Exploring Cognitive Plausibility of Neural NLP Models: Cross-Linguistic and Discourse-Level Studies

ニューラル自然言語処理モデルの認知的妥当性:
言語横断的分析と談話処理

Tatsuki Kuribayashi

Graduate School of Information Sciences
Tohoku University

This dissertation is submitted for the degree of
Doctor of Information Science

January 2022

Acknowledgements

First of all, I would like to appreciate my dissertation committee members: Prof. Kentaro Inui, Prof. Jun Suzuki, Prof. Mitsuyuki Nakao, and Prof. Akinori Ito. My supervisor, Prof. Kentaro Inui, provided me with valuable research advice and taught me a lot about the general way of thinking. I would also like to express my sincere gratitude to Prof. Jun Suzuki for his precise advice to my research and paper writing.

In addition to the committee members, Dr. Yohei Oseki (the University of Tokyo) provided me with the opportunity to learn about NLP and a wide range of related fields such as psycholinguistics, as well as sharp feedback on my research. I would like to thank Dr. Hiroki Ouchi (Nara Institute of Science and Technology), Dr. Naoya Inoue (Stony Brook University), and Dr. Paul Reisert for their valuable advice, especially when I was just getting started with my research. I am also grateful to Prof. Masayuki Asahara (National Institute for Japanese Language and Linguistics) for his valuable help in handling corpus data and research feedback.

The members of the Inui Laboratory at Tohoku University have supported me in my daily research life. In particular, I am grateful to Ana Brassard and Takumi Ito for their long discussions with me especially when I was compiling my research. Furthermore, I would like to express my gratitude to Sugawara-san, Aizawa-san, and Isobe-san for their help with the office work related to the research.

I would also like to express my gratitude to the people at Langsmith Inc. for allowing me to immerse myself in the research. I am also grateful to the JASSO Scholarship, the JSPS Research Fellowship for Young Scientists, and the Tohoku University International Joint Graduate School of Data Science for financial support.

In addition, we thank the members of the Tohoku University Wind Orchestra (especially, my contemporaries). They gave me a refreshing break in my research life.

Finally, I would like to express my utmost gratitude to my family for their warm support.

Abstract

Constructing a model mimicking human language processing is a central goal in cognitive science of language. In recent years, computational systems for handling natural language have made considerable progress in the natural language processing (NLP) field; exploring their cognitive plausibility is a necessary step to exchange the engineering and cognitive perspectives of linguistic research. With this in mind, this study explores the similarities and discrepancies between machine and human language processing. Specifically, this thesis tackles the following questions: *are modern neural-based models close to "human-like" computational models*, and *how could we make these models behave more human-like?*

In particular, we expand the scope of cognitive plausibility analysis in neural NLP models in two directions: (i) cross-linguistic, and (ii) discourse-level studies. Current cognitively-motivated NLP studies have typically focused on a limited scope, such as English sentence processing. While narrowing the focus and deepening the understanding are important steps, broadening the scope could complement the studies in terms of exploring the generality of the findings and incorporating multiple perspectives into the analysis. For the cross-language analysis, this study incorporates Japanese as a representative of languages that are typologically different from English into the analysis. We discovered a surprising asymmetry between the results of these languages and investigated the source of this discrepancy.

To widen the targeted linguistic phenomena, we focus on discourse processing—intepreting the meaning of text beyond a single sentence level. The NLP fields struggle with handling discourse with computers, for example, existing neural discourse parsing models exhibit cognitively implausible behaviors such as predicting shallow, near-linear discourse structures although human annotators determine deeper hierarchical trees. Using classical psychological theory as a hint, we designed an inductive bias (neural architecture design) for effective neural discourse processing model and guided hierarchical generalization in discourse parsing.

To summarize, this study clarified the (dis)similarity of language processing by humans and machines, such as disconnection between the linguistic accuracy of language models and their cognitive plausibility. We provide several implications for bridging the gap between them in terms of, for example, memory limitations and architectural bias.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The natural language processing (NLP) field, a branch of artificial intelligence, makes computers handle natural language. Recent progress of this field was remarkable; reaping the development of deep learning, NLP systems have achieved human-level scores in various benchmark tasks (Kiela et al., 2021), and NLP-powered applications (e.g., machine translation) have become more ubiquitous. Does this mean that NLP has been "getting solved?" To begin with, we briefly look back at the recent progress of NLP and frame the goal of this thesis, aligning with the traditional goals in artificial intelligence.

Referring to the encyclopedia of computer science (Shapiro, 2003), the goals of artificial intelligence are categorized into three:

I. Computational Psychology: *understand human intelligent behavior by creating computer programs that behave in the same way that people do.*

II. Computational Philosophy: *form a computational understanding of human-level intelligent behavior, without being restricted to the algorithms and data structures that the human mind actually does ("Is intelligence a computable function?").*

III. Machine Intelligence: *program computers to do what, until recently, only people could do.*

As for the II. computational philosophy, the field would have advanced considering that the performance of NLP models has been getting closer to human-level although it is unclear whether they perfectly mimic human-like processing. As for the III. machine intelligence, the field could also be progressed considering recent many useful NLP-powered applications (e.g., machine translation systems). In contrast, what about I. computational psychology? How well do recent neural NLP systems exactly mimic human language processing (cognitively plausible)? Has this neural-based NLP advanced our understanding of

humans and language? Actually, such a direction has been actively explored in computational psycholinguistics (Crocker, 2010), but we would like to emphasize that the research from the perspectives of computational psychology is still limited compared to the majority of engineering-oriented studies in NLP (e.g., aiming for a high score on a specific downstream task). In particular, the scope of such cognitively-motivated NLP studies is typically limited to, for example, English sentence processing (Marvin and Linzen, 2018; Warstadt et al., 2020a). Of course, narrowing the focus and deepening the understanding of specific phenomena is important, but widening the scope of analysis can also play a complementary role to explore the generality of the findings and incorporate multiple perspectives into the analysis.

In this thesis, we set a goal of constructing a human-like language processing model and expand the scope of cognitive plausibility analysis in the neural NLP models in two directions: (i) cross-linguistic and (ii) discourse-level studies. Regarding cross-language analysis, we are concerned that narrowing down the target languages might lead to biased conclusions. In psycholinguistic studies, cross-linguistic validation of linguistic theories has led to more sophisticated conclusions (e.g., expectation-based theory v.s. memory-based theory in sentence processing). In this study, the Japanese language is used as a representative of the language typologically differing from the English language and explored human-like computational models across languages. We observed an unexpected asymmetry in the validity of existing psycholinguistic findings and explored the source of this gap.

In terms of widening the targeted linguistic properties, we focus on discourse processing—understanding the meaning of text beyond a single sentence level. Discourse is indispensable in language communication; however, the integration of neural-based NLP and the computational psycholinguistics perspectives at the discourse level is limited in the current research community. Considering the cognitively implausible behaviors of existing discourse parsing models that tend to predict near-linear discourse structures, we design an effective inductive bias (e.g., neural architecture) for the neural discourse parsing model to lead to capture the hierarchical nature of discourse structures.

1.1 Research issues and thesis overview

We concretely focus on the following topics:

1. **Chapter 3: Sentence acceptability judgment by neural language models in flexible word order language.**
   From Chapters 3 to 5, we compared the sentence processing by neural language models (LMs) and humans. As the first requirement for a human-like model, LMs' human-like

sentence-level acceptability judgment (i.e., word order preference) was analyzed. Using controlled materials that have been exploited in Japanese word order research, we have profiled the linguistic preference of LMs from multiple linguistic perspectives.

2. **Chapter 4: Cross-linguistic (Japanese and English) analysis of *incremental* sentence processing by humans and neural language models.**
   While Chapter 3 analyzed inter-sentential preferences of LMs, in this chapter, we shift the focus into token-by-token incremental sentence processing. Based on the linking hypothesis between surprisal and human cognitive load, token-by-token surprisal computed by LMs was compared to the reading time human subjects took during reading. Our cross-linguistic studies have pointed out surprising asymmetry between languages in the similarities between human and machine incremental sentence processing.

3. **Chapter 5: Memory-based view of the gap between incremental sentence processing by humans and neural language models**
   The difference between humans and neural LMs suggested in Chapter 4 was further explored through the lens of their working memory capacity. Whereas humans have limited working memory, modern neural LMs could have excessive working memory capacity. We hypothesized that the difference between LMs and humans could stem from the difference in their working memory buffers.

4. **Chapter 6: Inductive bias facilitating discourse processing.**
   Finally, as a complementary approach to widening the target languages, we investigated *the discourse-level* (meaning of text beyond a single sentence) processing model. The NLP fields struggle with handling discourse with computers, for example, the existing neural discourse parsing model exhibits cognitively implausible behaviors such as predicting shallow, near-linear discourse structures despite the plausible structures being deep, hierarchical ones. We hypothesized that this stems from the lack of a human-like inductive bias leading the model to capture hierarchical nature of discourse. We investigated a cognitively motivated inductive bias that leads neural discourse parser to more hierarchically generalize.

1.2 Contributions

1. **Neural language models well simulate human-like Japanese word order preferences.**
   Our analysis demonstrated that the neural LMs accurately simulate human-like word

order preferences. We confirmed that their preference is consistent with that in crowd workers and quantitative linguistic theories on Japanese word order. The analysis also suggests that the human-like word order preference and next-word prediction performance of LMs are highly correlated.

2. **Neural language models cannot exhibit human-like behavior in modeling *incremental* reading behavior.**
   Despite the parallel in sentence-level word order preference between humans and LMs, we found a discrepancy in incremental processing difficulties exhibited by humans and neural LMs. Our cross-linguistic analysis shows that this discrepancy is language-dependent; the reading behavior of the Japanese readers is poorly simulated by accurate neural LMs. That is, better modeling language corpus does not always entail achieving human-like sentence processing.

3. **Filling the discrepancy between incremental human and language model sentence processing stems from their different working memory capacities.**
   Human working memory is limited, but modern neural LMs could have extremely powerful working memory capacity. We empirically demonstrated that this difference could lack the psychometric predictive power of neural LMs in modeling human reading behavior. In particular, the neural LMs with severely limited context access (mirroring the human-like memory limitation) accurately simulate human reading behavior in Japanese data; excessive working memory capacity could be one of the reasons for the deviation of accurate Japanese LMs from human reading behavior.

4. **Inductive bias associated with psycholinguistic theory leads the neural discourse parsing model to behave cognitively plausible discourse processing.**
   We found that neural network architecture modification associated with class discourse processing theory (schema theory) improves its discourse parsing performance. Specifically, the proposed architecture design imposes some modularity distinguishing content- and formal-level flows within a text and macro-level contextualization of clause representations on the model. While existing models tend to predict shallow, linear discourse structures, our model successfully parses hierarchical, complex discourse structures underlying argumentative texts.

Chapter 2

Background

Recently, there has been a successful exchange between neural-based NLP studies and the cognitive science of language. We briefly review these interdisciplinary studies.

2.1 Why neural NLP is exploited for linguistic studies

Since the 2010s, deep learning has significantly impacted various fields such as computer vision and NLP. Neural-based NLP models have outperformed classical methods in various tasks; the NLP fields get enthusiastic about successfully applying the neural networks to language tasks (Goldberg, 2017). Along with engineering studies, neural NLP models were also exploited in cognitively-motivated studies. Such use of neural models are mainly categorized into two:

Neural NLP for computing complexity metrics Constructing a model that simulates human language processing is an important goal from the cognitive science viewpoint. In general, theory needs quantitative evidence ensuring its validity; researchers need to implement a model representing the theory and evaluate their predictive power of human behavior data. In other words, if a theory cannot be technically implemented, proving its validity is prohibitively difficult.

Recent neural NLP models seem to better handle language than classic methods and broadened the possibilities of linguistic theorizing. In particular, neural LMs enable the computation of various information-theoretic complexity metrics (e.g., next-word prediction probabilities) more accurately than classic approaches. For example, it is reported that surprisal ($-\log p(\mathrm{word}|\mathrm{context})$) computed by more accurate neural LMs could better simulate human reading behavior data, which supports the surprisal theory of human incremental

syntactic comprehension Levy (2008); Smith and Levy (2013). That is, human sentence processing is tuned to language statistics; humans predict the next word during reading, and if the prediction is wrong, the processing load increases.

