Incremental Prediction of Sentence-final Verbs: Humans versus Machines

Alvin C. Grissom II,[1] Naho Orita,[2] and Jordan Boyd-Graber[1]

[1]University of Colorado Boulder, Department of Computer Science
[2]Tohoku University, Graduate School of Information Sciences
{Alvin.Grissom,Jordan.Boyd.Graber}@colorado.edu
naho@ecei.tohoku.ac.jp

Abstract

Verb prediction is important in human sentence processing and, practically, in simultaneous machine translation. In verb-final languages, speakers select the final verb before it is uttered, and listeners predict it before it is uttered. Simultaneous interpreters must do the same to translate in real-time. Motivated by the problem of SOV-SVO simultaneous machine translation, we provide a study of incremental verb prediction in verb-final languages. As a basis of comparison, we examine incremental verb prediction with human participants in a multiple choice setting using crowdsourcing to gain insight into incremental human performance in a constrained setting. We then examine a computational approach to incremental verb prediction using discriminative classification with shallow features. Both humans and machines predict verbs more accurately as more of a sentence becomes available, and case markers—when available—help humans and sometimes machines predict final verbs.

1 The Importance of Verb Prediction

Humans predict future linguistic input before it is observed (Kutas et al., 2011). This predictability has been formalized in information theory (Shannon, 1948)—the more predictable a word is, the lower the entropy—and has explained various linguistic phenomena, such as garden path ambiguity (Den and Inoue, 1997; Hale, 2001). Such instances of linguistic prediction are fundamental to statistical NLP. Auto-complete from search engines has made next-word prediction one of best known NLP applications.

Long-distance word prediction, such as verb prediction in SOV languages (Levy and Keller, 2013; Momma et al., 2015; **?**) is important in simultaneous machine translation from subject-object-verb (SOV) languages to subject-verb-object (SVO) languages. In SVO languages such as English, for example, the main verb phrase usually comes after the first noun phrase—the main subject—in a sentence, while in verb-final languages such as Japanese or German, it comes very last. Human simultaneous translators must make predictions about the unspoken final verb to incrementally translate the sentence. Minimizing interpretation delay thus requires making constant predictions and deciding when to trust those predictions and commit to translating in real time.

Such prediction can also aid machines. Matsubara et al. (2000) use pattern-matching rules; Grissom II et al. (2014) use a statistical n-gram approach; and Oda et al. (2015) extend the idea of using prediction by predicting entire syntactic constituents for English-Japanese translation. These systems require fast, accurate verb prediction to further improve simultaneous translation systems. We focus on verb prediction in verb-final languages such as Japanese with this motivation in mind.

In Section 2, we present what is, to our knowledge, the first study of humans' ability to incrementally predict the verbs in Japanese. We use these human data as a yardstick to compare computational incremental verb prediction. Incorporating some of the key insights from our human study into a discriminative model—namely, the importance of case markers— Section 3 presents a better incremental verb classifier than existing verb prediction schemes. Having established both human and computer performance on this challenging and interesting task, Section 4 reviews our work's relationship to other studies in NLP and linguistics.

2 Human Verb Prediction

We first examine human verb selection in a constrained setting to better understand what performance we should demand of computational approaches. While we know that humans make incremental predictions across sentences, we do not know how skilled they are in doing so. While it's possible that machines—with unbounded memory and access to Internet-sized data—could do better than humans, this study allows us to appropriately gauge our expectations for computational systems.

We use crowdsourcing to measure how well novice humans can predict the final verb phrase of incomplete Japanese sentences in a multiple choice setting. We use Japanese text of the Kyoto Free Translation Task corpus (Neubig, 2011, KFT), a collection of Wikipedia articles in English and Japanese, representing standard, grammatical text and readily usable for future SOV-SVO machine translation experiments.

2.1 Extracting Verbs and Sentences

This section describes the data sources, preparation, and methodology for crowdsourced verb prediction. Given an incomplete sentence, participants select a sentence-final verb phrase containing a verb from a list of four choices to complete the sentence, one of which is the original completion.

We randomly select 200 sentences from the development set of the KFT corpus (Neubig, 2011). We use these data because the sentences are from Wikipedia articles and thus represent widely-read, grammatical sentences. These data are directly comparable to our computational experiments and readily usable for future SOV-SVO machine translation experiments.

We ask participants to predict a "verb chunk" that would be natural for humans. More technically, this is a sentence-final *bunsetsu*.[1] We identify verb *bunsetsu* with a dependency parser (Kurohashi and Nagao, 1994). Of interest are *bunsetsu* at the end of a sentence that contain a verb. We also use *bunsetsu* for segmenting the incomplete sentences we show to humans, only segmenting between *bunsetsu* to ensure each segment is a meaningful unit.

