An Attentive Recurrent Model for Incremental Prediction of Sentence-final Verbs

Wenyan Li
Comcast AI Research Lab
wenyanl9562@gmail.com

Alvin Grissom II
Haverford College
agrissom@haverford.edu

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Abstract

Verb prediction is important for understanding human processing of verb-final languages, with practical applications to real-time simultaneous interpretation from verb-final to verb-medial languages. While previous approaches use classical statistical models, we introduce an attention-based neural model to incrementally predict final verbs on incomplete sentences in Japanese and German SOV sentences. To offer flexibility to the model, we further incorporate synonym awareness. Our approach both better predicts the final verbs in Japanese and German and provides more interpretable explanations of why those verbs are selected.

1 Introduction

Final verb prediction is fundamental to human language processing in languages with subject-object-verb (SOV) word order, such as German[1] and Japanese, (Kamide et al., 2003; Momma et al., 2014; Chow et al., 2018) particularly for simultaneous interpretation, where an interpreter generates a translation in real time. Instead of waiting until the entire sentence is completed, simultaneous interpretation requires translation of the source text units while the interlocutor is speaking.

When human simultaneous interpreters translate from an SOV language to an SVO one incrementally—without waiting for the final verb at the end of a sentence—they must use strategies to reduce the lag, or delay, between the time they hear the source words and the time they translate them (Wilss, 1978; He et al., 2016). One strategy is final verb prediction: since the verb comes late in the source sentence but early in the target translation, if the verb is predicted in advance, it can be translated before it is heard, allowing for a more

German	Cazeneuve dankte dort den Männern und sagte, ohne deren kühlen Kopf hätte es vielleicht ein "furchtbares Drama" **gegeben**.
English	Cazeneuve thanked the men there and said that without their cool heads **there might have been** a "terrible drama".
Japanese	また大和国奈良県の葛城山に篭り密教の宿曜秘法を習得したとも **言わ**.
English	It also **said** that he was acquainted with a secret lodging accommodation in Katsuragiyama in Nara Prefecture of Yamato.

Figure 1: An example of the verb position difference between SOV and SVO languages, where the final verb in German and Japanese is expected much earlier in their English translation.

"simultaneous" (or **monotonic**) translation (Jörg, 1997; Bevilacqua, 2009; He et al., 2015). Furthermore, Chernov et al. (2004) argue that simultaneous interpreters' probabilty estimates and predictions of the verbal and semantic structure of preceeding messages facilitates simultaneity in human simultaneous interpretation.

Like for human translation, simultaneous machine translation (SMT), becomes more monotonic for SOV–SVO with better verb prediction (Grissom II et al., 2014; Gu et al., 2017; Alinejad et al., 2018). Earlier work used pattern-matching rules (Matsubara et al., 2000), n-gram language models (Grissom II et al., 2014), or a logistic regression with linguistic features (Grissom II et al., 2016). Recent neural simultaneous translation systems have integrated prediction into the encoder-decoder model or argued that these predictions, including verb predictions, are made implicitly by such models (Gu et al., 2017; Alinejad et al., 2018), but they have not systmatically studied the late-occurring verb predictions themselves.

[1] German is rich in both SOV and SVO sentences. It has been argued that its underlying structure is SOV (Bach, 1962; Koster, 1975), but this is not immediately relevant to our task.

German Auch die deutschen Skispringer können sich Hoffnungen auf ihre erste Medaille bei den Winterspielen in Vancouver [**machen, schaffen, tun**].
English The German ski jumpers can also **hope for** their first medal at the Winter Games in Vancouver.

Figure 2: An example of alternatives of final verbs ("machen", "schaffen", "tun") that preserve same general meaning in German and do not influence its translation in English.

While neural models can identify complex patterns from feature-rich datasets (Goldberg, 2017), less research has gone into problem of *long-distance* prediction, particularly for sentence-final verbs, where predictions must be made with incomplete information. We introduce a neural model, **A**ttentive **N**eural **V**erb **I**nference for **I**ncremental **L**anguage (ANVIIL) for verb prediction, which predicts verbs earlier and with higher accuracy. Moreover, we make ANVIIL's predictions more flexible by introducing synonym awareness. Self-attention also allows visualizes why a certain verb is selected and how it relates to specific tokens in the observed subsentence.

2 The Problem of Verb Prediction

Given an SOV sentence, we want to predict the final verb *as soon as possible* in an incremental setting. For example, in Figure 1, the final verb, "gegeben", in German is expected to be translated together with "hätte es" as "there would have been" in the middle of the English translation.

Human interpreters will often predict a related verb rather than the *exact* verb in a reference translation, while preserving the same general meaning, since predicting the exact verb in a reference translation is difficult (Jörg, 1997). For instance, in Figure 2, besides "machen", verbs such as "schaffen" and "tun" also offen pair with "Hoffnungen" to express "hope for" in English. We therefore include two verb prediction tasks: first, we learn to predict the exact verb; second, we learn to predict verbs semantically similar to the exact reference verb. We describe these two tasks below.

2.1 Exact Prediction

We follow Grissom II et al. (2016), who formulate final verb prediction as sequential classification: a sentence is revealed to the classifier incrementally, and the classifier predicts the exact verb at each time step. While Grissom II et al. (2016) use logistic regression with engineered linguistic features, we use a recurrent neural model with self-attention, which learns embeddings[2] and a context representation that captures relations between tokens, regardless of the distance. Verbs are predicted by classifying on the learned representation of incomplete sentences.