More broadly, the search for models that simulate human language processing can be positioned to the cognitive science of language. Using the terminology of Marr (1982), the "understanding" of information processing model is categorized into three levels as follows:

- Computational level — *What is the goal of the computation, and why is it appropriate?*
- Algorithmic level — *What is the representation for the input and output, and what is the algorithm for the transformation?*
- Implementation level — *How can the representation and algorithm be realized physically?*

Cognitive science emphasized the need of understanding information processing models from multiple levels, especially, higher ones (*computational* or *algorithmic* levels). Exploring the cognitive plausibility of modern neural LMs could be aligned with this context. For example, the finding that the human reading behavior (e.g., reading time) is well modeled by the next-word prediction probability (Clark, 2013; Levy, 2008; Smith and Levy, 2013) could be at the computational-level understanding—*what do humans compute during reading?* Comparing the neural model architecture in terms of their psychometric predictive power could also be at the computational and/or algorithmic level—*e.g., what algorithm or constraints are adopted to achieve the computational goal?* From Chapters 3 to 5, we pursue the computational- and algorithmic-level understanding of human sentence processing through constructing models that could well simulate human reading behavior data.

Neural NLP for *in silico* simulation of language learning Is language acquired only from exposure to external language? If not, what innate knowledge (inductive bias) in humans facilitates language acquisition? Such questions attract huge scientific interest; for example, the argument from the poverty of the stimulus implies that humans innately have some strong inductive biases facilitating language acquisition (Chomsky, 1980). One unrealistic approach for investigating such biases could be to create humans (infants) while controlling their inductive biases and compare their generalization capability during language acquisition. However, such studies are prohibitively difficult from the ethical and technical perspectives.

An alternative approach could be using some simulators that can emulate language acquisition (McCoy et al., 2018). For example, if some simulators without any prior linguistic

knowledge acquire language from the data alone, it would provide sufficient conditions for language acquisition; language could be acquired empirically. A neural network is not designed to have language-specific prior; one typical approach is to imitate the neural nets as the simulator of "blank slate" language learners without any prior knowledge of the language. Their language learning ability could suggest what aspect of language can be learned from the text only. Notably, the way of creating such a blank slate simulator was not obvious in the previous NLP era (e.g., statistical NLP), where researchers have to manually design linguistic features (priors) in advance and then train a model. Some studies using neural models provide interesting suggestions on inductive bias successfully leading to linguistic generalization; models that merely see the language sequences fail to perform hierarchical generalization (McCoy et al., 2020), a huge amount of texts were needed to make LMs prefer linguistically plausible generalization against exploiting shallow cues (Warstadt et al., 2020b), and particular syntactic properties could be acquired by neural models under specific conditions without explicit syntactic supervision (Wei et al., 2021a).

Note that the language learnability of neural networks has been historically discussed especially in the Morphology fields. A neural network is inspired by the psychological theory of parallel distributed processing (Rumelhart et al., 1988), and their cognitive plausibility was tested in the English inflection (e.g., present→past form transformation) task (Pinker and Prince, 1988; Rumelhart and McClelland, 1986). This "past-tense debate" is still running in the NLP field (Kirov and Cotterell, 2018; McCurdy et al., 2020).

Furthermore, classical studies suggested that the human inductive bias is related to the connection of neurons in the brain (Elman et al., 1996). Thus cognitively-motivated NLP studies are also interested in the effective "wiring" of neurons in neural models (i.e., neural model architecture design). For example, it is suggested that integrating hierarchical modules (humans might have) into neural LM architecture improves language processing (Dyer et al., 2016; Hale et al., 2018; Yoshida et al., 2021). In the line with this, some claim that it is inappropriate to consider neural networks to be "blank slates" altogether and that researchers should consider what inductive biases each architecture (e.g., Transformer (Vaswani et al., 2017b)) reflects (Baroni, 2021; Kharitonov and Chaabouni, 2021). In Chapter 6, we investigate psychologically-motivated effective inductive bias for achieving discourse (intersentential) processing.

2.2 Probing linguistic knowledge in neural NLP models

Our studies could also be related to so-called probing studies investigating linguistic knowledge in neural NLP models. A classic paradigm in the NLP is to manually design linguistic

features and train/implement a model using these features. Conversely, the modern neural-based approach is to train a neural network without explicit language prior, and then researchers inspect what they know about and how they use linguistic features after training. There are diverse methods (e.g., observing their behavior, supervised probing) and targets (e.g., grammatical knowledge, commonsense) for such analyses (Belinkov et al., 2020; Belinkov and Glass, 2019; Rogers et al., 2020). One typical direction related to this thesis is to probe the syntactic knowledge of neural-based models via observing their behavior (e.g., generation probability) using controlled materials (Marvin and Linzen, 2018; Warstadt et al., 2020a, 2019). Although whether such neural models acquire proper syntactic knowledge is still controversial, some research found non-trivial clues that the models might have a good sense of learning languages (e.g., successfully achieve the generalization about subject-verb agreement (Wei et al., 2021a)). As a complementary direction to analyze the discrete grammatical decision by these models, Chapter 3 investigates nuanced, soft linguistic preferences in LMs through the lens of Japanese word order.

2.3 Engineering view for cognitive plausibility analysis

Finally, data on human language processing behavior and/or cognitive view have provided guidelines for NLP systems/studies to follow. Such directions are investigated, for example, in evaluating word embeddings (Hollenstein et al., 2019), language models (Ettinger, 2020; Misra et al., 2020; Upadhye et al., 2020), or reading comprehension studies (Sugawara et al., 2021). Furthermore, the use of cognitive features (e.g., eye-tracking data) to improve NLP systems has been investigated (Mathias et al., 2020; Vickers et al., 2021). Exploiting human behavior data improves NLP systems in tasks such as sequence labeling (Barrett et al., 2016; Barrett and Hollenstein, 2020; Klerke and Plank, 2019), named entity recognition (Hollenstein and Zhang, 2019), and sentiment analysis (Mishra et al., 2016). It is also explored to use eye-tracking information for evaluating annotator's behavior (Mitsuda et al., 2013).

Additionally, simulating human language processing could be useful in human-centric applications. For example, developing the model for automatic (incremental) text readability evaluation considerably overlaps the computational cognitive modeling studies that aim to simulate the cognitive effort exhibited by readers.

Chapter 3

Sentence Acceptability Judgments by Neural Language Models in Flexible Word Order Language

3.1 Introduction

Neural language models have achieved progressive performance in downstream tasks. However, what these models exactly know and how they use language is unclear. This study investigates how these models behave like humans against sentence acceptability judgement and are consistent with linguistic reports. Typical analysis probes their linguistic knowledge of discrete, strict syntactic rules (e.g., subject-verb agreement) required in English sentence processing. Complementary to these analyses, this study explores whether neural LMs have nuanced, soft linguistic preferences exhibited by humans. For example, speakers sometimes have a range of options for word order in conveying similar meaning. A typical case in English is dative alternation:

(1) a. *A teacher gave a student a book.*

b. *A teacher gave a book to a student.*

In this study, we investigate whether such human-like word order preference is replicated by neural LMs even in the language with flexible word order. We specifically focus on the Japanese language, where word order is a matter of human abstract preference; the Japanese language has less strict rules in word ordering except placing the verb at the end of the sentence (Tsujimura, 2013).

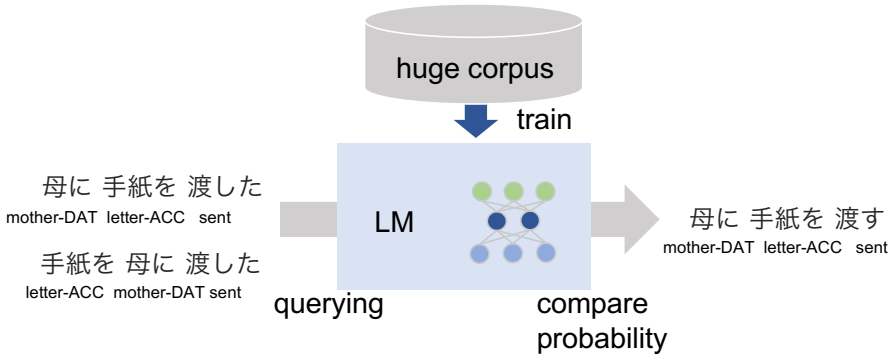


Fig. 3.1 Evaluating word order preferences of neural LMs.

The word order preference analysis in the Japanese language could contribute to three directions. First, as we mentioned, our analysis could profile the linguistic capacity of LMs. Compared to the task of acceptability rating that is not exactly based on a clear linguistic perspective (Lau et al., 2017) (i.e., ratings of the sentence pairs generated by machine translation systems), our controlled analyses tied to linguistic theories on word order could facilitate determining which linguistic aspects LMs could capture.

Second, our analysis could contribute to the linguistic studies of Japanese word order. If LMs could successfully model the human-like word order preference, this implies that the large part of Japanese word order preference could be explained by the statistical property of language. Our results demonstrate that neural LMs exhibit surprisingly well word order preference, and support the dominance of statistical accounts explaining the human preference.

Third, our analysis supports the application use of LMs for evaluating Japanese word order fluency. Especially for non-native speakers of Japanese, it is not easy to learn the natural word order of Japanese. Such automatic evaluation could potentially be useful for helping these learners. In addition, although LMs are widely used in querying the plausibility of texts (e.g., post-re-ranking in the translation task), it is not clear what aspects of plausibility LMs could accurately estimate. Our study separately tested the ability of LMs in choosing plausible word order and demonstrated that they exhibit proper preference against word order variations.

3.2 Background

3.2.1 Word order preference

There has been a significant linguistic effort to reveal the factors determining the word order preference in humans (Bresnan et al., 2007; Hoji, 1985). A typical methodology for testing the theories on word order is to observe the human reactions against several word order (Bahlmann et al., 2007; Shigenaga, 2014) or analyze a large corpus (Sasano and Okumura, 2016).

The motivations for modeling the word order preference range from linguistic interests to those involved in various other fields—it relates to language acquisition and production in psycholinguistics (Akhtar, 1999; Slobin and Bever, 1982), second language education (Alonso Belmonte et al., 2000), and natural language generation (Visweswariah et al., 2011) or error correction (Cheng et al., 2014) in NLP. In Japanese, there are also many studies on its word order (Hoji, 1985; Koizumi and Tamaoka, 2004; Saeki, 1998; Sasano and Okumura, 2016).

The word order of Japanese is basically subject-object-verb (SOV) order, but there is no strict rule except placing the verb at the end of the sentence (Tsujimura, 2013). For example, the following three sentences have the same denotational meaning (*"A teacher gave a student a book."*):

(2) a. *Sensei-ga* *seito-ni* *hon-wo* *ageta*.
teacher-NOM student-DAT book-ACC gave.

b. *Sensei-ga* *hon-wo* *seito-ni* *ageta*.
teacher-NOM book-ACC student-DAT gave.

c. *Hon-wo* *seito-ni* *sensei-ga* *ageta*.
book-ACC student-DAT teacher-NOM gave.

This order-free nature suggests that the position of each constituent does not represent its semantic role (case). Instead, postpositional case particles indicate the roles. Table 3.1 shows typical constituents in a Japanese sentence, their canonical order, and the sections of this study where each of them is analyzed.

3.2.2 Language models

In this chapter, we used unidirectional language models. That is, given a symbol sequence $[w_1, w_2, \cdots, w_N]$ with the length of N, the model is trained to predict the upcoming symbol w_i

Table 3.1 Overview of the typical cases in Japanese, their typical case marker, and the sections where the corresponding case is analyzed. The well-known canonical word order of Japanese is listed from left to right in the table.

	Time	Location	Subject	(Adverb)	Indirect object	Direct object	Verb
Notation	TIM	LOC	NOM	-	DAT	ACC	-
Related section	3.4.3	3.4.3	3.4.3	3.4.2	3.4.1	3.4.1	3.4.1

given a preceding sub-sequence $w_{<i} = [w_1, w_2, \cdots, w_{i-1}]$. With the chain rule of probability, this model could compute the plausibility of given a sequence (e.g., sentence):

$$p(w_1, w_2, \cdots, w_N) = \prod_{i=1}^{i=N} p(w_i | w_{<i}) \ . \tag{3.1}$$

Besides, the inverse geometric mean of the next-word probabilities $p(w_i|w_{<i})$ in a text $[w_1, w_2, \cdots, w_N]$ is a typical evaluation metric for the unidirectional LMs, perplexity (PPL):

$$\text{PPL}(w_1, w_2, \cdots, w_N) = \prod_{i=0}^{N} p(w_i | w_{<i})^{-\frac{1}{N}} \ . \tag{3.2}$$

One approach to analyze the linguistic knowledge in LMs is to test whether the model assigns a higher probability to a valid (e.g., grammatical) sentence than the other Belinkov et al. (2020). This study also adopts this behavioral test to investigate the word order preference in LMs. Notably, Futrell and Levy (2019), the closest work to ours, investigated word order preference in neural LMs using English data and suggested that they have human-like preferences. This study complements their studies in language with more flexible word order and previously overlooked trends related to, for example, omission.