[1] A *bunsetsu* is a commonly used linguistic unit in Japanese, roughly equivalent to an English phrase: a collection of content words and zero or more functional words. Japanese verb *bunsetsu* often encompass complex conjugation. For example, a verb phrase 読みたくなかった (read-DESI-NEG-PAST), meaning 'didn't want to read', has multiple tokens capturing tense, negation, etc. necessary for translation.

Answer Choice Selection We display the correct verb *bunsetsu* and three incorrect *bunsetsu* completions as choices that occur in the data with frequency close to the correct answer in the overall corpus. We manually inspect the incorrect answers to ensure that these choices are semantically distant, i.e., excluding synonyms or troponyms.

Sentence Presentation We create two test sets of truncated sentences from the KFT corpus: The first, the **full context set**, includes all but the final *bunsetsu*—i.e., the verb phrase—to guess. The second set, the **random length set**, contains the same sentences truncated at predetermined, random *bunsetsu* boundaries. The average sentence length is nine *bunsetsu*, with a maximum of fourteen and minimum of three. We display sentences in the original Japanese script.

Participants view the task as a game of guessing the final verb. Each fragment has four concurrently displayed completion options, as in the prompt (2) and answers (3). Users receive no feedback from the interface.

We use CrowdFlower[2] to collect participants' answers, at a total cost of approximately USD$300. From an initial pool of fifty-six participants, we remove twenty via a Japanese fluency screening. We verify the efficacy of this test with non-native but highly proficient Japanese learners; none passed. We collect five judgments per sentence from each participant.

(2) 谷崎潤一郎は
Junichiro Tanizaki-TOP
すき家を
tea-ceremony house-OBJ

(3) (a) 好んだ
like-COP
(b) 変えれられた
change-PASS CAP-PAST
(c) 始まったとされている
begin-PAST-COMP-suppose-AUX.PRES
(d) 増やしていた
increase-AUX.PAST

2.2 Presenting Partial and Complete Sentences

The first task, on the **full context set**, shows how humans predict the sentence-final verb chunk with all context available. The second task, on the

[2] http://www.crowdflower.com/

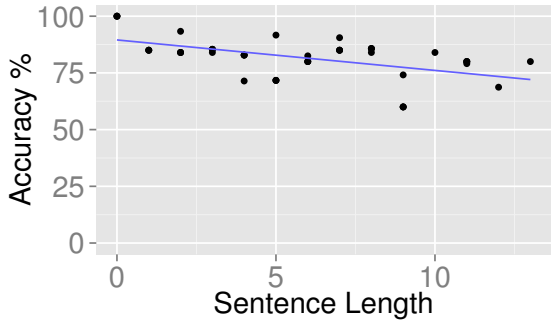

Figure 1: Full context set: Accuracy is generally high, but slightly decreases on longer, more complicated sentences, averaging 81.1%.

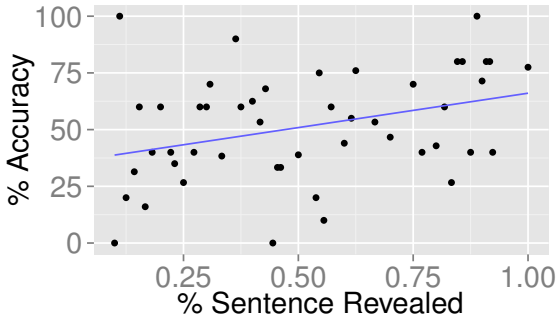

Figure 2: Random length set: The accuracy of human verb predictions reliably increases as more of the sentence is revealed.

random length set, shows how the amount of revealed data affects the predictability of the final verb chunk. We examine a correlation between the length of the pre-verb sentence fragment and participants' accuracy (Figure 1).

Psycholinguistic experiments using lexical decision tasks suggest Japanese speakers start syntactic processing by using case—the type and number of case-marked arguments—before the verb's availability (Yamashita, 2000). We also examine the correlation between the number of case markers[3] and accuracy. It is likely that the number of case markers and the length of the sentence fragment are confounded; so, we create a measure, the proportion of case markers to the overall sentence information (the number of case markers in the fragment divided by the number of *bunsetsu* chunks). We call this **case density**.

2.3 Results of Human Experiments

In the **full context set**, average accuracy over 200 sentences is 81.1%, significantly better than chance ($p < 2.2 \cdot 10^{-16}$). Figure 1 shows the accuracy per sentence length as defined by the *bunsetsu* unit. A one-way ANOVA reveals a significant effect of the sentence length ($F(1, 998) = 7.512, p < 0.00624$), but not the case density ($F(1, 998) = 1.2, p = 0.274$).