2.2 Synonym-aware Prediction

We also extend the idea in Section 2.1 to allow for synonym-aware predictions: for example, the verb synonym "give", used in place of "provide", preserves the intended meaning in most circumstances and can be considered a successful prediction. Instead of training the model to focus on one fixed verb for each input, we encourage the model to be confident about a *set of* verb candidates which are generally correct in the context.

3 A Neural Model for Verb Prediction

This section describes ANVIIL's structure. Gated recurrent neural networks (RNNs), such as LSTMs (Hochreiter and Schmidhuber, 1997) and gated recurrent units (Cho et al., 2014, GRUs), can capture long-range dependencies in text, which we need for effective verb prediction.

We construct an RNN-based classifier with self-attention (Lin et al., 2017) for predicting sentence-final verbs (Figure 3). This is a natural encoding of the problem, as it explicitly models how interpreters might receive information and update their verb predictions. The hidden states of the sequence model can be either at the word or character level.

3.1 BiGRU Sequence Encoder

Following Yang et al. (2016), we encode input sequences using the bidirectional GRU (BiGRU).[3] Given an incomplete sentence prefix $\boldsymbol{x} = (x_1, x_2, \cdots, x_l)$ of length l, BiGRU takes as input the embeddings $(\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_l})$, where $\boldsymbol{w_i}$ is the d-dimensional embedding vector of x_i. At time

[2]Character and word embeddings are learned from scratch, as pretrained embeddings (Bojanowski et al., 2017) did not improve prediction.

[3]While it may be initially counterintuitive to use a BiGRU for an incremental task, since we make predictions at each time step independently—i.e., without consulting prior predictions—there is no need to restrict ourselves to a unidirectional model.

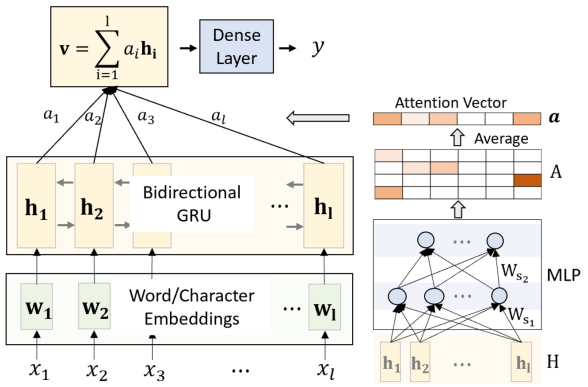

Figure 3: ANVIIL. Token sequences at the input layer are mapped to embeddings, which go to the GRU. The dot product of attention weights and hidden states pass through a dense layer to predict the verb.

step t, the forward and backward hidden states are:

$$\begin{aligned}\overrightarrow{\boldsymbol{h}_t} &= \overrightarrow{\text{GRU}}(\boldsymbol{w}_t, \overrightarrow{\boldsymbol{h}_{t-1}}) \\ \overleftarrow{\boldsymbol{h}_t} &= \overleftarrow{\text{GRU}}(\boldsymbol{w}_t, \overleftarrow{\boldsymbol{h}_{t+1}}).\end{aligned} \tag{1}$$

These are concatenated as $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}]$ and we represent the input sequence as

$$H = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_l). \tag{2}$$

As we only use a prefix of the sentence as input for prediction, we won't be able to see backward messages from unrevealed. However, once we see those words, later words in the prefix *do change* the internal representation of earlier words in H, creating a more powerful overall representation that uses more of the available context.

Embedding vectors for the input can be word embeddings or character embeddings, yielding a word-based or a character-based model; we try both in Section 4.

3.2 Structured Self-attention

Following Lin et al. (2017), we apply self-attention with multiple views of the input sequence to obtain a weighted context vector $\boldsymbol{v}$. By viewing the sequence multiple times, it allows different attentions to be assigned at each time. Using a two layer multilayer perceptron (MLP) without bias and a softmax function over the sequence length, we have an r-by-l attention matrix A, which includes r attention vectors extracted from r views of $\boldsymbol{x}$:

$$A = \text{softmax}(W_{s_2} \tanh(W_{s_1} H^T)) \tag{3}$$

We sum over all r attention vectors and normalize, yielding a single attention vector $\boldsymbol{a}$ with normalized weights (Figure 3). By assigning each hidden state its attention a_t, we acquire an overall representation of the sequence:

$$\boldsymbol{v} = \sum_{t=1}^{l} a_t \boldsymbol{h_t}. \tag{4}$$

3.3 Verb Predictor

For an incomplete input prefix $\boldsymbol{x}$, the target verb is $y \in \mathcal{Y} = \{1, 2, \ldots, K\}$. Based on the high-level representation $\boldsymbol{v}$ of the input sequence, we compute the probability of each verb k and select the one with the highest probability as the predicted verb:

$$p(y \mid \boldsymbol{v}) = \frac{e^{f_y(\boldsymbol{v})}}{\sum_{k=1}^{K} e^{f_k(\boldsymbol{v})}} \tag{5}$$

where $f_k(\boldsymbol{v})$ is the logit from the dense layer.