3.3 Experiment1: comparing human and LMs word order preference

We examine the parallels between the LMs and humans on the task of determining the plausibility of word order (Figure 3.2). First, we created data for this task (Section 3.3.1). We then compared the word order preference of LMs and that of humans (Section 3.3.2 and 3.3.3).

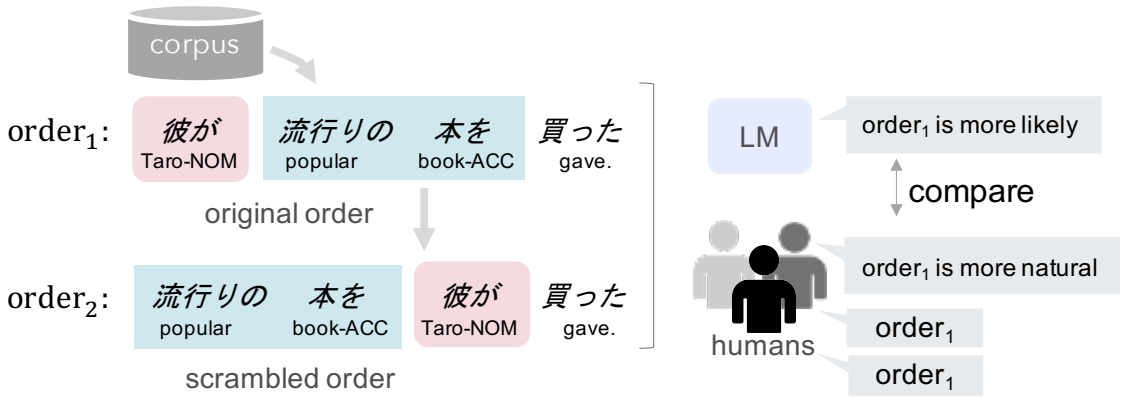


Fig. 3.2 Overview of the experiment of comparing human and LMs word order preference. First, we created data for analyzing word order preference (left part), then using this data set, we compared the preference of LMs and humans (right part).

3.3.1 Human preferences

Data We randomly collected 10k sentences from 3B web pages, which are not overlapped with the LM training data. To remove overly complex sentences, we extracted sentences that must: (i) have less than or equal to five words[1] and one verb, (ii) have words in a sibling relationship in dependency tree, and both of them accompany a particle or adverb, (iii) not have special symbols such as parentheses, and (iv) not have a backward dependency path. For each sentence, we created its scrambled version.[2] The scrambling process is as follows:

1. Identifying the dependency structure using JUMAN[3] and KNP[4].
2. Randomly selecting a word that has several syntactic dependents.
3. Shuffling the position of these dependents along with their descendants in the syntactic tree.

Crowdsourcing We used the crowdsourcing platform Yahoo Japan![5]. For our task, we showed crowd workers a pair of sentences ($order_1$, $order_2$), where one sentence has the original word order, and the other sentence has a scrambled word order.[6] Each annotator was

[1] Henceforth, "word" refers to a phrasal unit called *bunsetsu* in the Japanese language.
[2] When several scrambled versions were possible for a given sentence, we randomly selected one of them.
[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP
[5] https://crowdsourcing.yahoo.co.jp/
[6] Crowdworkers did not know which sentence was the original sentence.

instructed to label the pair with one of the following choices: (1) order_1 is better, (2) order_2 is better, or (3) the pair contains a semantically broken sentence. Only the sentences (order_1, order_2) were shown to the annotators, and they were instructed not to imagine a specific context for the sentences. We filtered unmotivated workers by using check questions.[7] For each pair instance, we employed 10 crowd workers. In total, 756 unique, motivated crowd workers participated in our task.

From the annotated data, we collected only the pairs satisfying the following conditions for our experiments: (i) none of 10 annotators determined that the pair contains a semantically broken sentence, and (ii) nine or more annotators preferred the same order. The majority decision is labeled in each pair; the task is binary classification. We collected 2.6k pair instances of sentences.

3.3.2 Preference of large-scale LMs

LMs settings: We used auto-regressive, unidirectional LMs with Transformer (Vaswani et al., 2017b). The input sentences are once divided into morphemes by MeCab (Kudo, 2006) with a UniDic dictionary,[8] and then these morphemes are split into subword units by byte-pair-encoding. (Sennrich et al., 2016)[9]. 160M sentences randomly selected from 3B web pages were used to train the LMs. Hyperparameters are shown in Appendix A.1.1. We trained two variants of unidirectional LMs: left-to-right and right-to-left LMs (133M parameters). They achieved a perplexity of 28.51 and 28.25 in validation 10k sentences, respectively.

Given a sentence s, we calculate its generation probability $p(s) = \overrightarrow{p}(s) \cdot \overleftarrow{p}(s)$, where $\overrightarrow{p}(\cdot)$ and $\overleftarrow{p}(\cdot)$ are generation probabilities calculated by a left-to-right LM and a right-to-left LM, respectively.[10] We compare the generation probabilities assigned to s and its variants with different word orders. The order with the highest generation probability is assumed to be favored by LMs.

Results: We compared the word order preference of LMs and that of the workers by using the 2.6K pairs created in Section 3.3.1. The task is to select the word order that is favored by the crowd workers given a pair of word order variations. The accuracy of selecting the

[7] We manually created check questions considering the Japanese speakers' preference in trial experiments in advance.

[8] https://unidic.ninjal.ac.jp/

[9] Implemented in sentencepiece (Kudo and Richardson, 2018) We set character coverage to 0.9995, and vocab size to 100,000.

[10] In this experiment, human subjects were not forced to read from left to right in judging the acceptability, thus we tentatively used bi-directional scores.

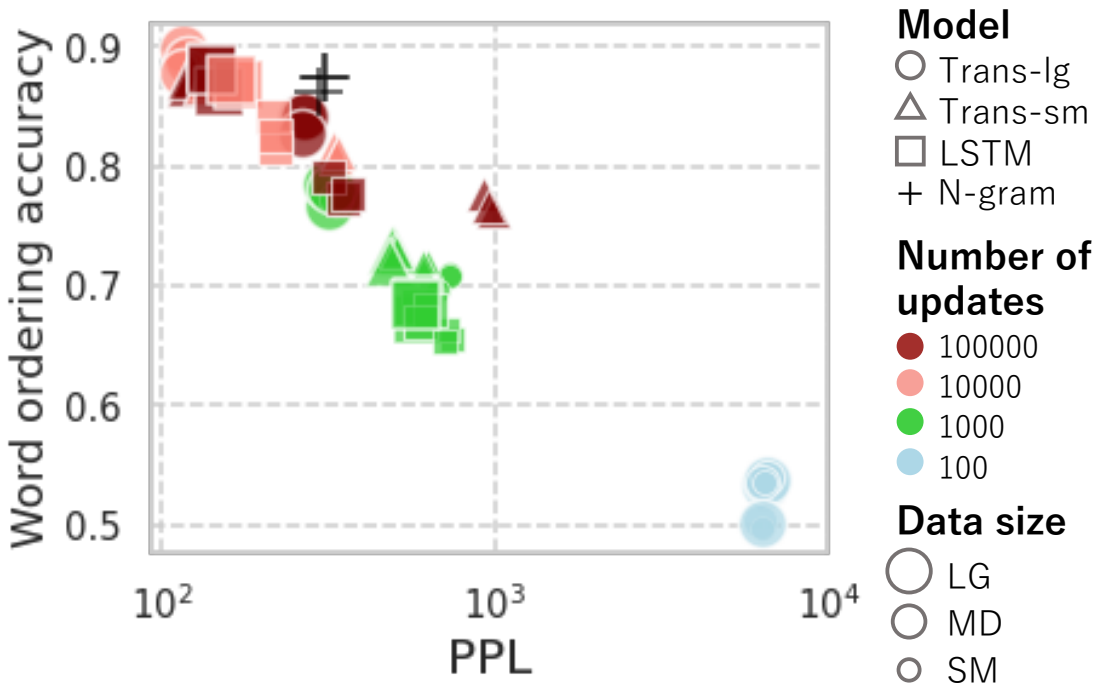


Fig. 3.3 Relationship between the word ordering accuracy and perplexity of LMs.

order preferred by the crowd worker is 95.0 (the chance rate is 50.0). This indicates that LMs exhibit surprisingly well parallels with humans in word order preferences.

3.3.3 Relationship between perplexity and word order preference

We additionally investigated the relationship between the LM quality (i.e., perplexity; PPL) and their word ordering accuracy. Here, higher word ordering accuracy indicates that the model's word order preference is more consistent with humans. This accuracy is evaluated in the same manner as Section 3.3.2. We investigate whether a good PPL, a typical evaluation metric of LMs, can be regarded as an indicator of LM's human-like word order preference.

LMs settings: To examine the word order preference of various LMs differing their perplexity, we trained 111 variants of LMs with different training settings. Namely, the following four variants of LMs were used: Transformer-large (TRANS-LG) (Vaswani et al., 2017a), Transformer-small (TRANS-SM), LSTM (LSTM) (Hochreiter and Schmidhuber, 1997), and N-gram LMs (N-GRAM).[11] The parameters of each neural LM were saved at four different

[11]The neural LMs were trained with the fairseq toolkit (Ott et al., 2019). N-GRAM LMs were trained using KenLM https://github.com/kpu/kenlm.

points during training: 100, 1K, 10K, and 100K parameter updates. As the N-GRAM models, 3-gram, 4-gram, and 5-gram LMs were used. For each neural LM architecture (TRANS-LG, TRANS-SM, and LSTM), three variants were trained using different training data sizes: LG (full training data), MD (1/10 training data), and SM (1/100 training data). Note that the LG data consists of approximately 5M sentences, which is 1/30 smaller than the data used in the previous Section 3.3.2 due to computational resource constraints. The N-gram LMs were trained on the LG datasets. Hyperparameters of neural LMs are shown in Appendix A.1.2.

To summarize, 39 LM training settings were attained (3 architectures $\times$ 3 data size $\times$ 4 parameter updates = 36 neural LMs, plus 3 N-GRAM LMs). In addition, our experiments use three LMs trained using different random seeds for each neural LM training configure; hence, 111 LMs (36 neural LMs $\times$ 3 seeds, plus 3 N-GRAM LMs) were tested for each language. In this section, we evaluated the LMs' word order preferences only in terms of the likelihoods computed by the left-to-right manner $\vec{p}(s)$.

Results: Figure 3.3 shows the relationship between the two metrics; the better PPL LMs have, the more human-like word order preference LMs exhibit. Note that low PPL indicates an accurate prediction of the upcoming word. Based on this observation, in the next section, we further investigate how well the large-scale, accurate LMs (used in Section 3.3.2) exhibit reasonable word order preference.

3.4 Experiment2: consistency with linguistic studies

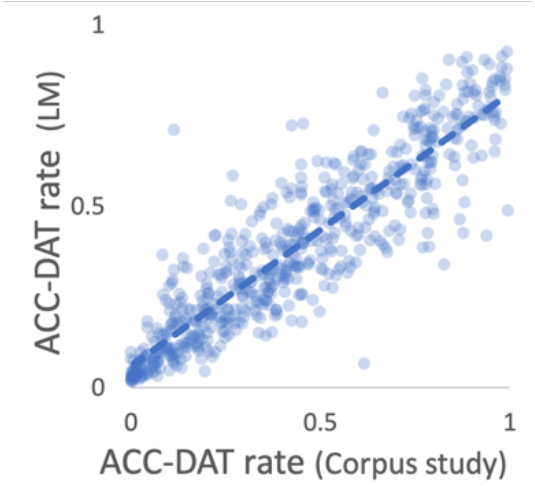
This section examines whether LMs show word order preference that is consistent with previous psycholinguistic and quantitative linguistic studies using controlled materials. The results show that the LM preference is entirely consistent with these studies, which support that LMs have proper word order preference.

3.4.1 Double objects

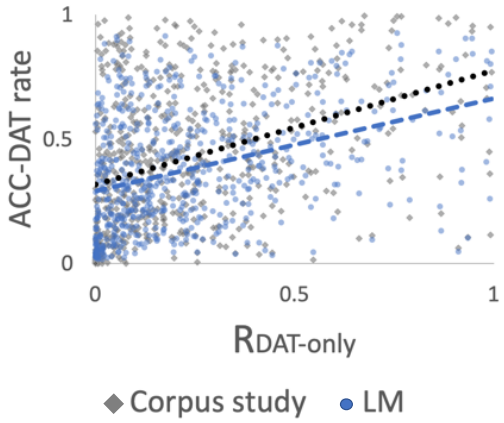
The order of double objects is one of the most controversial topics in Japanese word order studies. Examples of the possible order are as follows:

(5) DAT-ACC: *Seito-ni* *hon-wo* *ageta*
student-DAT book-ACC gave.