In the **random length set**, average accuracy over 200 sentences is 54.2%, significantly better than chance ($t(199) = 11.8205, p < 2.2 \cdot 10^{-16}$). Figure 2 shows the accuracy per percentage of length of the presented sentence fragment. A one-way ANOVA reveals a significant effect of the sentence length ($F(1, 998) = 57.44, p < 7.94 \cdot 10^{-14}$). We also find a significant effect of the case density ($F(1, 998) = 5.884, p = 0.0155$).

2.4 Discussion

Predictability increases with the percent of the sentence available in all of our experiments. By the end of the sentence, the verb chunks are highly predictable by humans in the multiple choice setting. Participants choose the final verb more accurately as they gain access to more case markers in the **random length set** but not in the **full context set**.

Case density is a significant factor in predictive accuracy on the **random length set** for humans, suggesting that the case is more helpful in predicting a sentence-final verb when the preceding contextual information is insufficient. The following example illustrates how case helps in prediction. The nominative and accusative markers greatly narrow the choices, as shown in (4).[4] Our results further support the proposition case markers modulate predictability in SOV verb-final processing.

(4) 江戸幕府区-**が** 成立すると
Edo shogunate-**NOM** establish-do-**CONJ**
寺院法度-**に**-より —
temple-prohibition-etc.-ACC-for

[3] In this study, we counted case markers that mark nominative (*-ga*), accusative (*-wo*), and ablative (*-kara*), and dative (*-ni*).

[4] A recent psycholinguistics study on incremental Japanese verb-final processing (Momma et al., 2015) argues that native Japanese speakers plan verbs in advance, before the articulation of object nouns, but not subject nouns. Since case markers assign the roles of subject and object in Japanese, we expect that a high ratio of case markers to words will increase predictability of verbs. In addition, (Yamashita, 1997) argues that the *variety* of case markers increases predictability just before the verb.

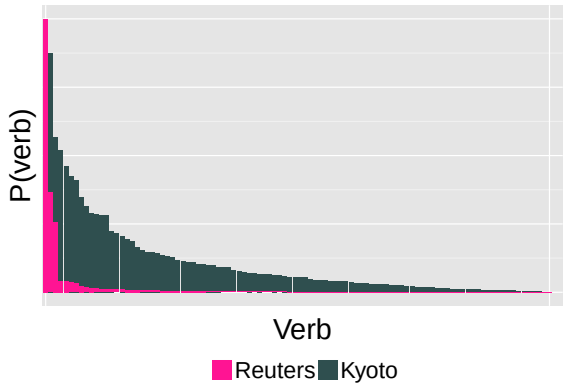

Figure 3: Distribution of the top 100 content verbs in the Kyoto corpus and the Reuters Japanese news corpus. Both are Zipfian, but the Reuters corpus is even more skewed, even with the common special cases excluded.

> 'After Edo shogunate has established, due to the temple prohibition etc. —'

In other cases, there exist choices, which, while incorrect, could naturally complete the sentence. These questions are frequently missed. For instance, in one 90% revealed sentence, the participant has the choices: (i) 収め る (put-PRES), (ii) 厳しくなる (strict-become), (iii) 収録されている (record-do.PASS-AUX.PRES), and (iv) 務める (work-PRES). Choice (i) is the correct answer, but choice (iii) is a reasonable choice for a Japanese speaker. All participants missed this question, and all chose the same wrong answer (iii). We leave a cloze task where participants can freely fill in the sentence-final to future work.

These results provide a basis of comparison for automatic prediction. In the next section, we examine whether computational models can predict final verbs and compare the models' performance to that of humans.

3 Machine Verb Prediction

Now that we have the results of the previous section, we have baselines against which we can compare computational verb prediction approaches. In this section, we introduce incremental verb classification with a linear classifier.[5] For our investigation of computational verb classification, we use two very different languages that both have verb-final syntax—Japanese, which is agglutinative, and German, which is not—and show that discriminative classifiers can predict final verbs with increasing accuracy as more context of sentences is revealed.