3.3.1 Exact Verb Prediction

As there is only one ground-truth verb y for the input, we maximize the log-likelihood of the correct verb with cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{K} q(k \mid \boldsymbol{v}) \log p(k \mid \boldsymbol{v}) \tag{6}$$

where $q(k \mid \boldsymbol{v})$ is the ground-truth distribution over the verbs, which equals 1 if $k = y$, or 0 otherwise.

3.3.2 Synonym-aware Verb Prediction

In addition to the exact verb y, we add verbs that are of similar meaning to y in to a synonym set $\mathcal{Y}' \subset Y$, creating a verb candidate pool for each input sample. Instead of maximizing the log-likelihood of the fixed verb y, we maximize the log-likelihood of the most probable verb candidate $y' \in \mathcal{Y}'$ dynamically through training:

$$\mathcal{L} = -\sum_{k=1}^{K} q'(k \mid \boldsymbol{v}) \log p(k \mid \boldsymbol{v}) \tag{7}$$

where

$$q'(k \mid \boldsymbol{v}) = \begin{cases} 1, & \text{if } k = \underset{k \in \mathcal{Y}'}{\text{argmax}}\, p(k \mid \boldsymbol{v}) \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

As the candidate can be different in each step, overall the likelihood of any verb candidate in the synonym set is maximized in the training process.

Most Frequent Verbs		Thousand of Verbs	Coverage (%)
DE (Inflected)	100	1286.7	16.0
	200	2243.7	28.0
	300	2577.3	32.2
JA (Normalized)	100	70.2	56.8
	200	85.2	68.9
	300	93.2	75.4

Table 1: Dataset for final-verb prediction. We extract sentences with the most frequent 100–300 verbs in German and Japanese verb final sentences. Using normalized Japanese verbs reduces the sparsity of the verbs and improves coverage of sentences.

4 Exact Prediction Experiments

We first test exact prediction on both Japanese and German verb-final sentences with both word-based and character-based models.

4.1 Datasets

We use German and Japanese verb-final sentences between ten and fifty tokens (Table 1) that end in the 100 to 300 most common verbs (Wolfel et al., 2008). For each sentence, the extracted final verb becomes the label; the token sequence preceding it (the **preverb**) is the input. We split sentences into train (64%), evaluation (16%) and test (20%) sets.

For Japanese, we use the Kyoto Free Translation Task (KFT) corpus of Wikipedia articles. Since Japanese is unsegmented, we use the morphological analyzer MeCab (Kudo, 2005) for tokenization. Like Grissom II et al. (2016), we strip out post-verbal copulas and normalize verb forms to the dictionary *ru* (non-past tense) form. We also consider *suru* light verb constructions a single unit.

For German, we use the Wortschatz Leipzig news corpus from 1995 to 2015 (Goldhahn et al., 2012). German sentences ending with a verb (we throw out verb medial sentences) are tokenized and POS-tagged with TreeTagger (Schmid, 1995). Since German sentences may end with two verbs—for example, a verb followed by *ist*, we only predict the content verb, i.e., the first verb in the two-verb sequence. Unlike Japanese, we leave German verbs inflected, as there is less variation (usually past participle or infinitive form).

4.2 Training Data Representation

Because we predict from partial input, we train on incrementally longer preverb subsequences. Each subsequence is an independent input sample during training, and each preverb is truncated into five progressively longer subsentences: 30%, 50%, 70%, 90%, and 100%.[4]

4.3 Training Details

We train both word- and character-based models for German and Japanese verb prediction. We use the dev sets to manually tune hyperparameters for accuracy—word embedding size, hidden layer size, dropout rates and learning rate.

Character-based Model For input character sequences, we learn 64-dimensional embeddings and encode them with a two-layer BiGRU of 256 hidden units. The embeddings are randomly initialized with PyTorch defaults and updated during training jointly with other parameters. Mini-batch sizes are 256 for German but 128 for Japanese's smaller corpus. We use the evaluation set for tuning and set the embedding dropout rate as 0.6 and the RNN dropout rate as 0.2 while averaging from five views for attention vectors. We optimize with Adam (Kingma and Ba, 2015) with an initial learning rate of 10^{-4}, decaying by 0.1 when loss increases. Training takes approximately two (Japanese) and four (German) hours on one 6GB GTX1060 GPU.

Word-based Model We use a vocabulary of 50,000 for German and Japanese; we use the <*UNK*> token for out-of-vocabulary tokens. The embedding size is 300. We encode the input embeddings with a two-layer BiGRU with 512 hidden units. Other hyperparameters are unchanged from the character-based model.

4.4 Results

We compare ANVIIL to the logistic regression model[5] in Grissom II et al. (2016) on the 100 most frequent verbs in the corpus (Figure 4). For both languages, ANVIIL has higher accuracy than previous work (Figure 5), especially early in the sentence. While word-based models work best for German, character-based models work best for Japanese, perhaps because it is agglutinative.

Figure 6 compares other encodings of preverbs (at a character level) in Japanese. In general, ANVIIL has higher accuracy on verb prediction tasks.

[4] As input sequence lengths vary, we pad input samples with zeros and train in minibatches *a la* neural MT (Doetsch et al., 2017; Morishita et al., 2017).

[5] This model uses token unigrams and bigrams, case marker bigrams, and the last observed case marker as features.