ACC-DAT: *Hon-wo* *seito-ni* *ageta*
book-ACC student-DAT gave.



(a) Correlation between each verb's ACC-DAT rate estimated by LMs and corpus counts.



(b) Relationship between each verb's $R^v_{\texttt{DAT-only}}$ and the ACC-DAT rate estimated by LMs and corpus counts.

Fig. 3.4 Overlap of the results of corpus study (Sasano and Okumura, 2016) and that of LMs. In figures (a) and (b), each plot corresponds to each verb.

Henceforth, DAT-ACC (ACC-DAT) denotes the word order in which the DAT (ACC) argument precedes the ACC (DAT) argument. We assess the double object order preference from the viewpoint of the five linguistic aspects.

1. Double-object order for each verb: First, we focus on the double object order trends for each verb. It is known that different verb has different trends; whether LMs imitate such verb-dependent trends is interesting to see.

For a particular verb, the ratio of LMs preferring the DAT-ACC order than the ACC-DAT order is computed. Specifically, using the set of examples S^v for verb v, we: (i) created an instance with the swapped order of ACC and DAT for each example, (ii) compared the generation probabilities of the original and swapped instance, and (iii) summarized the ratio DAT-ACC order is preferred:

$$R^v_{\texttt{ACC-DAT}} = \frac{N^v_{\texttt{ACC-DAT}}}{N^v_{\texttt{ACC-DAT}} + N^v_{\texttt{DAT-ACC}}} \ .$$

Here, the list of verbs v and their examples S^v are collected by Sasano and Okumura (2016). The $R^v_{\texttt{ACC-DAT}}$ for 620 verbs is calculated, and we analyzed the consistency between the obtained preference in LMs and the corpus trends.

Figure 3.4-(a) shows the relationship between $R^v_{\texttt{ACC-DAT}}$ determined by LMs (Y-axis) and one estimated in the existing corpus study (X-axis) (Sasano and Okumura, 2016). Each dot corresponds to the result for each verb. These results strongly correlate with the Pearson correlation coefficient of 0.88. That is, LMs could successfully model the verb-dependent trends in the order of their double objects.

2. Argument omission: It is claimed that a rarely omitted case is placed near a verb (Sasano and Okumura, 2016). Intuitively, for a particular verb v, we computed the two metrics of how frequently ACC argument is (i) **not** omitted for the verb, and (ii) placed near the verb. Then the relationship between the two metrics is investigated.

Specifically, we first quantified how rarely ACC argument is omitted for a verb v as follows:

$$R^v_{\texttt{DAT-only}} = \frac{N^v_{\texttt{DAT-only}}}{N^v_{\texttt{DAT-only}} + N^v_{\texttt{ACC-only}}} \ ,$$

where $N^v_{\texttt{DAT-only}}$ ($N^v_{\texttt{ACC-only}}$) denotes the number of examples in which the DAT (ACC) case appears, and the other case does not in S^v. A large $R^v_{\texttt{DAT-only}}$ score indicates that the DAT argument is less frequently omitted than the ACC argument in S^v.

Then, we compared the $R^v_{\texttt{DAT-only}}$ and how frequently the DAT argument is placed nearer a verb than ACC argument ($R^v_{\texttt{ACC-DAT}}$). Corpus study indicated that there is a positive correlation between the two metrics.

Figure 3.4-(b) shows that LMs reproduced the results reported in corpus-based study (Sasano and Okumura, 2016), where the correlation coefficient was 0.391. The Pearson correlation coefficient between $R^v_{\texttt{DAT-only}}$ and $R^v_{\texttt{ACC-DAT}}$ is 0.374 in the results obtained by LMs.

3. Verb type: There are two types of causative-inchoative alternating verbs in Japanese: show-type verbs and pass-type verbs. Matsuoka (2003) claimed that the double object order depends on their verb types; show-type verb prefers the DAT-ACC order, while the pass-type verb prefers the ACC-DAT order.

The verb types are categorized by the sentence structure with the corresponding inchoative verb. For the show-type verbs, the DAT argument of a causative sentence becomes the subject in its corresponding inchoative sentence (Example (10)). On the other hand, the ACC argument of a causative sentence becomes the subject in its corresponding inchoative sentence for the pass-type verbs (Example (11)):

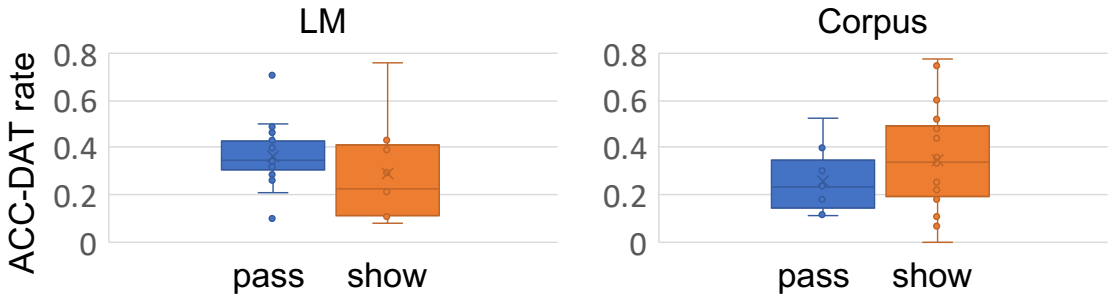


Fig. 3.5 ACC-DAT rate for pass-type and show-type verbs. The left part corresponds to the preferences by LMs, and right part corresponds to those by humans. Both results suggest that there is no significant difference in double object order preference between the two verb types.

(10) **Causative:** *Seito-ni* *hon-wo* *miseta*
student-DAT book-ACC showed.
(ϕ_I showed a student a book.)

Inchoative: *Seito-ga* *mita*
student-NOM saw.
(A student saw $\phi_{\text{something}}$.)

(11) **Causative:** *Seito-ni* *hon-wo* *watashita*
student-DAT book-ACC showed.
(ϕ_I passed a student a book.)

Inchoative: *Hon-ga* *watatta*
book-NOM passed.
(A book passed to $\phi_{\text{something}}$.)

It is hypothesized that the show-type verb prefers the DAT−ACC order, while the pass-type verb prefers the ACC−DAT order.

Figure 3.5 shows the $R^v_{\texttt{ACC-DAT}}$ distributions of the show-type and pass-type verbs. Existing empirical studies reported that there is no significant difference between the two groups (Matsuoka, 2003; Sasano and Okumura, 2016). This is replicated by LMs; it is also observed that the two distributions of $R^v_{\texttt{ACC-DAT}}$ determined by LMs (show-type and pass-type verbs) are not significantly different (the Wilcoxon rank-sum test p-value is 0.12). Notably, LMs showed moderately contrasting trends with the corpus statistics; pass-type verb more prefers ACC−DAT order than show-type. Exploring the source of this discrepancy could be our future work.

4. Animacy of argument: It is claimed that the canonical word order differs depending on the animacy of the arguments (Matsuoka, 2003; Sasano and Okumura, 2016); animate argument tends to be positioned earlier. We used minimal pair of two sentences where only the dative argument is different collected by Sasano and Okumura (2016):

(6) Type-A: *Gakkô-ni* *hon-wo* *kaeshita*
school-DAT book-ACC returned.

Type-B: *Sensei-ni* *hon-wo* *kaeshita*
teacher-DAT book-ACC returned.

Here, Type-A has an inanimate goal (*school*) as the DAT argument, while Type-B has an animate processor (*teacher*). The ACC-DAT order should be more favored by Type-A construction than Type-B because the inanimate dative argument in Type-A should tend to be introduced later. Following Sasano and Okumura (2016), we analyzed 113 verbs.[12]

For each verb, we calculated the two scores: (i) ratio that the ACC-DAT order is preferred by LMs in its Type-A examples, and (ii) the ratio in its Type-B examples. Then we categorized the verb into two groups: the ACC-DAT order is more frequently preferred in its Type-A than its Type-B (group 1), and those with opposite trends (group 2). The number of verbs in group-1 is significantly larger (a two-sided sign test $p < 0.05$). That is, LMs could reproduce the trend that animate argument is likely to be positioned earlier (i.e., the Type A sentence, where an DAT argument is inanimate, tends to be ACC-DAT order, where inanimate DAT argument is positioned later).

5. Co-occurrence of verb and arguments: It is claimed that an argument that frequently co-occurs with the verb tends to be placed near the verb (Sasano and Okumura, 2016). To quantify this trend, we compared the two metrics: (i) the ACC-DAT ratio, and (ii) how the DAT noun more strongly co-occurs with the verb than the ACC noun. Given a triple of (v, DAT noun, ACC noun), the latter co-occurrence score is calculated as follows:

$$\Delta\text{NPMI} = \text{NPMI}(n_{\texttt{DAT}}, v) - \text{NPMI}(n_{\texttt{ACC}}, v) \ ,$$
$$\text{where NPMI}(n_c, v) = \frac{\text{PMI}(n_c, v)}{-log(p(n_c, v))} \ ,$$
$$\text{PMI}(n_c, v) = log\frac{p(n_c, v)}{p(n_c)p(v)} \ ,$$

[12] Among the 126 verbs used in Sasano and Okumura (2016), 113 verbs with data that do not overlap with the LM training data were selected.

where, v is a verb and n_c ($c \in \{\texttt{DAT}, \texttt{ACC}\}$) is its argument. A higher ΔNPMI means that the DAT noun more strongly co-occurs with the verb than the ACC noun. It is suggested that high ΔNPMI (DAT noun highly co-occurs with verb) leads to a high ACC−DAT ratio (DAT noun is positioned near the verb).

The correlation between ΔNPMI and the ACC−DAT ratio among the examples was 0.521 with Pearson's correlation coefficient. This replicates the corpus trend (Sasano and Okumura, 2016); the correlation was 0.577.

3.4.2 Adverb position

Next, the preference of the adverb position in LMs is investigated. Note that the adverb position has no strict restriction except that it must be before the verb. Koizumi and Tamaoka (2006) suggested that the canonical position of an adverb depends on its type. They investigated the position of the following four types of adverbs: MODAL, TIME, MANNER, and RESULTIVE.

We collected the same examples as Koizumi and Tamaoka (2006), and for each example, we identified the most plausible adverb position estimated by LMs. Specifically, for each example, we identified the order with the highest probability in the following three candidates:

(10) **ASOV:** *Ranbôni tomodachi-ga dôgu-wo atsukatta*
roughly friend-NOM tools-ACC handled.

SAOV: *Tomodachi-ga ranbôni dôgu-wo atsukatta*
friend-NOM roughly tools-ACC handled.

SOAV: *Tomodachi-ga dôgu-wo ranbôni astukatta*
friend-NOM tools-ACC roughly handled.

Then, we aggregated the most frequently preferred order in each adverb group. Table 3.2 shows the preferred adverb position by LMs. In the MANNER adverb, the SAOV and SOAV order is selected the same number of examples. There is a remarkable parallel between the position LMs and humans assumed to be natural. The human results are from Koizumi and Tamaoka (2006).

3.4.3 Order of constituents representing time, location, and subject

It is well-known that the constituent representing time information (TIM), location information (LOC), and the subject (NOM) is ordered in the TIM-LOC-NOM order (Saeki, 1960, 1998).

Table 3.2 Overlap of the preference of LMs and that of humans (Koizumi and Tamaoka, 2006) in the adverb position.

Model	MODAL	TIME	MANNER	RESULTIVE
LM	ASOV	SAOV	SAOV, SOAV	SOAV
Koizumi(2016)	ASOV	ASOV, SAOV	SAOV, SOAV	SAOV, SOAV

These categories (e.g., location) do not exactly correspond to the particle types; for example, some constituents with particle "de" indicates the place information, while others with "de" do not. We examine whether LMs could capture the categories of constituent content, specifically through the lens of position and time order.

Data First, we randomly collected 50M sentences from 3B web pages. Note that there is no overlap between the collected sentences and the training data of LMs. Next, we obtained the sentences that satisfy the following criteria: (i) there is a verb (placed at the end of the sentence) with more than two directly depending arguments, and (ii) each argument (with its modifiers) has fewer than 11 morphemes. Each example sentence is created by ordering the arguments with their descendants and verb, preserving their original order.