A simple verb prediction scheme applied to German (Grissom II et al., 2014) achieves poor accuracy. Their approach creates a Kneser-Ney n-gram language model for the prior context associated with each verb in the corpus; i.e., 50 n-gram models for 50 verbs. Given pre-verb n-gram context c in a sentence S_t, and verb prediction $v^{(t)} \in V$, the verb selection is defined by the following equation:

$$v^{(t)} \equiv \arg\max_{v} \prod_{c \in S_t} p(c \mid v) p(v). \qquad (1)$$

It chooses the verb that maximizes the probability of the observed context, scaled by the prior probability of the verb in the overall corpus. Unsurprisingly, given the distribution of verbs in real data (Figure 3), this n-gram-based approach has low accuracy and tends to predict the most common verb. For a translation system, this often degenerates into the less interesting problem of whether to trust whether the final verb is indeed a common one. While this improves translation delay, better predictions will lead to more significant improvements. We instead opt for a one-vs-all discriminative classification approach.[6]

3.1 Classification on Human Data

We first incrementally classify verbs on the same 200 sentences from Section 2. Since the answer choices are often complex verb *bunsetsu* and since many of these verb phrase answer choices do not appear among the most common verbs, lemmatizing the verbs and performing one-vs-all classification yields accuracy close to random chance. Thus, we use binary classification with a single linear classifier to produce a probability for each candidate answer, encoding the verb phrase itself into the feature vector.

3.1.1 Training a Morphological Model

The processing is as follows: We train on 463,716 verb-final sentences extracted from the training data. We use both **context features** and **final verb features**. Our context features, i.e., those preceding the final verb, are represented as follows: the context **unigrams** and **bigrams** take a value of 1

[5] While we use logistic regression, using hinge loss achieves similar accuracy.

[6] One-vs-all classification builds a classifier for each class versus the aggregate all other classes.

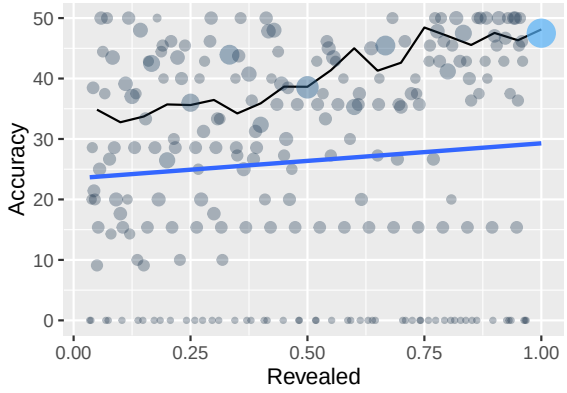

Figure 4: Verb classification results on crowdsourced sentences. Despite many out-of-vocabulary items and significant noise, the average accuracy, shown in the non-monotonic line in the plot, increases over the course of the sentence. Larger, darker circles indicate more examples for a given position. Accuracy was calculated by aggregating the guesses at 5% intervals.

if they are present and 0 otherwise; **case markers** observed in the sentence context are represented as unigrams and bigrams in the order that they appear; and we reserve a distinct feature for the **last observed case marker** in the sentence. Our **verb features** consist of the final verb tokens given by the morphological analyzer, which, in addition to the verb stem itself, typically include tense and aspect information. These are represented as unigrams and bigrams in the feature vector.

To allow the classifier to learn, we must encode the interactions between the verb features and the context features. Thus, we use the Cartesian product of sentence and verb features to encode interactions between them: each training sentence generates both a positive and a negative example. The example with the correct verb phrase is labeled as a positive example ($+1$), and we uniformly select a random verb phrase from one of the 500 most common verb phrases and label it as negative (-1) example for the same sentence context,[7] yielding 927,432 training examples and 267,037,571 features.

For clarity, we describe this feature representation more formally. Given sentence S_t with a pre-verb context consisting of unigrams, bigrams, and case marker tokens, $C = \{c_0, ..., c_n\}$, and *bunsetsu* verb phrase tokens $A = \{a_0, ..., a_k\}$, the feature vector consists of $C \times A = \{c_0 \wedge a_0, c_0 \wedge a_1, ..., c_n \wedge a_k\}$, where $\wedge$ concatenates the two context and answer strings. During learning, the weights learned for the concatenated tokens are thus based on the relationship between a context token and a *bunsetsu* token and mapped to $\{+1, -1\}$. More concretely, individual morphemes of the Japanese verb phrase are combined with the pre-verb unigrams, bigrams, and uniquely identified case marker tokens. Accuracy improves when the morphemes used in the negative examples and positive examples are disjoint; so, we enforce this constraint when selecting negative examples. For example, if the positive example includes the past tense morpheme (た), the negative example is disallowed from having this morpheme altogether.

[7] We experimented with several numbers of weighted negative examples and found that one negative example with of equal weight to the positive gave the best results of the configurations we tried.