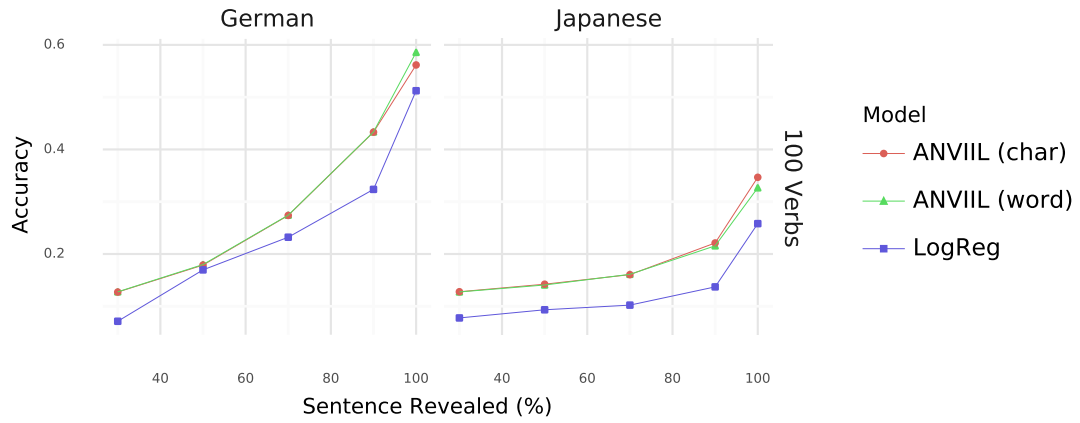

Figure 4: Comparing word and character representations for German (inflected) and Japanese (normalized) verb prediction. ANVIIL consistently has higher accuracy than LogReg from Grissom II et al. (2016), and word-based prediction is slightly better for German but worse for Japanese.

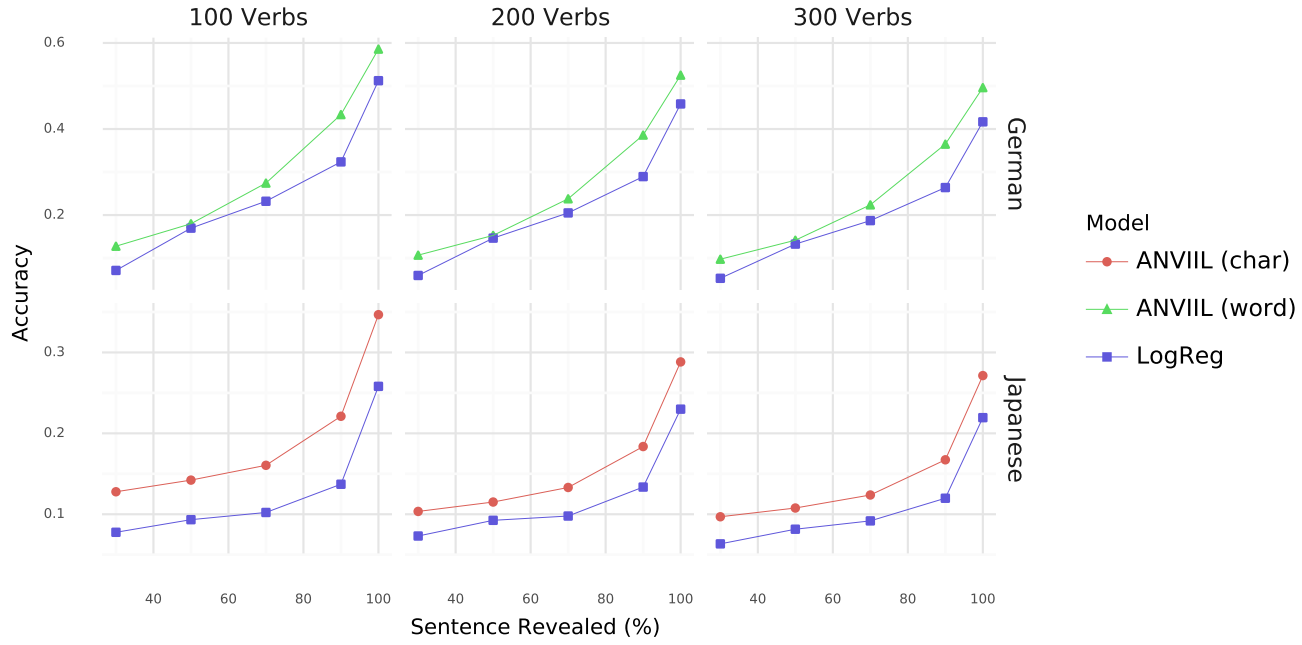

Figure 5: Accuracy when classifying among the most common 100, 200, and 300 verbs. ANVIIL consistently outperforms the best-performing model described in Grissom II et al. (2016), especially early in the sentences.

5 Synonym-aware Prediction

We now describe synonym-aware verb prediction (Section 4). We use 2,214,523 German sentences ending with 100 most frequent lemmatized verbs. For each sentence, we extract the preverb as in Section 4.1, but in this case, the target is not just a single verb. For each lemmatized verb, we extract its synonyms among the 100 verbs using Germanet synsets (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010). If synonyms exist, we include them all in a list as candidate target verbs for the input as in Figure 2. Synonyms exist for 40.79% of the sentences in the dataset.

Similarly, we train incrementally on subsequences of the preverb as in Section 4.3. We learn high-level representations of the preverb using word-level embeddings and use the same training parameters as in Section 4.3

During training, instead of maximizing the exact verb's log-likelihood, we maximize the log-likelihood of any verb in the synonym-set, encouraging the model to be confident about *any* verb that fits in the context.

