMT model with decay and monolingual data against a baseline system with monolingual data. Values left of bars indicate the percentage of the full test set that each subset constitutes.

pertags facilitates better handling of agreement and word order (especially re-orderings) from Turkish to English.

Figure 5.2 shows a marginal improvement on the control/raising subset (around 0.1 BLEU). Control constructions usually involve long-distance agreement (Nădejde, 2018); the results therefore suggest CCG supertags may not carry enough information to usefully model the agreement and dependencies in these complex constructions. CCG supertags do generally help with long-distance agreement however, demonstrated by improvements for conjunctions.

I note the current subset divisions are coarse-grained and have relatively few sentences; an in-depth comparison using larger subsets and finer-grained divisions, e.g. via constituency parses, may prove enlightening. In sum, this analysis demonstrates how CCG supertags improve the general handling of word order and agreement in translation, reinforcing the central thesis. I now discuss examples illustrating these improvements.

## 5.3 Discussion

The experiments illustrate the proposed method of predicting attention over CCG supertags improves Turkish→English translation quality. Here I discuss the explicit impact of CCG supertags on the handling of word order and agreement, through two test set examples in Table 5.9 for which the tag attention model predicts more grammatical translations.

Example I shows a more adequate and fluent translation for a subject *wh*-question

| I | **TR→EN — Question & PP** |
|---|---|
| Source | Bunların sorumlusu kimdir ? |
| Target | Who is responsible for them ? |
| Trg-CCG | S[wq]/(S[dcl]\NP) (S[dcl]\NP)/(S[adj]\NP) (S[adj]\NP)/PP PP/NP NP . |
| NMT | Is they responsible ? |
| TA-NMT | **Who** is responsible **for them** ? |

| II | **TR→EN — Coordination & Subordinate** |
|---|---|
| Source | Kontrol istasyonumuz TRT yayınlarının geri geldiğini ve bildiri okunmaya başladığını söyledi . |
| Target | Our control station said that TRT broadcasts came back on-air and the announcement started to be read . |
| Trg-CCG | NP/N N/N N (S[dcl]\NP)/S[em] S[em]/S[dcl] N/N N (S[dcl]\NP)/(S[adj]\NP) (S\NP)\(S\NP) S[adj]\NP conj NP/N N (S[dcl]\NP)/(S[to]\NP) (S[to]\NP)/(S[b]\NP) (S[b]\NP)/(S[pss]\NP) S[pss]\NP . |
| NMT | our unions have come back and say the statement is beginning being read . |
| TA-NMT | our **Control stations say that TRT publications** have come back and the statement **started to be read** . |

Table 5.9: Examples comparing baseline (NMT) and tag attention NMT (TA-NMT) system output for Turkish→English for a question (I) and coordination (II). Hypothesised targets are shown with source, target and CCG tag references. TA-NMT improvements against NMT predictions are in bold.

with a prepositional phrase. The TA-NMT system correctly predicts the *wh*-question, whereas the baseline predicts a yes-no question. Correctly predicting the S[wq] category would help the TA-NMT system predict an initial *wh*-word. Next, the TA-NMT system translates the sentence structure correctly by predicting the copula supertag (S[dcl]\NP)/(S[adj]\NP), indicating a declarative verb, a following adjective and agreement with the NP subject. The syntax-unaware baseline fails to reconcile the third-person plural pronoun and singular copula, predicting incorrect agreement and argument structure. Third, the baseline fails to translate the final PP *for them*; if the TA-NMT system predicts (S[adj]\NP)/PP for *responsible*, it will know to predict a PP, whose tag indicates a constituent NP. This also demonstrates proficient reordering of the Turkish sentence-initial PP to sentence-final. The TA-NMT system's CCG supertag prediction thus improves agreement, word order and argument handling in Example I.

For Example II, the TA-NMT correctly predicts the coordination structure inside a subordinate/embedded S[em] clause, likely because it predicts the S[em] dependency on *said*'s supertag, while the baseline predicts an embedded clause inside coordina-

tion. The TA-NMT system also correctly translates the full subject and object phrases by predicting multiple N dependencies. Finally, the TA-NMT system correctly translates the past tense to-infinitival final clause rather than two S[ng] verbs as the baseline predicts, likely by predicting the (S[dcl]\NP)/(S[to]\NP) category and dependency correctly prior to predicting *to*. This demonstrates how the proposed method vastly improves overall sentence structure for various syntactic constructions by predicting future word order and agreement dependencies.