3.1.2 Choosing an Answer

At test time, we test progressively longer fragments of each sentence, extracting the aforementioned features online until the entire pre-verb context is available. For every sentence fragment, the classifier determines the probability of each of the four possible verbs by adding their verb features to the feature vector of the example. The answer choice with the highest probability of $+1$ (or the lowest probability of -1) is chosen as the answer. By taking this approach, we can model complex verbs and their context jointly. Intuitively, the probability of a ($+1$) is the model's prediction of how well the *bunsetsu* verb phrase fits in with the sentence context (represented by the feature vector).

Some verbs are absent from the training data, forcing the classifier to rely on morphemes to distinguish between them. The alternative—e.g., in a typical one-vs-all classification approach—is that the classifier could reason from nothing whatsoever when a fully-inflected verb is absent from the training data. Given the complexity of *bunsetsu*, this happens often even in large corpora for a language such as Japanese.

3.1.3 Multiple Choice Results

Despite only choosing among four choices, this task is in many ways more difficult than the 50-label classification problem described in the next section because of the added complexity inherent modeling the effect of morphemes and missing examples. These limitations notwithstanding, the

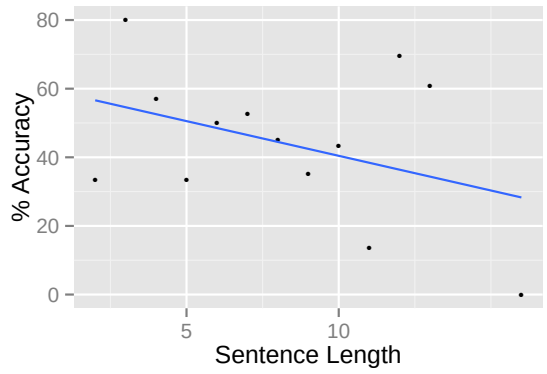

Figure 5: Classification accuracy as a function of sentence length on the full context set. The shortest sentences have higher accuracy, though the trend is less clear for sentences of medium length. Compare to Figure 1.

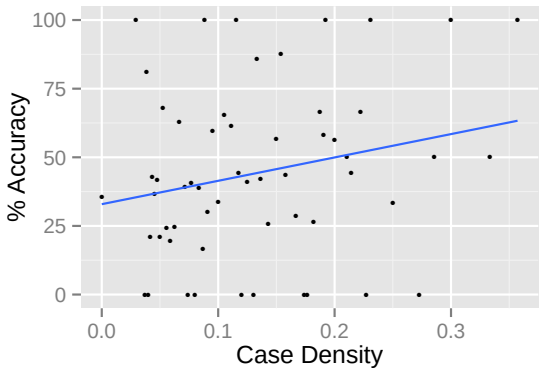

Figure 6: Classification accuracy as a function of case density on the incremental sentences. The accuracy is correlated with case density, but the data are extremely noisy. Full-context accuracy shows a similar trend (not shown).

accuracy does improve as more of the sentence is revealed (Figure 4), indicating that the algorithm learns to use these features to rank verbs, though the accuracy significantly lags that of both the human participants and our later experiments. Additionally, on the **full context set**, sentence length is slightly negatively correlated with accuracy (Figure 5), as in the much more convincing results of our human experiments (Figure 1), though the trend is not entirely consistent,, making it difficult to draw firm conclusions. Case density is again positively correlated with accuracy on both the random (Figure 6) and full context sets.

An Illustrative Example To gain some insight into how features can influence the classifier, we here examine an example of the classifier's behavior on the multiple choice data.

(5) 少年時代-は 熊本藩-の
childhood days-TOP Kumamoto domain-GEN
藩校-で 儒学-を
clan school-LOC Confucianism-ACC
学び、 後-に
study:MED subsequently-LOC
西本願寺-において 修行-に
Nishihongan Temple-LOC discipline-ALL

(6) (a) 励ん-だ
strive-PAST
(b) 創刊-さ-れ-る
issue-do-PASS-NPST
(c) 加え-られ-てい-る
add-PASS-CONT-NPST
(d) 勤め-る
serve-NPST

In Example (5), the classifier incorrectly chooses "issue" as the verb until observing the accusative case marker attached to "Confucianism". At this point, the classifier's confidence in the correct answer rises to 0.74—and correctly chooses "strive". This answer goes unchanged for the remainder of the sentence, although "study" attaches to "Confucianism", not the final verb. The combined evidence, however, is enough for the classifier to select correctly, and indeed, most of the following tokens only increase the classifier's confidence. Adding "subsequently" increases confidence to 0.84, an intuitive increase given the likely tense information contained in such a word. The redundant case marker in this case only increases confidence to 0.86. Adding the reference to the temple decreases confidence again to 0.79. Adding the **final case marker** results in a huge increase in confidence, to 0.90.