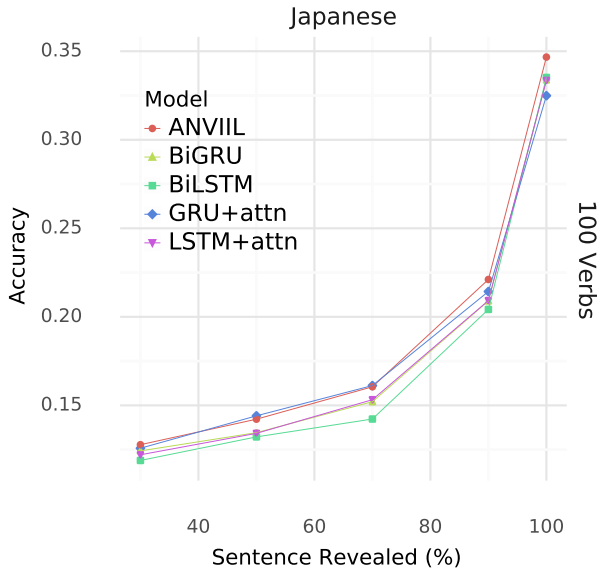

Figure 6: ANVIIL's BiGRU with self-attention outperforms other most settings on predicting the 100 most common verbs in Japanese.

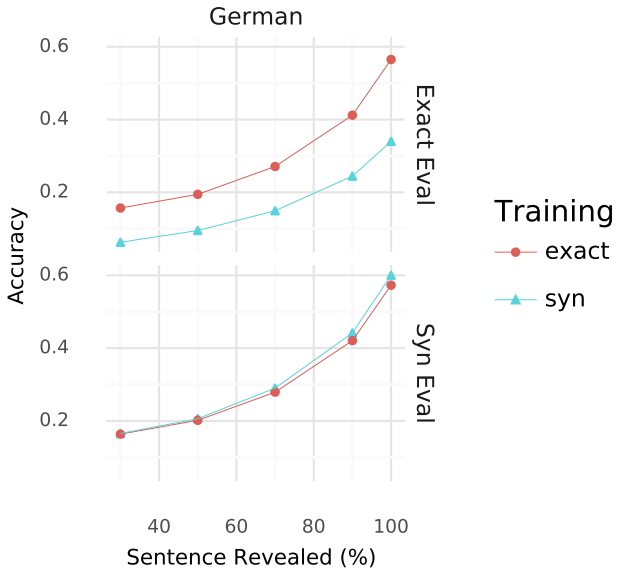

Figure 7: Accuracy across time on exact/synonym-aware match with exact/synonym-aware training. Accuracy increases slightly with the addition of the synonym-aware matching.

5.1 Verb Prediction Results

We compare accuracy for predicting exact and synonym-aware verbs with different objects in training. In synonym-aware prediction, we consider the prediction successful if it is one of the candidate verbs. Compared to predicting the exact verb, while being less focused on the fixed verb, synonym-aware prediction further improves the predication accuracy (Figure 7), but only slightly. ANVIIL clearly outperforms the feature engineering linear models on Japanese across the entire sentence, even when the number of verbs to choose from is larger; and on German, ANVIIL outperforms previous models when the number of verbs to choose from is the same (Figure 4). This is may be due to the long-range dependencies which are not captured in the logistic regression model.

6 Visualization and Analysis

We now analyze our model's predictions. While previous work (Grissom II et al., 2016) examines the contribution of features by examining the model itself, our approach does not rely on feature engineering. To examine our model, we instead use a heatmap to visualize the time course attention values in sentences, allowing us to see on what the model focuses when predicting.

6.1 Visualization of the Prediction Process

We visualize how our model makes its predictions in Figure 8 and Figure 9. In both languages, the model not only focuses on the most recent revealed word, but also focuses attention to relevant long-distance dependencies.

Predictions are, as expected, also more confident and accurate when approaching the end of the preverb. This is consistent with the verb prediction process for human interpreters (Wilss, 1978) and with previous work (Grissom II et al., 2016). With increasing information, the number of possible alternatives gradually declines. Figure 10 visualizes how the model makes synonym-aware predictions.

6.2 Character-based versus Word-based

As described in Section 4.3, we implement both character-based and word-based models for verb prediction. For Japanese final-verb prediction, the character-based model has higher prediction accuracy. Unlike the word-based model, it does not require use of a morphological analyzer and has a smaller vocabulary size. The word-based model, however, works better for German verb prediction and word-based heatmaps are more interpretable than character-based ones for German. We show word-based heatmaps for exact prediction in Figure 8 and Figure 11.