We then regard the constituent (argument and its descendants) satisfying the following condition as the `TIM` constituent: (i) accompanying the postpositional particle "ni," and (ii) containing time category morphemes (identified by JUMAN). We regard the constituent (argument and its descendants) satisfying the following condition as the `LOC` constituent: (i) accompanying the postpositional particle "de," and (ii) containing location category morphemes (identified by JUMAN). 81k examples with `TIM` or `LOC` constituents were created. The average number of characters in a sentence was 45.1 characters.

Word order analysis For each example s, we created all possible word orders (scrambling the arguments) and obtained the word order with the highest generation probability ($\hat{s}$). Given a set of LMs-preferred examples $\hat{S}$, the ratio $o(a \prec b)$ that the type a constituent precedes the type b is calculated:

$$o(a \prec b) = \frac{N_{a \prec b}}{N_{a \prec b} + N_{b \prec a}} \ ,$$

where $N_{a \prec b}$ is the number of examples where the type a constituent precedes the type b in $\hat{S}$. The $o(a \prec b)$ score more than 0.5 indicates that the type a is more likely to be placed before the type b. If the `TIM-LOC-NOM` order is preferred in LMs, $o($`TIM` $\prec$ `LOC`$)$, $o($`TIM` $\prec$ `NOM`$)$,

Table 3.3 The columns $a \prec b$ show the score $o(a \prec b)$, which indicates the rate of constituent type a being more likely to be placed before b. The row "Corpus" shows the trends in original corpus.

	TIM≺LOC	TIM≺NOM	LOC≺NOM
LM	.708	.632	.615
Corpus	.686	.666	.681

and $o(\texttt{LOC} \prec \texttt{NOM})$ should be larger than 0.5. The results are shown in Table 3.3. The corpus statistics agrees with the`TIM-LOC-NOM` order, and the LM also replicates this preference. This indicates that LMs have word order preference that is consistent with corpus statistics and known linguistic description.

3.5 Conclusion

From the two experiments, we tentatively concluded that neural LMs exhibit human-like word order preference even in a word order flexible language. Their preferences were successfully aligned with human acceptability judgment and several reports in linguistic studies. In addition, we observed that LMs with better next-word prediction performance show more human-like word order preference. From a scientific viewpoint, these results also imply that the human word order preference could be acquired only through exposure to language data, and *why*/*how* these particular trends emerged could be an interesting next question.

Chapter 4

Cross-Linguistic Analysis of Incremental Sentence Processing by Neural Language Models

4.1 Introduction

Whereas the previous chapter analyzed the sentence-level acceptability judgment by humans and language models (LMs), this chapter analyzes the more in-depth, token-by-token incremental processing difficulties exhibited by humans and LMs. It is well known that the probability of a word in context (i.e., surprisal) impacts its processing difficulty in incremental human language comprehension (Demberg and Keller, 2008; Hale, 2001; Levy, 2008; Smith and Levy, 2013). Building on this basis, researchers have compared a variety of language models (LMs) in terms of how well their surprisal correlates with human reading behavior (Aurnhammer and Frank, 2019; Fossum and Levy, 2012; Frank and Bod, 2011; Goodkind and Bicknell, 2018; Hale et al., 2018; Merkx and Frank, 2021; Roark et al., 2009; Wilcox et al., 2020). For example, recent studies reported that LMs with better performance for next-word prediction could also better predict the human reading behavior (i.e. more human-like) (Fossum and Levy, 2012; Goodkind and Bicknell, 2018; Wilcox et al., 2020).

In this chapter, we re-examine whether the recent findings on human-like computational models can be generalized across languages. Despite the community's ongoing search for a language-independent model (Bender, 2011), existing studies have focused almost exclusively on the English language. Having said that, broad-coverage cross-linguistic evaluation of the existing reports is prohibitively difficult. In fact, data on human reading behavior

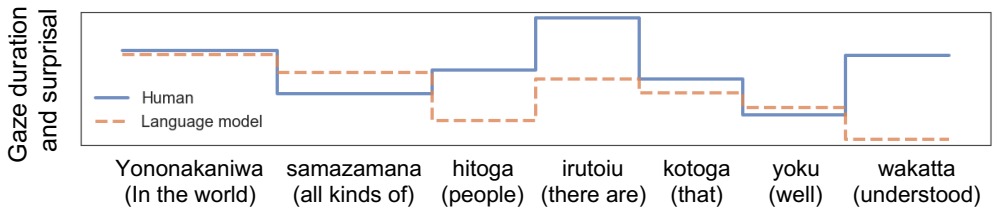


Fig. 4.1 Gaze duration from human subjects and surprisal from language models for the Japanese sentence "Yononakaniwa samazamana hitoga irutoiu kotoga yoku wakatta." (*I understood well that there are all kinds of people in the world.*)

(e.g., eye movement) is available only in limited languages. As an initial foray, this study focuses on the Japanese language as a representative of languages that have typologically different characteristics from the English language. If the observation is different between English and Japanese, the current findings on English data might lack a universality across languages.

We specifically revisit the recent report—*the lower perplexity a LM has, the more human-like the LM is*—in the English and Japanese languages (Fossum and Levy, 2012; Goodkind and Bicknell, 2018; Wilcox et al., 2020). In addition to the importance of cross-linguistic evaluation, the report itself is worth investigating. Recent studies in the machine learning field have reported that more parameters, training data, and computation cost can result in better PPL (Brown et al., 2020; Kaplan et al., 2020). Our investigation has implications for whether a human-like model might exist beyond such improvements.

More concretely, over three dozen of LMs were trained for each language, with variants in their architecture, training data size, and the number of parameter updates. Then, the surprisals computed by each LM were compared to human eye movement data (Figure 4.1). The analysis of the relationship between PPL and the psychometric predictive power revealed substantively different trends between the Japanese and English LMs. In Japanese, a lower PPL of a LM does not indicate better performance for modeling reading behavior. By contrast, in English, there was a clear relationship between the two metrics as reported in the prior studies.

This opens a remaining and important question: why are English and Japanese different in this aspect? We discuss the differing results between English and Japanese from the perspective of the uniform information density hypothesis (Genzel and Charniak, 2002; Jaeger and Levy, 2007; Levy, 2005). We find that the processing difficulty (i.e., gaze duration) of words is less uniformly distributed within a Japanese sentence. Given this, the discrepancy of the results between English and Japanese might stem from a mismatch between the information uniformity of the target language and the LM's training objective. We demonstrate

that tuning Japanese LMs to this training objective collapses the human-like nonuniformity of the processing difficulty observed in Japanese subjects.

4.2 Related work

4.2.1 Human sentence processing and LMs

What factor determines the incremental difficulty of human language processing? At present, surprisal theory (Hale, 2001; Levy, 2008) has been widely adopted in the field of computational psycholinguistics. This theory suggests that the processing difficulty of a word is determined by how predictable the word is in its preceding context ($-\log p(\mathrm{word}|\mathrm{context})$).

Existing studies have compared various computational models by checking the effectiveness of their surprisals in modeling human reading behavior (Fossum and Levy, 2012; Frank and Bod, 2011; Goodkind and Bicknell, 2018; Hale, 2001; Hale et al., 2018; Merkx and Frank, 2021; Roark et al., 2009; Wilcox et al., 2020). Data such as eye movement (Kennedy et al., 2003) and brain activity (Brennan et al., 2016; Frank et al., 2015) are used as measures of human reading behavior. For example, using eye movement data, Frank and Bod (2011) compared the surprisals from phrase-structure grammars (PSGs) with those from a non-hierarchical, sequential model, tentatively concluding that human sentence processing was insensitive to hierarchical structures since non-hierarchical models displayed better psychological predictive power than PSGs. Recently, researchers reported that surprisals from LMs with low PPL correlate well with human reading behaviors (Aurnhammer and Frank, 2019; Fossum and Levy, 2012; Goodkind and Bicknell, 2018; Wilcox et al., 2020).

The work most closely related to this study is Wilcox et al. (2020). They examined the relationship between PPL, psychometric predictive power, and syntactic knowledge in LMs using a variety of models, including modern neural LMs (Radrof et al., 2018). They found a tight relationship between PPL and psychometric predictive power in the English corpora. This study investigates whether this relationship can be generalized across languages.

4.2.2 Reading behavior in Japanese

In comparison to English speakers, Japanese speakers display different patterns in sentence processing. For example, an anti-locality effect (the more modifiers a word has in its preceding context, the easier the word is to process) has typically been observed in head-final languages, including Japanese (Konieczny, 2000). Such differences between the languages are assumed to be more or less due to their different sentence structures. Recently, eye

movement data for naturally occurring Japanese texts have recently become available (Asahara et al., 2016) and was extensively annotated with various linguistic properties (Asahara, 2017, 2018; Asahara and Kato, 2017).

4.3 Methods

This section describes the settings of LMs, eye movement data, and evaluation metrics.

4.3.1 Language models

A variety of sentence-level, left-to-right sequential LMs was used. Note that this setting is the same as that in Section 3.3.3.

Training data of English LMs: We used the WikiText-103 dataset to train the English LMs. Based on the reports that subword-level English LMs exhibits superior psychometric predictive power (Wilcox et al., 2020), input texts were divided into subwords by a byte-pair encoding (BPE) (Sennrich et al., 2016).[1] The training data consist of approximately 4M sentences (114M subwords units).

Training data of Japanese LMs: We used news articles and the Japanese part of Wikipedia to train the Japanese LMs. Input texts were first segmented into morphemes by MeCab (Kudo, 2006) with unidic dictionary, and then further divided into subwords by BPE.[2] The training data consist of approximately 5M sentences (146M subwords units).

Architectures: The following four variants of LMs were used: Transformer-large (TRANS-LG) (Vaswani et al., 2017a), Transformer-small (TRANS-SM), LSTM (LSTM) (Hochreiter and Schmidhuber, 1997), and N-gram LMs (N-GRAM).[2] The parameter size was almost the same for TRANS-SM and LSTM. With respect to the N-GRAM models, 3-gram, 4-gram, and 5-gram LMs were used.

Training data size: For each neural LM architecture (TRANS-LG, TRANS-SM, and LSTM), three variants were trained using different training data sizes: LG (full training data), MD

[1] Implemented in SentencePiece (Kudo and Richardson, 2018). We set character coverage to 0.9995, and vocabulary size to 32,000 in English. In Japanese, the vocabulary size is 100,000, reflecting its rich morphemes.

[2] The neural LMs were trained with the fairseq toolkit (Ott et al., 2019). N-GRAM LMs were trained using KenLM https://github.com/kpu/kenlm.

(1/10 training data), and SM (1/100 training data). The N-gram LMs were trained on LG datasets.

Number of updates: The parameters of each neural LM were saved at four different points during training: 100, 1K, 10K, and 100K parameter updates.

To summarize, 39 LM training settings were attained for each language (3 architectures $\times$ 3 data size $\times$ 4 parameter updates = 36 neural LMs, plus 3 N-GRAM LMs). In addition, our experiments use three LMs trained using different random seeds for each neural LM training configure; hence, 111 LMs (36 neural LMs $\times$ 3 seeds, plus 3 N-GRAM LMs) were tested for each language. Appendix A.1.2 shows the detailed hyperparameters of TRANS-SM, TRANS-LG, and LSTM.

4.3.2 Eye movement data

English: The Dundee Corpus (Kennedy et al., 2003), which contains gaze duration annotation for each word, was used. Following Smith and Levy (2013), the first-pass gaze duration was analyzed. Then, following Goodkind and Bicknell (2018), the data points that met any of the following criteria were excluded:

- data points with zero gaze duration or that beyond three standard deviations
- words with punctuation or numeric characters
- words whose next word has punctuation or numeric characters
- first or last word in a line

In total, the analysis included 214,955 data points in the corpus.

Japanese: The BCCWJ-EyeTrack (Asahara et al., 2016), which contains gaze duration annotation for each phrasal unit, was used. Note that the phrasal unit (i.e., bunsetsu) consists of at least one content morpheme and its postpositional function morphemes. Henceforth, an English word and a Japanese phrasal unit are referred to as a "word." The same exclusion criteria as the Dundee Corpus was applied to the BCCWJ-EyeTrack data.[3] In total, the analysis included 6,009 data points in the corpus. Note that the BCCWJ-EyeTrack data was deliberately designed to address language-specific issues in Japanese such as the lack of segmentation spaces in Japanese texts (Asahara et al., 2016).

[3]Strictly speaking, the exclusion criteria was slightly different between Japanese and English data. In the Japanese data, we included the words whose next word had punctuation or a numeric character, as there is no spillover effect in Japanese (see Section 4.3.3)

Table 4.1 Statistics of the corpora used for evaluating the psychometric predictive power of LMs. "#articles" and "#sents." are the number of articles and sentences in each corpus. "#words" denotes the number of words annotated with human reading time in each corpus. "#data points" is the number of reading time annotations used in our experiments. Each word has the reading time annotations from multiple subjects (#subjects per article). "Avg. GD per word" is the averaged gaze duration per word. "Avg. #subwords per word" denotes the averaged number of subwords consisting of each word.