3.2 Multiclass Verb Prediction

While the multiple choice experiment was more open-ended (predicting random verbs), we now focus on a more constrained task: how well can we predict the most frequent verbs. This is the central conceit of Grissom II et al. (2014): if you can do a good job of this, you can improve simultane-

ous translation. They show a slight improvement in simultaneous translation by using n-gram language model-based verb prediction. We show a large improvement over their approach to verb prediction using a discriminative multiclass logistic classifier (Langford et al., 2007).

Data Preparation Our classes for multiclass classification are the fifty most common verbs in the KFT (Japanese, as in the human study) and Wortschatz corpora (Biemann et al., 2007, German).

We use data from the training and test sets of the KFT Japanese corpus of Wikipedia articles and a random split of the German Wortschatz web corpus, from which we extract the verb-final sentences. Grissom II et al. (2014) use an n-gram model to distinguish between the fifty most common German verbs for SOV-SVO simultaneous machine translation, which we replicate as our baseline. Following this study, we train a model on the fifty most common verbs in the training set.

In Japanese, due to the small size of the standard test set, we split the data randomly, training on 60,926 verb-final sentences ending in the top fifty verbs and testing on 1,932. Our total feature count is 4,649,055. We use the MeCab (Kudo, 2005) morphological analyzer for segmentation and verb identification. We consider only verb-final sentences. We skip semantically vacuous post-verbal copulas when identifying final verbs.

Finding Verbs We identify verbs in the German text with a part-of-speech tagger (Toutanova et al., 2003) and select from the top fifty verbs. We consider the sentence-ending set of verbs to be the final verbs. We train on 76,209 verb-final sentences ending in the top fifty verbs and test on 9,386. In German, to approximate the case information that we extract in Japanese, we test the inclusion of equivalent unigram and bigram features for German **articles**, the surface forms of which determine the case of the next noun phrase.

In Japanese, we omit some special cases of light verbs that combine with other verbs, as well as ambiguous cases and copulas.[8]

Features All features are encoded as binary features indicating their presence or absence. For Japanese, we again include **case unigrams**, and **case bigrams**, which encode as distinct features the for case markers observed thus far.[9] We also include a feature for the **last observed case marker**. For both Japanese and German, we normalize the verbs to the non-past, plain form, both providing more training data for each verb and simplifying the job of our classifier.

German case is conveyed primarily through articles and pronouns, so we include special features for articles. For example, for the sentence "Es wurde ihnen von einem alten Freund geholfen", we add the features `ART_es_ihnen` and `ART_ihnen_einem` to convey case information beyond individual words and bigrams.

Individual tokens are also used as binary features, as well as token bigrams.

An Example for Every Word In a simultaneous interpretation a person or algorithm receives a constant stream of words, and each new word provides new information that can aid in prediction. Previous predictive approaches to simultaneous machine interpretation have taken this approach, and we also use it here: as each new word is observed, we make a prediction. This is a generalization of *random* presentation of prefixes in the human study.

3.3 Classification Results and Discussion

Better at the End A discriminative classifier does better than an n-gram classifier, which has a tendency to over-predict frequent verbs. By the end of the sentence, accuracy reaches 39.9% for German (Figure 7) and 29.9% Japanese (Figure 8), greatly exceeding choosing the most frequent class baseline of 3.7% (German) and 6.05% (Japanese). The n-gram language model also outperforms this baseline, but not by much. It also improves over the course of the sentence, but the model cannot reliably predict more than a handful of verbs in either language.

Richer Features Help (Mostly at the End) Bigram features help both languages, but Japanese

[8] In Japanese, we omit some ambiguous cases and variants of "is" and "do": excluded are variants of *suru* ("to do"), which combines with nouns to form new verbs, *aru* ("is", inanimate case), and *iru* ("is", animate case). The tokens *aru* and *iru* also combine with other verbs to change tense and aspect, in which case they are not verbs, and can form the copula *de aru*. Distinguishing between all of these cases is beyond the scope of this study; so, they are excluded. We also omit duplicates that are spelled differently (i.e., the same word but spelled without Chinese (*kanji*) characters and slightly different forms of the same root).

We also omit the light verb *naru* ("to become" or "to make up") for similar reasons to *suru*. The increasing trend shown in the results does not change with their inclusion.

[9] For instance, given a sentence fragment X-に Y-を, representing X-DAT Y-ACC, the case bigram would be に∧を.