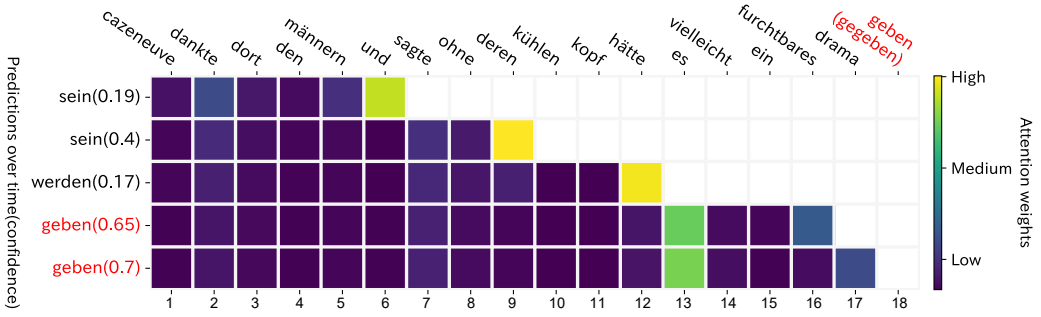

Figure 8: Attention during German verb prediction. The model usually attends to the most recent word, but focuses on "es", which can be used as the *subject* of an existential phrase (Joseph, 2000) in combination with the verb "geben". Thus, it focuses on an interpretation of "es" as the subject, consistently attends to "es" throughout the sentence, and correctly predicts "geben" (for consistency with the Japanese examples, we show the model that predicts the normalized—infinitive—form of the verb).

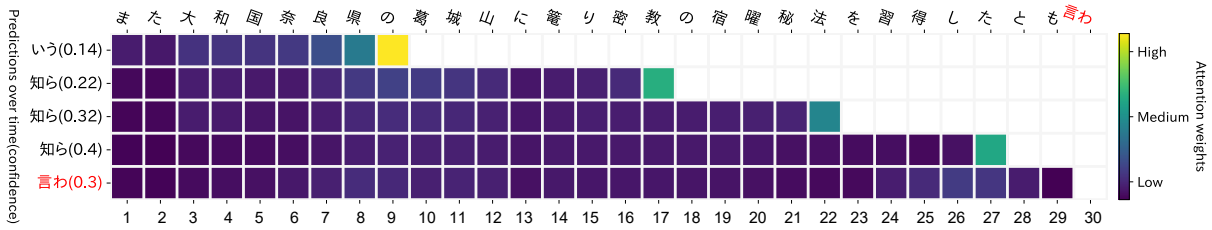

Figure 9: Attention during Japanese verb prediction. Attention and prediction transition through time on a Japanese sentence. The genitive case marker *no*, in bright yellow, has a high attention weight, as do the characters making in the noun before it. Case marker-adjacent nouns, including before the genitive *no* (twice) and the accusative *wo* have slightly less. Toward the end of the sentence, attention shifts to the quotative particle *to*, which significantly limits possible completions.

6.3 Synonym-aware versus Exact Prediction

We show an example of how synonym-aware prediction can make the task easier in Figure 12. By providing synonyms during training, the model makes an alternative prediction "zeigen" (present, show) for the original verb "einsetzen" (use).

6.4 Case Markers

Previous work suggests that case markers play a key role in both human and machine verb prediction for Japanese (Grissom II et al., 2016). Japanese has explicit postposition case markers which mark the roles of the words in a sentence. By examining the accuracy of predictions when the most recent token is a case marker, we can gain insight into their contributions to the predictions.

Figure 13 considers the instances where the most recent token observed is the given case marker; in these situations, the accuracy of predicting one of the 100 most frequent verbs is much higher than in general. It is unsurprising that the quotative particles have higher accuracy at the end of the sentence, since the set of verbs that follow them is highly constrained—e.g., *say*, *think*, *announce*, etc. Quotative particles for the entire sentence occur immediately before to final verb. More general particles, such as *ga* (NOM) and *wo* (ACC) show a smaller increase in accuracy.

7 Related Work

This section examines previous work on prediction in humans, simultaneous interpretation, and simultaneous machine translation.

Psycholinguistics has examined argument structure using verb-final *bǎ*-construction sentences in Chinese (Chow et al., 2015, 2018). Kamide et al. (2003) find that case markers facilitate verb predictions for humans, likely because they provide clues about the semantic roles of the marked words in sentences. In sentence production, Momma et al. (2015) suggest that humans plan verbs after selecting a subject but before objects.

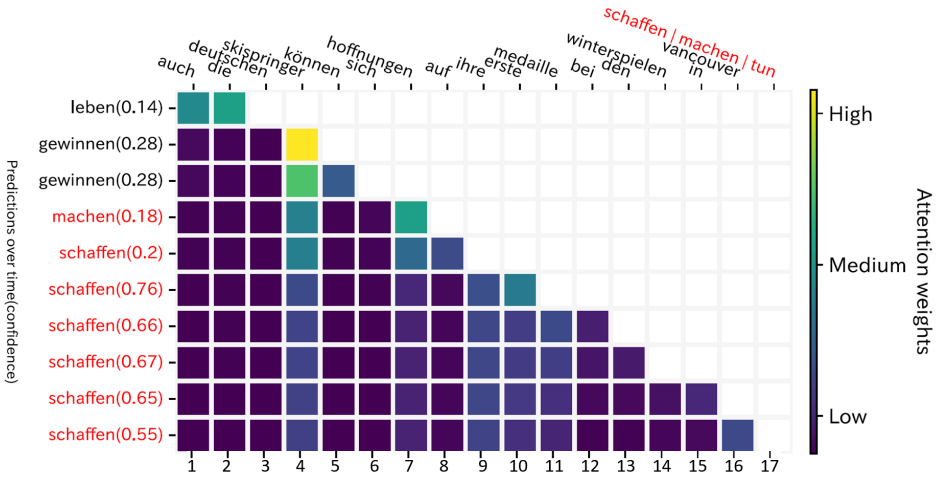

Figure 10: Attention during German synonym-aware verb prediction. The model constantly focuses on "skispringer" (ski jumpers), which is the subject of the verb and predicts "machen" and "schaffen" from three of the verb candidates.