Corpus	#articles	#sents.	#words	#data points (used)	#subjects per article	Avg. GD per word	Avg. #subwords per word
Dundee Corpus	20	2,478	51,501	214,955	10	227.1	1.3
BCCWJ-EyeTrack	20	218	1,643	6,009	12	361.6	3.4

Statistics: Table 4.1 shows the statistics of the Dundee Corpus and BCCWJ-EyeTrack data. The BCCWJ-EyeTrack has more than 10 times a smaller number of data points than the Dundee Corpus. Notably, the word annotated with eye movement information differs between English and Japanese. On average, a Japanese word consists of 3.4 subwords, while an English word consists of 1.3 subwords.

4.3.3 Evaluation metrics

Perplexity (PPL): Again, PPL of N symbols $(w_1, w_2, \cdots, w_N)$, a typical evaluation metric for unidirectional LMs, is defined as follows (Eq. 4.1):

$$\mathrm{PPL} = \prod_{i=0}^{N} p(w_i|w_{<i})^{-\frac{1}{N}} \quad . \tag{4.1}$$

Low PPL indicates that the model can accurately predict the upcoming signal based on its preceding context. The training objective of LMs works to minimize the PPL computed by the model. In the experiments, the PPL of a LM is evaluated with the texts in the eye movement data, which do not overlap with the training data. A model with low PPL is also called a *linguistically accurate* model (Frank and Bod, 2011).

Psychometric predictive power: The surprisal measure, a negative logarithmic probability of a word in context ($-\log p(\mathrm{word}|\mathrm{context})$), is a widely used information-theoretic complexity metric. Intuitively, a model is considered to have high psychometric predictive power (i.e., *psychological accuracy*) if the surprisals of words computed by the model have trends similar to the human subject's cognitive load (e.g., measured by gaze duration). Following the existing studies (Goodkind and Bicknell, 2018; Merkx and Frank, 2021; Wilcox et al.,

2020), the psychometric predictive power of a model was measured by comparing surprisal from the model and gaze duration from human subjects.

While LMs process a text subword-by-subword, gaze duration is annotated in a larger unit. Following the study using subwords (Wilcox et al., 2020), the surprisal of each word was calculated using the joint probability of its constituent subwords. Formally, given a text consisting of N subwords $w_{1:N} = (w_1, w_2, \cdots, w_N)$, surprisal $I(\cdot)$ of a word $s_k = (w_l, w_{l+1}, \cdots, w_m)$, where $1 \leq l \leq m \leq N$, was calculated as follows:

$$\begin{aligned} I(s_k) &= -\log p(w_l, \cdots, w_m | w_{<l}) \\ &= -\sum_{k=l}^{m} \log p(w_k | w_1, \cdots, w_{k-1}) \ . \end{aligned} \tag{4.2}$$

The effect of surprisals for modeling human reading behavior was calculated using a linear mixed-effects regression (Bates et al., 2015). Specifically, the gaze duration (`GD`) was modeled using the following formula:

$$\begin{aligned} \texttt{GD} \sim\ & \texttt{surprisal} + \texttt{surprisal\_prev\_1} \\ & + \texttt{surprisal\_prev\_2} + \texttt{freq} * \texttt{length} \\ & + \texttt{freq\_prev\_1} * \texttt{length\_prev\_1} \\ & + \texttt{screenN} + \texttt{lineN} + \texttt{wordN} \\ & + \texttt{(1|article)} + \texttt{(1|subj)} \ . \end{aligned} \tag{4.3}$$

The regression model includes baseline factors (e.g., frequency of a word) that are of no interest in the comparison of LMs. A collection of factors used in the existing studies (Asahara et al., 2016; Wilcox et al., 2020) were initially examined and the factors that were not significant ($p > 0.05$) for gaze duration modeling both in the Dundee Corpus and BCCWJ-EyeTrack were excluded. The frequency of a word (`freq`) was calculated using the entire training data for LMs. Table 4.2 shows the details of each factor. The frequency of a word (`freq`) was estimated using the full training data for the LMs.

In English experiments, surprisals of preceding words (`surprisal_prev_1` and `surprisal_prev_2`) were included in order to handle the spillover effect (the processing cost of a certain word is affected by its preceding words) (Rayner and Well, 1996; Smith and Levy, 2013). In Japanese experiments, the surprisals of preceding words were not included because our preliminary experiment showed that these factors were not significantly effective

Table 4.2 Factor names and their descriptions.

Factor name	Type	Description
`surprisal`	num	surprisal caluzulted by LMs
`GD`	num	reading time (first pass time)
`article`	factor	article ID
`screenN`	int	screen display order
`lineN`	int	the serial number of line the segment is displayed
`wordN`	int	the serial number of segment in a screen
`sentN`	int	the serial number of sentence the segment belongs to
`tokenN`	int	the position of segment in sentence
`length`	int	number of characters
`freq`	num	geometric mean of the frequencies of subword constituents in a segment
`subj`	factor	participant ID
`syn_category`	factor	syntactic category the segment falls into (`nominal`, `verbal`, `modifier`, or `other`)
`sem_category`	factor	semantic category the segment falls into (`relation`, `subject`, `action`, `product`, or `nature`)
`n_dependents`	int	number of dependents before the segment

for modeling gaze duration in the BCCWJ-EyeTrack.[4] All the regression models used in our experiments were converged.

To isolate the effect of surprisal for gaze duration modeling, a baseline regression model was trained without surprisal information (excluding the `surprisal`, `surprisal_prev_1`, and `surprisal_prev_2` terms from Eq. 4.3). Following Wilcox et al. (2020), the mean by-word difference of log-likelihood between the model using surprisal values (Eq. 4.3) and the baseline model was calculated. Henceforth, this metric is called ΔLogLik. When surprisal from a LM is not effective for gaze duration modeling, the ΔLogLik score becomes zero. A high ΔLogLik means that the surprisal values obtained by the LM are effective for modeling gaze duration (i.e., the LM has a high psychometric predictive power).

4.4 Experiments

The relationship between PPL and psychometric predictive power is investigated. Furthermore, the relationship is analyzed with respect to the training configures of LMs (e.g., the

[4]The reason is probably that a Japanese phrasal unit (i.e., bunsetsu) could be a larger unit than an English word.

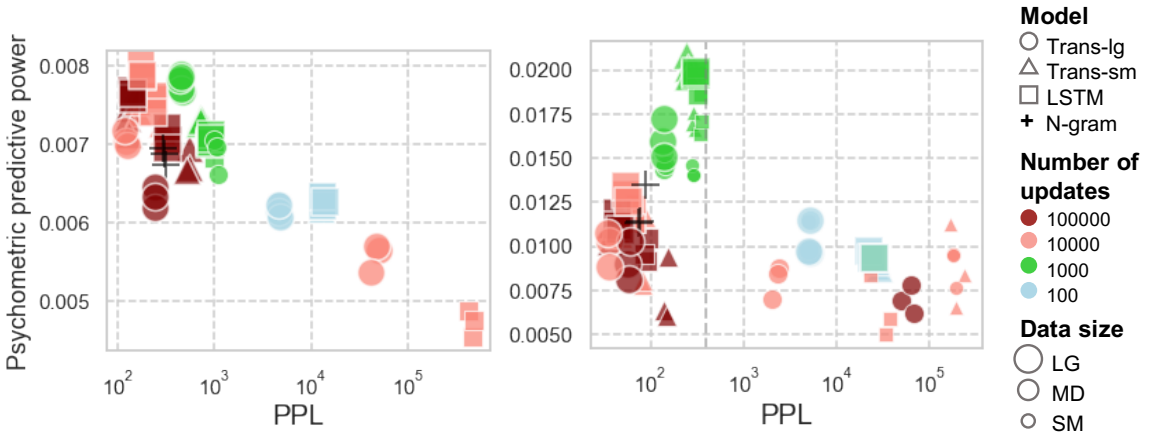


Fig. 4.2 Relationship between PPL (X-axis) and psychometric predictive power, i.e., ΔLogLik (Y-axis) in the English and Japanese languages. Each point corresponds to each LM. A low score on the X-axis indicates the high linguistic accuracy of the model. The PPL was calculated on the eye movement data, and the LMs with PPL more than 10^6 was excluded from the figure. A high score on the Y-axis indicates that the model has a high psychometric predictive power. Note that the X-axis is on a log scale.

number of parameter updates). Then, we discuss the results from the perspective of the uniformity of information density.

4.4.1 Psychometric predictive power and PPL

Figure 4.2 shows the relationship between PPL and psychometric predictive power (i.e., ΔLogLik) of LMs in each of the languages. Each point corresponds to each LM, and a score on the Y-axis indicates the psychometric predictive power of a LM (higher is better). The X-axis is PPL on a log scale (lower is better).

Dundee Corpus: First, the results of the data from the Dundee Corpus show a clear relationship between PPL and psychometric predictive power; namely, lower PPL corresponds to more psychometric predictive power, as reported by prior studies (Goodkind and Bicknell, 2018; Wilcox et al., 2020). Spearman's rank correlation coefficient between the two metrics was –0.87.

BCCWJ-EyeTrack: By contrast, in BCCWJ-EyeTrack, there was no clear, consistent trend between the PPL and psychometric predictive power. While LMs with PPL over 400 show the correlation between PPL and psychometric predictive power (–0.68 with Spearman's ρ), there is a positive correlation (0.53 with Spearman's ρ) for LMs with PPL below

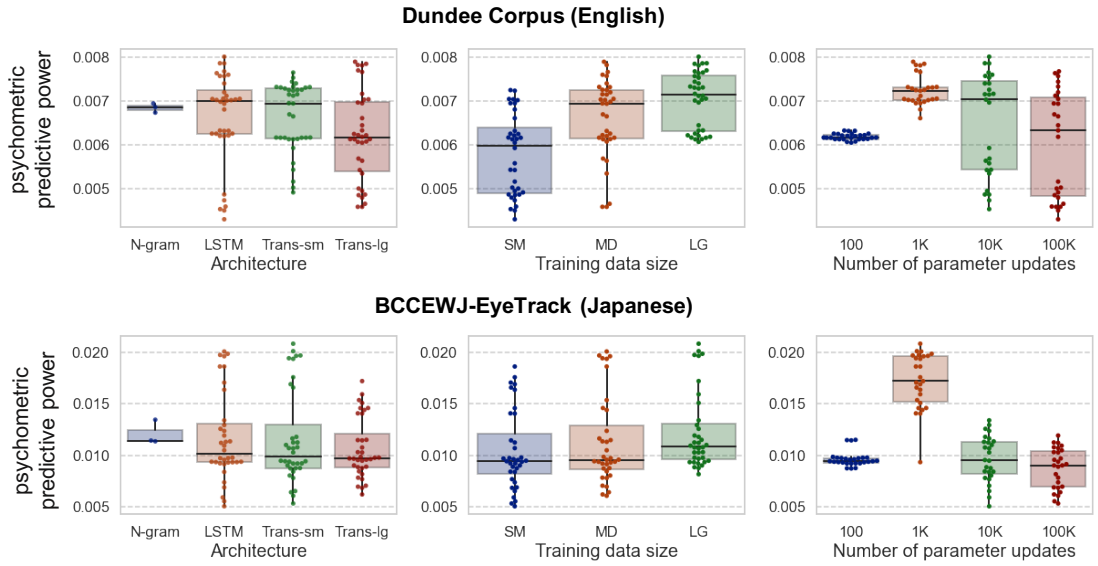


Fig. 4.3 Separate effect of model architecture, training data size, and the number of parameter updates for LMs' psychometric predictive power in each language. Each point corresponds to each LM. The box shows the quartiles of the data. The whiskers show 1.5 times interquartile range.

400. The positive correlation means that the more accurately the LMs can predict the upcoming word, the *worse* the psychometric predictive power of the LMs is. These results demonstrate the non-universality of the recent report across languages; *lower perplexity is not always human-like*. The LSTM LM trained using the MD dataset with 1K updates achieved the best psychometric predictive power. Notably, surprisal was effective for gaze duration modeling in all the Japanese LMs. ΔlogLik scores were significantly higher than zero with the chi-square test (p <0.05).

4.4.2 Model architectures, data sizes, number of parameter updates

Which factor (e.g., model architecture, training data size, and the number of parameter updates) characterizes the psychometric predictive power of LMs? Is the collection of effective factors consistent between the two languages? This study takes a more in-depth look at the separate effects of (i) model architecture, (ii) training data size, and (iii) the number of parameter updates for the psychometric predictive power.