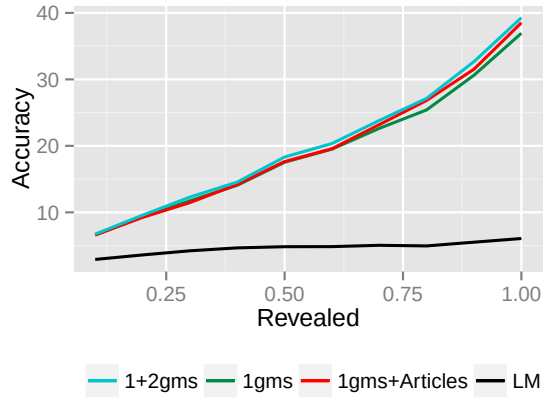

Figure 7: German average prediction accuracy over the course of sentences. Bigrams help slightly in the second half of the sentence. Adding special features for case-assigning articles to unigrams nearly matches the performance of adding all bigrams in the final 10%. All handily outperform the trigram language model.

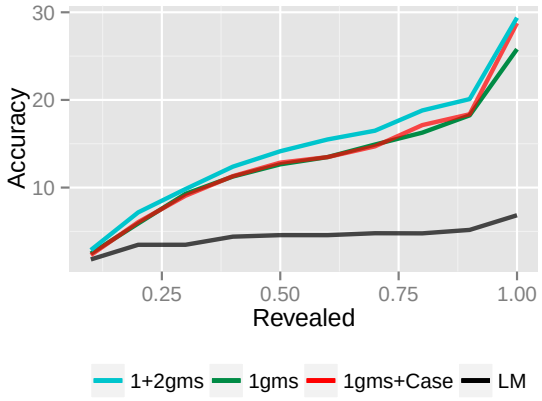

Figure 8: Japanese average prediction accuracy over the course of sentences. Adding bigrams consistently outperforms unigrams alone in Japanese, possibly due to the agglutinative nature of the language. The accuracies diverge the most toward the end of the sentences: Adding only explicit case markers to unigrams nearly matches performance of adding all bigrams toward the end. All outperform the trigram language model.

more than German; beyond bigrams, however, trigram and longer features overfit the training data and hurt performance. The better performance for Japanese bigrams is likely because word boundaries are not well-defined in Japanese, and individual morphemes can combine in ways that significantly add information. German word boundaries are more precise and words (particularly nouns) can carry substantial information themselves.

Richer features matter more toward the end of the sentence. In Japanese, adding bigrams consistently outperforms unigrams alone, but in both languages, adding special features for tokens with case information helps almost as much as adding the full set of bigrams. In Japanese, case markings always immediately follow the words marked, and in German the articles precede the nouns to which they assign case; thus, rather than relying on isolated unigrams, using bigrams provides opportunities to encode case-marked words that more narrowly select for verbs. In Japanese, the differences are more pronounced toward the very end of the sentences (and less so in German).

Richer features help more at the end not just because the last words of the sentence represent the densest feature vectors. In Japanese, the last word is usually a case-marked noun phrase or adverb that matches the final predicate. The final word is therefore immune to subclause interference and must modify the final verb, boosting the classifier performance in these final positions and amplifying the predictive discrepancies between the various feature sets. Accuracy spikes at the end of Japanese sentences, where case information helps nearly as much as adding the entire set of bigrams, further supporting case information's importance. Deeper processing—e.g., separating case-marked words in subclauses from those in the main clause—would likely be more useful. Features and feature-selection strategies that we tried which did not help included the following: adding case marker unigrams; using only case-marked words; only allowing one word per case marker in the feature vector (the most recent); adding part-of-speech tag n-grams; and adding the word nearest to the centroid of the observed context in a word embedding space.

4 Related Work

While to our knowledge our work is the first in-depth study of incremental verb prediction, it is not the first study of verb prediction in humans or machines. This section reviews that related work.

Human Verb Prediction Prediction is easier with more context and explicit case markings. Teramura (1987) shows that *next word prediction* in Japanese improves as more words are incrementally revealed. While only looking at verb prediction given the *complete* preceding context, Yamashita (1997) finds that scrambling word order in Japanese—a case rich language that allows such scrambling—does not harm final verb prediction, but that explicit case marking helps final verb prediction. Our results show that this is true even for incremental verb prediction. Levy and Keller (2013) also find that dative markers aid German verb prediction.

Neurolinguistic measurements by Friederici and Frisch (2000) suggest processing verb-final clauses in German use both semantic and syntactic information, but that they are processed differently. In Japanese, Koso et al. (2011) measure the effect of case markings on predicting verbs with strong case preferences. This is consistent with our use of case-based features and suggests that further gains are possible using richer syntactic representations. **?**) use N400 measurements to investigate two competing hypotheses for the initial prediction of an upcoming verb: whether predictions are dependent on all words equally (the Bag-of-words hypothesis), or alternatively, whether prediction is selectively modulated by the final verb's arguments (the Bag-of-arguments hypothesis). They argue for the latter.