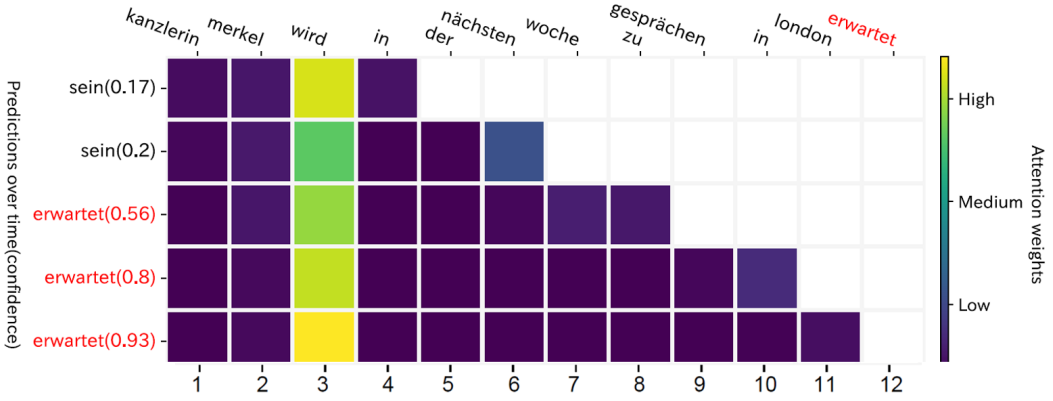

Figure 11: Progression of attention weights of a word-based model on a German sentence. The model successfully captures the passive voice in the sentence where "wird erwartet" is often translated together as "is expected". Full translation of the example is: Chancellor Merkel is expected to speak in London next week.

Empirical work on German verb prediction first investigated German–English simultaneous interpreters in Jörg (1997): professional interpreters often predict verbs. Matsubara et al. (2000) introduce early verb prediction into Japanese–English SMT by predicting verbs in the target language. Grissom II et al. (2014) and Gu et al. (2017) use verb prediction in the source language and learn when to trust the predictions with reinforcement learning, while Oda et al. (2015) predict syntactic constituents and do the same. Grissom II et al. (2016) predict verbs with linear classifiers and compare the predictions to human performance. We extend that approach with a modern model that explains which cues the model uses to predict verbs.

In interactive translation (Peris et al., 2017) and simultaneous translation (Alinejad et al., 2018; Ma et al., 2019) systems, neural methods for next word prediction improve translation. BERT (Devlin et al., 2019) uses masked deep bidirectional language models and contextualized representations (Peters et al., 2018) for pretraining and gain improvements in word prediction and classification. We incorporate bidirectional encoding to verb prediction.

Existing neural attention models for sequential classification are commonly trained on complete input (Yang et al., 2016; Shen and Lee, 2016; Bahdanau et al., 2014). Classification on incomplete sequences and long-distance sentence-final verb prediction remains difficult and under-explored.

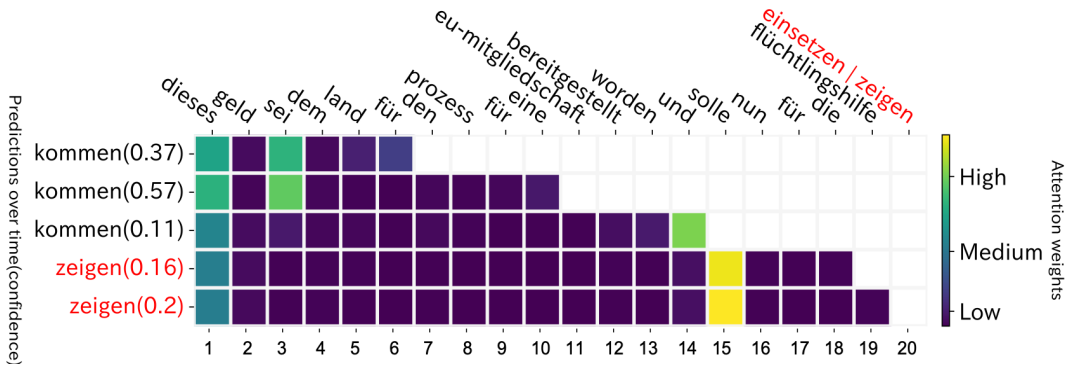

Figure 12: Imperfect synonym-aware prediction process on a German sentence. The predicted synonym "zeigen" (show/appear) in context is not a perfect replacement for the correct verb "einsetzen" (put in place), but it better preserves the general meaning of the sentence: "This money had been made available to the country for the process of EU membership and should now appear for refugee assistance."