This work has important implications. This discussion emphatically reinforces the error analysis, showing increased BLEU scores stem from improved prediction of future word order and agreement dependencies for subordinate, interrogative, coordination and prepositional constructions; the increased fluency and adequacy is accredited to CCG supertag prediction. Regarding the principal research question, the full results strongly support the central hypothesis. The secondary and tertiary hypotheses are generally supported, since target-side supertag attention is shown to be complementary to monolingual data, with improvements attributable solely to supertag attention. The main implication is that Transformer-based NMT should attempt to incorporate target-side syntax using this method, which proves beneficial for low-resource, syntactically divergent languages. This reinforces previous results for RNN-based NMT (Nădejde et al., 2017). Additionally, I underline for other work CCG supertags' utility as lexicalised chunks of syntactic information, thanks to their derivational role.

This study's main strength is the range of experiments focusing attribution of improvements to the specific CCG supertag attention method. An inherent drawback of this method is the lack of explicit interpretability of supertag predictions; a deeper inspection is left to future work. Finally, time constraints limited experiments with high-resource language pairs, which are also left to future work, with further recommendations offered in Section 6.2.

# Chapter 6

# Conclusion

## 6.1 Major Findings

In this work I proposed a method to directly incorporate target-side syntax in Transformer-based NMT by predicting attention over CCG supertags. The results in Chapter 5 affirm the central thesis and show that the method improves translation quality over baselines in a low-resource setting. Improvements are independent of source syntax and additional monolingual data, and complementary to the latter, largely supporting the secondary hypothesis. The model is only matched by an interleaving model and outperforms MTL NMT, and TA-NMT models with DG or random tags, supporting the tertiary hypothesis. The TA-NMT model with decaying loss and monolingual data produces the greatest improvement of 0.7 BLEU. Finally, the model shows significant improvements for sentences with questions, coordination, subordinate clauses and PPs, via improved reordering and agreement handling.

In sum, it is the specific combination of the dense categorial information in CCG supertags and the proposed tag attention method that provides the decayed CCG TA-NMT model with a tangible advantage over syntax-unaware NMT systems.

## 6.2 Future Work

This study begets plentiful avenues for future work. First, experiments with other language pairs, including high-resource source-side languages, are necessary to test consistency and scaling; of special interest is translation *into* morphologically underspecified languages, e.g. Mandarin Chinese, for which supertags are more informative than English given increased ambiguity, perhaps leading to greater improvements.

This would also permit testing source-side CCG supertags in Transformer-based NMT. Second, a deeper manual analysis of syntactic performance, especially given the large improvements for questions, could further elucidate this method's advantages. A third possible avenue could involve replacing tag loss decay with a reactivatable loss threshold to keep tag predictions in check. Finally, I intend to test this method with Minimalist Grammar supertags (Chomsky, 1995; Torr et al., 2019) to implement the first practical application of Chomsky's Minimalist Program in NLP.

# Bibliography

Roee Aharoni and Yoav Goldberg. Towards string-to-tree neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–140, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-2021.

Nader Akoury, Kalpesh Krishna, and Mohit Iyyer. 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, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1122.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108:331–342, 2017. URL https://ufal.mff.cuni.cz/pbml/108/art-ataman-negri-turchi-federico.pdf.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015. URL https://arxiv.org/abs/1409.0473.

Srinivas Bangalore and Aravind K. Joshi. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265, 1999. URL https://aclanthology.org/J99-2004.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi,

Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-5301.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.1.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/D17-1209.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclanthology.org/D16-1025.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100, 2014. URL https://www.sciencedirect.com/science/article/pii/S0167639313000988.

Satwik Bhattamishra, Arkil Patel, and Navin Goyal. On the computational power of transformers and its implications in sequence modeling. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 455–475, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.conll-1.37.

Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0702.

Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-4755.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-6401.

David Chiang. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1146.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. URL https://aclanthology.org/D14-1179.

Noam Chomsky. *The Minimalist Program*. MIT Press, Cambridge, MA, 1995.

Noam Chomsky. Minimalist inquiries: The framework. *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, pages 89–155, 2000.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the*

*Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-1198.

Anna Currey. *Moving beyond parallel data for neural machine translation*. PhD thesis, The University of Edinburgh, 2019.