The literature on incremental verb prediction is sparse. A key finding of Matsubara et al. (2002) is that Japanese-English simultaneous interpreters, when given access to lecture slides, would refer to them to predict the next phrase.

Prediction for Simultaneous Machine Translation The Verbmobil simultaneous translation system (Kay et al., 1992) uses deleted interpolation (Jelinek, 1990) to create a weighted n-gram models to predict dialogue acts—almost identical to predicting the next word (Reithinger et al., 1996). Konieczny and Döring (2003) predict verbs with a recurrent neural network, but Matsubara et al. (2000) was the first to use verb predictions as part of a simultaneous interpretation system. They use pattern matching-based predictions of English verbs. In contrast, Grissom II et al. (2014) use a statistical approach, using n-gram models to predict German verbs and particles (in Section 3 we show that this model predicts verbs poorly). However, their simultaneous translation system is able to learn when to trust these predictions. Oda et al. (2015) extend the idea of using prediction by predicting entire syntactic constituents for English-Japanese simultaneous machine translation. Both systems will likely benefit from our improved verb prediction presented here.

5 Conclusion

Verb prediction is hard for both machines and humans, but it is impossible for neither. Verbs become more predictable in discriminative settings as more of the sentence is revealed, and when all of the prior context is available, the verbs are highly predictable by humans when a limited number of choices is available, though even then not perfectly so. While we make no claims concerning upper or lower bounds of predictability in different settings, our dataset provides benchmarks for future verb prediction research on publicly available corpora: cognitive scientists can validate prediction, confusion, and anticipation; engineers have a human benchmark for their systems; and linguists can conduct future experiments on predictability. Shallow features can be used to predict verbs more accurately with more context. Improving verb prediction can benefit simultaneous translations systems that have already shown to benefit from verb predictions, as well as enable new applications that involve predicting future linguistic input.

6 Acknowledgments

We would like to thank the anonymous reviewers for their comments. We thank Yusuke Miyao for his helpful support. We would also like to thank James H. Martin, Martha Palmer, Hal Daumé III, Mans Hulden, Mohit Iyyer, John Morgan, Shota Momma, Graham Neubig, and Sho Hoshino for their invaluable discussions and input. This work was supported by NSF grant IIS-1320538. Boyd-Graber is also partially supported by NSF grants CCF-1409287 and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistics*.

Yasuhara Den and Masakatsu Inoue. 1997. Disambiguation with verb-predictability: Evidence from Japanese garden-path phenomena. In *Proceedings of the Cognitive Science Society*, pages 179–184. Lawrence Erlbaum.

Angela D Friederici and Stefan Frisch. 2000. Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language*, 43(3):476–507.

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 *Empirical Methods in Natural Language Processing*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Fred Jelinek. 1990. Self-organized language modeling for speech recognition. *Readings in speech recognition*, pages 450–506.

Martin Kay, Peter Norvig, and Mark Gawron. 1992. *Verbmobil: A translation system for face-to-face dialog*. University of Chicago Press.

Lars Konieczny and Philipp Döring. 2003. Anticipation of clause-final heads: Evidence from eye-tracking and srns. In *Proceedings of iccs/ascs*.

Ayumi Koso, Shiro Ojima, and Hiroko Hagiwara. 2011. An event-related potential investigation of lexical pitch-accent processing in auditory Japanese. *Brain research*, 1385:217–228.

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

Sadao Kurohashi and Makoto Nagao. 1994. Kn parser: Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*.

Marta Kutas, Katherine A DeLong, and Nathaniel J Smith. 2011. A look around at what lies ahead: prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future*, pages 190–207.

John Langford, Lihong Li, and Alex Strehl. 2007. Vowpal wabbit online learning project.

Roger P Levy and Frank Keller. 2013. Expectation and locality effects in german verb-final structures. *Journal of memory and language*, 68(2):199–222.

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*.

Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2002. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *LREC*.

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*.

Graham Neubig. 2011. The Kyoto free translation task. *Available online at http://www. phontron. com/kftt*.

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*, June.

Norbert Reithinger, Ralf Engel, Michael Kipp, and Martin Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. volume 2, pages 654–657. IEEE.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656.

Hideo Teramura. 1987. Kikitori ni okeru yosoku nouryoku to bunpouteki tisiki. *Nihongogaku*, 3;6:56–68.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 173–180.

Hiroko Yamashita. 1997. The effects of word-order and case marking information on the processing of Japanese. *Journal of Psycholinguistic Research*, 26(2):163–188.

Hiroko Yamashita. 2000. Structural computation and the role of morphological markings in the processing of japanese. *Language and speech*, 43(4):429–455.