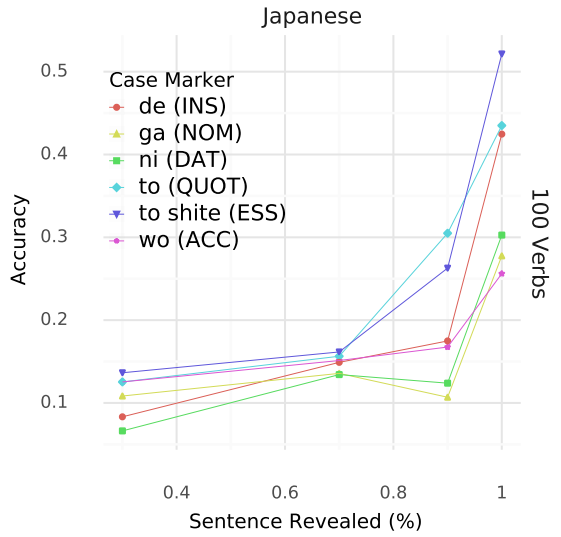

Figure 13: Case markers correlate with improved verb prediction compared to overall verb prediction (Figure 4). Some case markers, such as *to*, have large jumps in accuracy toward the end, while others, such as *wo* do not. We examine nominative (NOM), instructive (INS), accusative (ACC), dative (DAT), quotative (QUOT), and essive (ESS) markers.

8 Conclusion

We present a synonym-aware neural model for incremental verb prediction using BiGRU with self-attention. It outperforms existing models in predicting the most frequent sentence-final verbs in both Japanese and German. As we predict the verbs incrementally, our method can be directly applied to solve real-time sequential classification or prediction problems. SMT systems for SOV to SVO simultaneous MT can also benefit from our work to reduce translation latency. We show that larger datasets always help with predicting the sentence-final verbs, suggesting that larger corpora will further improve results.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1748663 (UMD). The views expressed in this paper are our own. We thank Graham Neubig and Hal Daumé III for useful feedback.

References

Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. Prediction improves simultaneous neural machine translation. In *Conference of Empirical Methods in Natural Language Processing*, pages 3022–3027.

Emmon Bach. 1962. The order of elements in a transformational grammar of German. *Language*, 38(3):263–269.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*.

Lorenzo Bevilacqua. 2009. The position of the verb in Germanic languages and simultaneous interpretation. *The Interpreters' Newsletter*, 14:1–31.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

G.V. Chernov, R. Setton, and A. Hild. 2004. *Inference and Anticipation in Simultaneous Interpreting: A Probability-prediction Model*. Benjamins translation library. J. Benjamins Publishing Company.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference of Empirical Methods in Natural Language Processing*.

Wing-Yee Chow, Ellen Lau, Suiping Wang, and Colin Phillips. 2018. Wait a second! delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7):803–828.

Wing-Yee Chow, Cybelle Smith, Ellen Lau, and Colin Phillips. 2015. A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, pages 1–20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Association for Computational Linguistics*.

Patrick Doetsch, Pavel Golik, and Hermann Ney. 2017. A comprehensive study of batch construction strategies for recurrent neural networks in MXNet. *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *International Language Resources and Evaluation*.

Alvin Grissom II, Naho Orita, and Jordan Boyd-Graber. 2016. Incremental prediction of sentence-final verbs: Humans versus machines. In *Conference on Computational Natural Language Learning*, pages 95–104.

Alvin C. Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Conference of Empirical Methods in Natural Language Processing*.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. *European Chapter of the Association for Computational Linguistics*.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

He He, Alvin Grissom II, Jordan Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In *Conference of Empirical Methods in Natural Language Processing*.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT—the GermaNet editing tool. In *International Language Resources and Evaluation*. European Languages Resources Association (ELRA).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Udo Jörg. 1997. Bridging the gap: Verb anticipation in German-English simultaneous interpreting. In M. Snell-Hornby, Z. Jettmarová, and K. Kaindl, editors, *Translation as Intercultural Communication: Selected Papers from the EST Congress, Prague 1995*.

Brian Joseph. 2000. What gives with es gibt? *American Journal of Germanic Linguistics and Literatures*, 12:243–265.

Yuki Kamide, Gerry Altmann, and Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1):133–156.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Jan Koster. 1975. Dutch as an SOV language. *Linguistic analysis*, 1(2):111–136.

T. Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the International Conference on Learning Representations*.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework.

Shigeki Matsubara, Keiichi Iwashima, Nobuo Kawaguchi, Katsuhiko Toyama, and Yasuyoshi Inagaki. 2000. Simultaneous Japanese-English interpretation based on early predition of English verb. In *Symposium on Natural Language Processing*.

Shota Momma, L Robert Slevc, and Colin Phillips. 2015. The timing of verb selection in japanese sentence production. *Journal of experimental psychology. Learning, memory, and cognition*.

Shota Momma, Robert Slevc, and Colin Phillips. 2014. The timing of verb selection in english active and passive sentences.

Makoto Morishita, Yusuke Oda, Graham Neubig, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2017. An empirical study of mini-batch creation strategies for neural machine translation. In *The First Workshop on Neural Machine Translation*.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Syntax-based simultaneous translation through prediction of unseen syntactic constituents. *Proceedings of the Association for Computational Linguistics*.

lvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Computer Speech and Language*, 45:201–220.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*.

Sheng-syun Shen and Hung-yi Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. In *Conference of the International Speech Communication Association*.

Wolfram Wilss. 1978. Syntactic anticipation in German-English simultaneous interpreting. In *Language Interpretation and Communication*.

M. Wolfel, M. Kolss, F. Kraft, J. Niehues, M. Paulik, and A. Waibel. 2008. Simultaneous machine translation of German lectures into English: Insvestigating research challenges for the future. In *IEEE Spoken Language Technology Workshop*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.