Figure 4.3 summarizes the effect of each factor, where the Y-axis denotes the psychometric predictive power. The most noticeable trend is that Japanese LMs with a relatively fewer number of parameter updates (1K) have better psychometric predictive power than the other Japanese LMs (bottom right part of Figure 4.3), while this trend does not exist in the En-

glish LMs (top right part). This implies that the training objective of the LMs, maximizing $\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})$, had a negative impact on the psychometric predictive power of LMs, at least in Japanese. We discuss this point in Section 4.4.3.

To quantitatively test the differences in Figure 4.3, a linear regression model was trained to estimate psychometric predictive power with the factors of the model architecture, the training data size, and the parameter update number in each language. The training data size and the parameter update number are represented as logarithmically transformed numerical factors. The following trends were found: (i) ; (ii) the training data size positively affects the performance in English alone; and (iii) the number of parameter updates positively affects the performance only in English. There was no factor that boosted the psychometric predictive power of LMs in both English and Japanese languages.

4.4.3 Discussion: uniform information density

The key question is: why do Japanese and English show different trends between PPL and psychometric predictive power? One possible interpretation connecting our results to the uniform information density is discussed in this section.

In computational psycholinguistics, it is commonly assumed that language is designed to enable efficient communication. This principle has been typically investigated under the uniform information density (UID) hypothesis (Genzel and Charniak, 2002; Jaeger and Levy, 2007; Levy, 2005). This hypothesis suggests that speakers seek to keep the amount of information constant across the signals (e.g., words).

Assuming this hypothesis holds for all languages, the reasonable expectation would be for human subjects to show a near-uniform gaze duration across words regardless of their native language. However, this study found that the coefficient of variation[5] in gaze duration over the whole corpus was around 1.7 times higher in Japanese compared to English (0.75 vs. 0.44). Specifically, in Japanese, the gaze duration tended to speed up towards the end of sentences, whereas the duration was near-uniform in English (Figure 4.4).[6] These observations imply that the Japanese language might have a less uniform information density than English. This phenomenon was also investigated through the lens of word order, where SOV languages such as Japanese are reported to show less uniformity of information density (Maurits et al., 2010).

[5]Coefficient of variation is $\frac{\sigma}{\mu}$, where σ and μ are the standard deviation and the mean of the first-pass gaze durations in the eye movement data.

[6]At least in our experimental setup, token position within the sentence was not significantly effective for gaze duration modeling in English sentences, whereas it was significant in Japanese sentences. We checked the coefficient of the factor of position in sentence `wordN` using the linear regression model of `GD ~ sengmentN`.

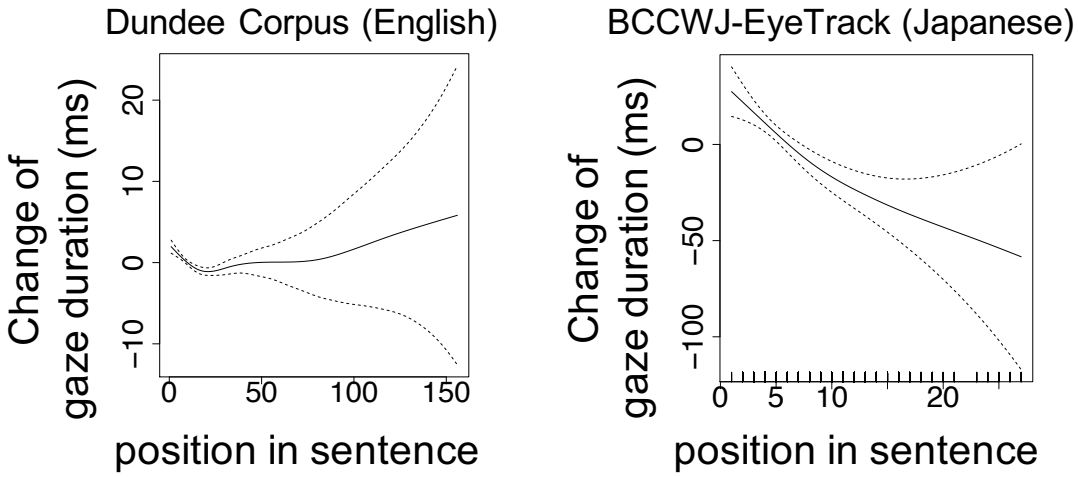


Fig. 4.4 Uniformity of gaze duration with respect to wordposition in a sentence. This plot is computed by the generalized additive model of `GD` $\sim$ `wordN`. Here, `wordN` is denoted as the position of a word in a sentence.

Based on this observation, the discrepancy between English and Japanese low-PPL LMs' psycholinguistic predictive power could stem from a mismatch between the LM's training objective and the information uniformity of the target language. The objective function, $\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})$, defines that the "ideal" is to maximize all next word probabilities to 1.0 (a *uniform* goal).[7] That is, LMs are, *in theory*, trained to approach a model satisfying the UID assumption (Bloem, 2016), where all surprisals from the LM are equally, sufficiently small across the words. Therefore, the objective function might lead to a worse approximation of human-like surprisal in languages that are further from the UID assumption, such as Japanese, while it might be more compatible with English, which has a more uniform processing difficulty across words. This explanation would be consistent with the observation that more tuning to the LM training objective (i.e., a lower PPL) had a negative impact on the psycholinguistic performance of the Japanese LMs (Section 4.4.2). Note the tendency of LMs to assign unreasonably high probabilities to words has also attracted attention from the viewpoint of memorization capability of LMs (Carlini et al., 2020). In addition, the connection of the UID hypothesis to the modern NLP techniques has been recently explored (Meister et al., 2020; Wei et al., 2021b). We further investigate our hypothesis in Section 4.5.

[7] PPL, $\prod_{i=1}^{N} P(w_i|w_{<i})^{-\frac{1}{N}}$, is minimized when the LM objective are maximized.

4.5 Probing nonuniform information density of Japanese LMs

This study hypothesized that tuning to the LM objective (i.e., uniform goal) obscures the nonuniform trend observed in the reading behavior of Japanese subjects. We investigated whether the nonuniformity of the processing difficulty observed in human reading time is mirrored by LM surprisals.

Settings: In a preliminary experiment, we observed that the *syntactic category* (similar to part-of-speech) was the most dominant linguistic factor for explaining the difference in human gaze duration in Japanese sentences (see Appendix 4.7). Based on this observation, we analyze the nonuniformity of surprisals in Japanese LMs with respect to the syntactic categories.

The words in BCCWJ-EyeTrack were classified into one of the following syntactic categories: (a) `nominal` (nouns), (b) `verbal` (verbs), (c) `modifier` (adjectives and adverbs), and (d) `other` entries, as follows:

Kanojo-ga	*akai*	*kaban-o*	*kat-ta*
She-NOM	red	bag-ACC	buy-PAST
nominal	modifier	nominal	verbal

As Asahara and Kato (2017) reported, `verbal` and `modifier` words have a shorter gaze duration than the other words in Japanese sentences. An analysis was conducted on how strongly the Japanese LM's surprisals on words are influenced by their syntactic category. This influence can be evaluated by examining how effectively syntactic category factors can model LM surprisals.

In this experiment, surprisal was regarded as "simulated gaze duration" from an "LM subject," and the importance of syntactic category information for modeling the simulated gaze duration (`simulated_GD`) was evaluated. To inspect the effect of the syntactic category information for modeling the simulated gaze duration, the following regression model[8] was used, including a factor defining which syntactic category the word falls into (`syn_category`):

$$\texttt{simulated\_GD} \sim \texttt{syn\_category} + \texttt{sentN} + \texttt{tokenN} + \texttt{freq} * \texttt{length} \ . \tag{4.4}$$

[8] `sentN` and `tokenN` denote the sentence position and the word position in a sentence (see Appendix 4.7). Note that the `tokenN` and syntactic category exhibit low correlation (0.02 with Pearson's r).

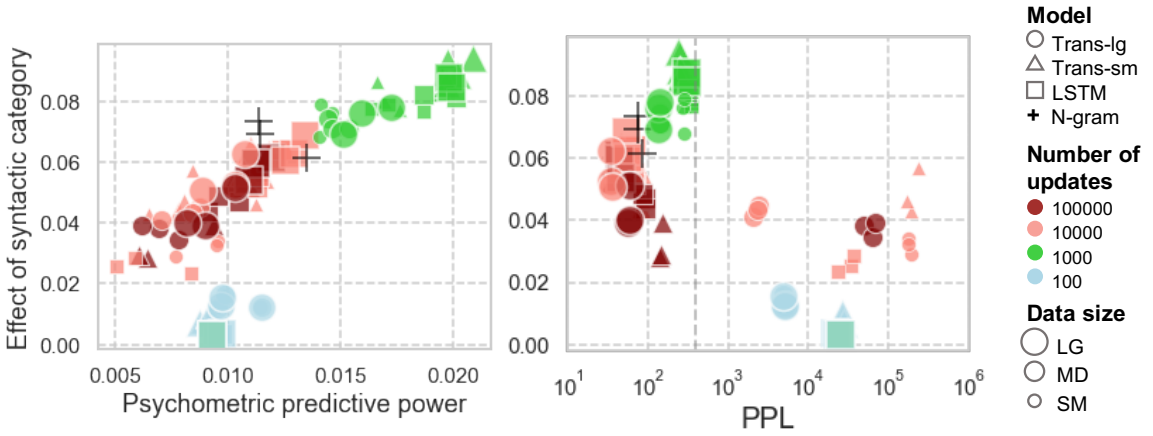


Fig. 4.5 Relationship between the LM's psychometric predictive power and the effect of the syntactic category on the surprisal computed by each LM (left part), and that between PPL and the effect of the syntactic category (right part). Each point corresponds to each LM. The PPL was calculated on the eye movement data, and the LMs with PPL of more than 10^6 was excluded from the right part of the figure.

From this regression model, a log-likelihood score for the simulated gaze duration was obtained. To evaluate the separate effect of `syn_category`, ΔLogLik between Eq. 4.4 and a baseline model was calculated. The baseline model was `simulated_GD` ∼ `sentN+tokenN+freq` ∗ `length`. The ΔLogLik is denoted as "Effect of syntactic category." A lower score means that the LM lacked the property of varying processing difficulty with respect to the syntactic category.

Results: The results are shown in Figure 4.5. First, the higher psychometric predictive power the LMs exhibit, the greater the effect of syntactic category on surprisals (left part in Figure 4.5). This means that, depending on the syntactic category of the word they processed, LMs with high psychometric predictive power computed surprisals with a more nonuniform trend. The right part of Figure 4.5 shows that, as PPL decreases below a certain value ($\mathrm{PPL} \sim 400$), the Japanese LMs compute surprisals that obscure the nonuniform trends with respect to the syntactic category of words.[9] This trend supports our hypothesis that tuning to LM objectives obscures the human-like nonuniformity of the processing difficulty. Even though LMs that are not fully tuned to the LM objective ($\mathrm{PPL} \sim 400$) acquire human-like trends with respect to syntactic category, these biases tend to be lost by further lowering their PPL.

[9]The correlation between PPL and the effect of syntactic category in the LMs with PPL less than 400 was 0.45 and 0.34 with Pearson's r and Spearman's ρ, respectively.

4.6 Conclusion

This study has investigated whether the recent reports on the psychometric predictive power of LMs can be generalized across languages. Our initial investigation has re-examined the recent report—*the lower PPL a LM has, the more human-like the LM is*—using Japanese eye movement data. Our experiments have demonstrated a surprising lack of universality of this report; lower perplexity is not always human-like. This discrepancy of the results between the languages reinforces the need for the cross-lingual evaluation of the psychometric predictive power of LMs. The discussion considers potential factors that make the observation different across languages from the viewpoint of the uniform information density hypothesis. We believe that this is an important first step for seeking a language-agnostic model of human sentence processing. Hopefully, this study encourages researchers to further investigate the universality of human language processing across languages. In addition, to summarize the results in Chapter 3 and this chapter, it is demonstrated that Japanese LMs exhibit human-like behavior at the sentence level, but their word-by-word behavior deviates from the human-like behavior.