Anna Currey and Kenneth Heafield. Multi-source syntactic neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2961–2966, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1327.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-4715.

Marie-Catherine de Marneffe and Joakim Nivre. Dependency grammar. *Annual Review of Linguistics*, 5(1):197–218, 2019. URL https://doi.org/10.1146/annurev-linguistics-011718-011842.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.

Sufeng Duan, Hai Zhao, Junru Zhou, and Rui Wang. Syntax-aware transformer encoder for neural machine translation. In *2019 International Conference on Asian Language Processing (IALP)*, pages 396–401. IEEE, 2019. URL http://www.colips.org/conferences/ialp2019/ialp2019.com/files/papers/IALP2019_101.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North*

*American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/N16-1024.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-2012.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-1012.

Sidney Greenbaum. *The Oxford English Grammar*. Oxford University Press, Oxford, 1996.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. URL https://aclanthology.org/2020.tacl-1.11.

Hany Hassan, Khalil Sima'an, and Andy Way. 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, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1037.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 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 Track)*, pages 200–207, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://aclanthology.org/W18-1820.

Beryl Ann Hoffman. *The computational analysis of the syntax and interpretation of "free" word order in Turkish*. PhD thesis, University of Pennsylvania, 1995.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint*, 2017. URL https://arxiv.org/abs/1611.01144.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-3014.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.560.

Eliyahu Kiperwasser and Miguel Ballesteros. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association for Computational Linguistics*, 6:225–240, 2018. URL https://aclanthology.org/Q18-1017.

Philipp Koehn. *Statistical machine translation*. Cambridge University Press, Cambridge, 2009.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-3204.

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. 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, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-2045.

Daniel Kondratyuk, Ronald Cardenas, and Ondřej Bojar. Replacing linguists with dummies: A serious need for trivial baselines in multi-task neural machine trans-

lation. *The Prague Bulletin of Mathematical Linguistics*, 113:31–40, 2019. URL https://ufal.mff.cuni.cz/pbml/113/art-kondratyuk-cardenas-bojar.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-3019.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. Scalable syntax-aware language models using knowledge distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1337.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, NM, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1054.

An Nguyen Le, Ander Martinez, Akifumi Yoshimoto, and Yuji Matsumoto. Improving sequence to sequence neural machine translation by utilizing syntactic dependency information. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 21–29, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1003.

Mike Lewis and Mark Steedman. 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, October 2014. Association for Computational Linguistics. URL https://aclanthology.org/D14-1107.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. URL https://aclanthology.org/Q16-1037.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, 2016. URL https://arxiv.org/abs/1511.06114.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclanthology.org/D15-1166.

Dominik Macháček. *Enriching neural MT through multi-task training*. Master's thesis, Charles University, Prague, 2018.

David Mareček and Rudolf Rosa. From balustrades to Pierre Vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-4827.

David Mareček, Hande Çelikkanat, Miikka Silfverberg, Vinit Ravishankar, and Jörg Tiedemann. Are multilingual neural machine translation models better at capturing linguistic features? *The Prague Bulletin of Mathematical Linguistics*, 115:143–162, 2020. URL https://ufal.mff.cuni.cz/pbml/115/art-marecek-et-al.

Igor Aleksandrovic Mel'čuk. *Dependency Syntax: Theory and Practice*. SUNY press, Albany, NY, 1988.

Maria Nădejde. *Syntactic and semantic features for statistical and neural machine translation*. PhD thesis, The University of Edinburgh, 2018.

Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-4707.

Jan Niehues and Eunah Cho. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-4708.

Joakim Nivre. Dependency grammar and dependency parsing. *MSI report*, 5133 (1959):1–32, 2005. URL https://cl.lingfil.uu.se/~nivre/docs/05133.pdf.

Hiroki Ouchi, Kevin Duh, and Yuji Matsumoto. Improving dependency parsers with supertags. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 154–158, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL https://aclanthology.org/E14-4030.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, July 2002. Association for Computational Linguistics. URL https://aclanthology.org/P02-1040.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-demos.14.

Alessandro Raganato and Jörg Tiedemann. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5431.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. Multi-representation ensembles and delayed SGD updates improve syntax-based NMT. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

*Linguistics (Volume 2: Short Papers)*, pages 319–325, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-2051.

Rico Sennrich. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2060.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. URL https://aclanthology.org/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, August 2016b. Association for Computational Linguistics. URL https://aclanthology.org/P16-1162.

Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclanthology.org/D16-1159.

Miloš Stanojević and Mark Steedman. Max-margin incremental CCG parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4111–4122, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.378.

Mark Steedman. A very short introduction to CCG. Unpublished manuscript, 1996. URL https://www.inf.ed.ac.uk/teaching/courses/nlg/readings/ccgintro.pdf.

Mark Steedman. *Taking scope: The natural semantics of quantifiers*. MIT Press, Cambridge, MA, 2012.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Lucien Tesnière. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959.

Antonio Toral and Víctor M. Sánchez-Cartagena. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1100.

John Torr, Miloš Stanojević, Mark Steedman, and Shay B. Cohen. Wide-coverage neural A* parsing for Minimalist Grammars. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2486–2505, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1238.

Gerjan van Schaaik. *The Oxford Turkish Grammar*. Oxford University Press, Oxford, 2020.

Eva Vanmassenhove and Andy Way. SuperNMT: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-3010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. In *25th Annual German Conference on Artificial Intelligence*, pages

18–32, Aachen, Germany, 2002. Deutsche Jahrestagung für Künstliche Intelligenz. URL https://link.springer.com/chapter/10.1007/3-540-45751-8_2.

# Appendix A

# JoeyNMT Training Hyper-parameters

In Table A.1 below I provide the hyper-parameters used to train the baseline and improved Transformer models in JoeyNMT. The codebase for the modified tag attention model with decay, plus various data pre-processing scripts, will be made publicly available.

| Parameter | TR→EN | Parameter | TR→EN |
|---|---|---|---|
| Tokenisation | BPE | Validation frequency | 1/epoch |
| BPE merges | 85000 | Shuffled | True |
| Max. input/output length | 70 | Initialiser | xavier |
| Optimiser | adam | Bias initialiser | zeros |
| Adam betas | [0.9, 0.999] | Initialisation gain | 1.0 |
| Scheduling | plateau | Embedding initialiser | xavier |
| Normalisation | token | Tied embeddings | True |
| Patience | 5 | Tied softmax | True |
| Learning rate | 0.0002 | Encoder/Decoder layers | 6 |
| LR decrease factor | 0.7 | Encoder/Decoder heads | 8 |
| Label smoothing | 0.1 | Embedding dimension | 512 |
| Batch size | 2048 | Scaled embeddings | True |
| Batch type | token | Embedding dropout | 0 |
| Evaluation metric | BLEU | Hidden size | 512 |
| Early stopping metric | BLEU | FFNN size | 2048 |
| Epochs | 20 | Dropout | 0.1 |
| Batch multiplier | 1 | Tag embedding dimension | 128 |

Table A.1: Full training hyper-parameter set for models reported in this study.

# Appendix B

# Supplementary Results

**Smaller Vocabulary —** This experiment does not directly investigate the principal research question or alternative hypotheses, thus I include it here in the appendix. I tested whether improvements are still observed for a small vocabulary with 30,000 BPE merge operations/tokens, as Ataman et al. (2017) use for Turkish to English translation, compared to 85,000 tokens as per Nădejde et al.'s (2017) interleaving model, meaning there will be more subwords in each sentence. The results in Table B.1 show that improvements from tag attention (about 0.3 BLEU) largely remain alongside the consistent overall 1.4 BLEU improvement for all smaller vocabulary models. The overall improvement for a smaller vocabulary is likely because making predictions over a small vocabulary is a much more constrained task, then the model can learn to associate CCG supertags with multiple sub-words. Therefore these results conform with expectations and are included for completeness. In future work, it may be appropriate to use a smaller BPE vocabulary alongside tag attention, given the overall improvement.

| | | TR→EN | |
|---|---|---|---|
| Model | Vocab | Dev | Test |
| Baseline | | 15.55 | 8.73 |
| Tag Attention | 85,000 | 15.77 | 8.80 |
| + Decay | | 15.90 | 9.11 |
| Baseline | | 16.99 | 10.07 |
| Tag Attention | 30,000 | 17.39 | 10.21 |
| + Decay | | 17.23 | **10.36** |

Table B.1: Target-side syntax experiments for Turkish→English translation with a smaller BPE vocabulary, reporting BLEU scores for baseline, TA-NMT and decayed TA-NMT models with 85,000 and 30,000 token BPE vocabularies.