4.7 Preliminary experiments in Section 4.5

Which linguistic factor is helpful for explaining the difference in gaze duration? We conducted experiments using linguistic annotation in the BCCWJ-EyeTrack. Following the existing studies, we checked the separate effect of syntactic category, semantic category (Asahara and Kato, 2017), and a particular aspect of hierarchical syntactic structure (i.e., the anti-locality effect) (Asahara et al., 2016). Specifically, we used the factors, `syn_category`, `sem_category`, and `n_dependents`, shown in Table 4.2. For each factor, we inspect the separate effect of each factor for modeling gaze duration. As Eq. 4.4, we first modeled the gaze duration using each factor (`factor_X`):

$$\texttt{GD} \sim \texttt{factor\_X} + \texttt{sentN} + \texttt{wordN} + \texttt{freq} * \texttt{length} \ . \tag{4.5}$$

Then, we calculated the ΔLogLik between X and a baseline model. The baseline model was $\texttt{GD} \sim \texttt{sentN} + \texttt{wordN} + \texttt{freq} * \texttt{length}$.

The ΔLogLik for each collection of factors are shown in 4.3. We found that syntactic category is the most influential factor for modeling gaze duration, at least in this experiment.

Table 4.3 The separate effect of each linguistic annotation for modeling gaze duration.

linguistic property	ΔLogLik
syntactic category	58.37
semantic category	17.08
number of dependents	13.84

Chapter 5

Memory-based View of the Gap between Incremental Sentence Processing by Humans and Neural Language Models

5.1 Introduction

The previous chapter highlighted the discrepancy between surprisals computed by accurate LMs and human reading behavior and explored the cause from an information-theoretic view. This chapter further investigates discrepancies between LMs and human sentence processing from psycholinguistic perspectives.

Again, computational models of human incremental sentence processing have long been a subject of psycholinguistic study (Crocker, 2010). One account of sentence processing is a forward-looking, expectation-based one; it assumes surprisal as the predictor of incremental processing cost. Recent studies thus compared computational models' surprisal, i.e., next-word prediction probability, to human reading behavior, e.g., gaze duration (Hale, 2001; Levy, 2008; Smith and Levy, 2013). Some such analyses found discrepancies between LMs and humans (Kuribayashi et al., 2021; Wilcox et al., 2021). For example, modern neural LMs underestimated the processing cost of specific syntactic constructions (e.g., filler-gap dependency) (van Schijndel and Linzen, 2021; Wilcox et al., 2021). In addition, simple n-gram LMs often exhibit a good parallel with human reading behavior, despite the n-gram LMs' rel-

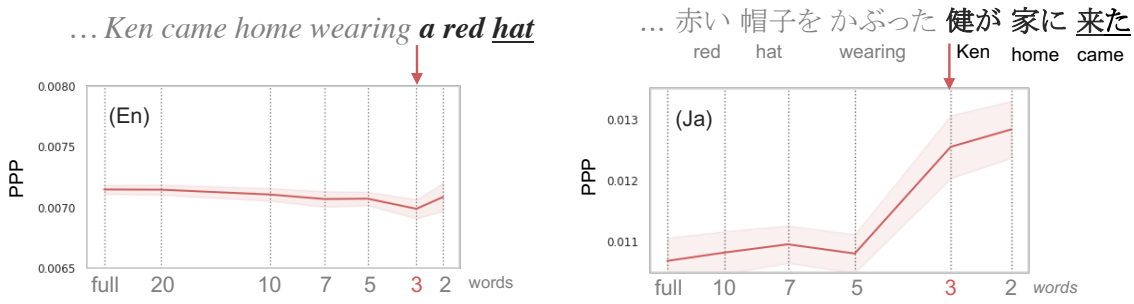


Fig. 5.1 Psychometric predictive power (PPP) of n-gram surprisal computed by neural LMs. A higher value on the Y-axis represents that human reading behavior is better simulated by the corresponding surprisal. The shaded area represents one standard deviation confidence interval. The X-axis corresponds to the context length. Limiting the context length either did not substantially hurt or rather improved the psychometric predictive power.

atively low accuracy on next-word prediction (Goodkind and Bicknell, 2018; Wilcox et al., 2020).

We suspect that this discrepancy could stem from the excessively large working memory capacity of modern LMs compared to humans. While psychological research argues that human working memory has a capacity of storing only seven (Miller, 1956) or even four chunks (Cowan, 2001), the inductive bias of neural LMs assumes access to hundreds or thousands of context tokens in parallel. Similarly, Merkx and Frank (2021) characterized Transformer-based processing as consistent with a cue-based sentence processing theory, emphasizing that typical variants of the theory also assume a memory decay (Lewis and Vasishth, 2005; Lewis et al., 2006), and such a human-like prior was not explicitly implemented.

In this study, we test our hypothesis by introducing input context restrictions mimicking human memory limitations and observing their effectiveness for modeling reading behavior data in two typologically different languages. Specifically, we compared n-gram-based surprisal estimated by neural LMs to human gaze duration while reading naturally occurring text in English and Japanese. Although recent psycholinguistic studies typically compare count-based n-gram LMs and neural full-context LMs (Goodkind and Bicknell, 2018; Kuribayashi et al., 2021; Wilcox et al., 2020), the effect of context limitation is not exactly analyzed, for example, because there are at least two orthogonal differences between the models: neural-based v.s. count-based, and limited-context v.s. full-context.

Overall, we found that limiting the context information either did not hurt or even improved the cognitive plausibility of neural LMs in modeling naturalistic reading data (Figure 5.1). This implies that a much smaller context than previously assumed might be cross-

lingually sufficient for simulating human behavior. More specifically, in Japanese, a clear trend emerged: the little amount of context leads LMs to human-like behavior on average. Although it is generally claimed that the reading behavior in the Japanese (head-final) language conflicts with a memory-based effect (e.g., anti-locality), our results suggest that memory limitations still come into play. We also found a mixed trend in the English data, implying that longer-context inputs are still sometimes necessary to achieve human-like reading behavior. Further analysis revealed that this is indeed the case in specific syntactic constructions, such as long subject-verb dependencies (Section 5.5).

To summarize, our results emphasize that full-context surprisal is a sub-optimal choice for explaining reading behavior; memory-limited n-gram surprisal (**memory bias**) and selective long-context access aligned with specific syntactic constructions (**syntactic bias**) could be more plausible.

5.2 Background

5.2.1 Incremental sentence processing

Human readers incrementally process text and exhibit different processing costs (e.g., reading time) for different tokens. Psycholinguistic theories on what influences such processing cost are mainly categorized into expectation-based and memory-based theories.

Expectation-based theories claim that humans predict upcoming words during incremental sentence processing (Clark, 2013). Recent studies have extensively analyzed this expectation-based aspect, comparing surprisal ($-\log p(\mathrm{word}|\mathrm{context})$) to human reading behavior (Hale, 2001; Levy, 2008).

Memory-based theories assert that human sentence processing is constrained by a limited working memory (Cowan, 2001; Lewis and Vasishth, 2005; Lewis et al., 2006; Miller, 1956). One fundamental view is that retrieving information from linearly distant items (i.e., processing long dependency) overloads human memory (Gibson, 2000). Furthermore, cross-linguistic studies reported that different languages incur different memory decay (Frank et al., 2016; Husain et al., 2014; Konieczny, 2000; Vasishth et al., 2010).

Integration of expectation and memory: Recently, Futrell et al. (2020) proposed to integrate both aforementioned theories. They theoretically introduced the concept of *lossy-context surprisal* ($-\log p(\mathrm{word}|\mathrm{noisy\ context})$), i.e., that the expectation of the next word

calculated with *noisy* context information should better predict human reading behavior than with complete context. The n-gram LMs could be a special case of lossy-context surprisal. Whereas the main focus of the study was the theoretical derivation of lossy-context surprisal, our study complements their claim with empirical experiments of modeling human naturalistic reading behavior. To observe language-dependent effects, we also include cross-linguistic analyses using languages with typologically different constructions: English (SVO) and Japanese (SOV).

Recent psycholinguistic studies typically compare count-based LMs and neural full-context LMs (Goodkind and Bicknell, 2018; Wilcox et al., 2020). In such an analysis, the effect of context limitation is not exactly analyzed because there are at least two orthogonal differences between the two models: neural-based v.s. count-based, and limited-context v.s. full-context. In this study, we control the input of the neural language model and analyze the context restrictions separately.

5.3 Method

Our experiments investigate how human-like neural LMs become with more or less context at their input. Specifically, we measured the *psychometric predictive power* of LMs with limited context access mirroring limited human working memory. This chapter follows the settings in Chapter 4 except that surprisal values conditioned with limited contexts instead of full contexts are used.

5.3.1 Lossy-context surprisal

Given a sentence consisting of symbols $[w_0, \cdots, w_n]$, the *surprisal* of the symbol w_i in its preceding context $c_{<i} = [w_0, \cdots, w_{i-1}]$ is computed by Left-to-right LMs θ as follows:

$$I(w_i, c_{<i}) = -\log p_\theta(w_i | \mathrm{BOS} \circ [w_0, \cdots, w_{i-1}]) \ , \tag{5.1}$$

where BOS denotes a special symbol representing the beginning of a sequence and $\circ$ is a concatenation function. Then, we control the LMs' access to contextual information (i.e., working memory) by externally deleting the input symbols of LMs with a particular pattern. Formally, we define the lossy-context surprisal of a word w_i as follows:

$$I_{\mathrm{lossy}}(w_i, c_{<i}, f) = -\log p_\theta(w_i | \mathrm{BOS} \circ f(c_{<i})) \ , \tag{5.2}$$

Table 5.1 An example of training data of neural LMs.

<s> _Prior _to _the _performance , _some _of _the _show <b> _on _4 _March _1990 _with _a _concert _performed _by _Ell a _Fitzgerald _at _the _Royal _Albert _Hal <b> _or y z omy ine _rodents _from _the _Pleistocene _of _B ona ire , _West _Indies . <s> _The _Har row _&

where the noising function f is applied to context c. Our experiment specifically investigates *n-gram surprisal* that is computed with immediately preceding $n-1$ words as context. That is, f leaves only the last several symbols in computing surprisal.

While LMs process text subword-by-subword, gaze duration is annotated in larger segments. Same as Section 4.3, given a sequence of subwords $[w_0, w_1, \cdots, w_n]$, the lossy-context surprisal of the k-th segment $s_k = [w_l, w_{l+1}, \cdots, w_m]$ was calculated as the cumulative surprisal of constituent subwords:

$$I_{\text{lossy}}(s_k, c_{<l}, f) = \sum_{j=l}^{m} I_{\text{lossy}}(w_j, c_{<j}, f) \ . \tag{5.3}$$

5.3.2 Language models

We used the Transformer LMs for computing the n-gram surprisal. The LM training setting is the same as TRANS-SM introduced in Section 4.3 except that training data is rearranged. When computing n-gram surprisal, the LMs have to predict the upcoming words with severely limited context from the middle of a sentence. However, such prediction is rarely enforced during ordinal LM training, thus we augmented such data points. Specifically, we randomly split each sentence into two sub-sequences; here, the former begins with a begin-of-sentence special token (<s>), and the latter begins with a special token representing the breakpoint (<b>). Then, training data is created by randomly patching these sub-sequences as shown in Table 5.1. In this data, the data points following <b> enforce LMs to predict the upcoming words using severely limited context without any prior token position within a sentence. The BOS token in Eq. 5.2 corresponds to the <b> token.

5.3.3 Psychometric predictive power

We used the same datasets as Chapter 4, but we included the Japanese data points that satisfy (i) the last segment in a line or (ii) contains punctuation. We fear that applying these criteria reduces the data points corresponding to the main verb, which could be important data points

Table 5.2 Comparison of PPP of N-gram surprisal (mean ± standard deviation). Values are multiplied by 1000. N denotes the length of the input: N-1 preceding context segments and one target segment.

N	En	Ja
2	7.1 ± 0.1	$\mathbf{12.8 \pm 0.5}$
3	7.0 ± 0.1	12.5 ± 0.5
5	7.1 ± 0.1	10.8 ± 0.3
7	7.1 ± 0.1	11.0 ± 0.3
10	7.1 ± 0.1	10.8 ± 0.3
all	7.1 ± 0.0	10.7 ± 0.4

especially in analyzing memory-based accounts of sentence processing. We used 217,876 data points from the Dundee Corpus and 9,217 from BCCWJ-EyeTrack.

5.4 Does limiting context length make LMs more human-like?

Motivated by the memory-based theory of sentence processing, we hypothesized that shorter contexts might bring LM surprisal closer to human reading behavior. To test our hypothesis, we compared the PPP of surprisal given by LMs conditioned on n-1 preceding words (not subwords), henceforth referred to as n-gram surprisal. We compared $N \in \{2, 3, 5, 7, 10\}$-grams and full, where full refers to using the entire sentence (up to w_i) as context. As the main focus of this study is sentence-level syntactic processing, only the context within the same sentence is used in all the settings.

Table 5.2 (visualized in Figure 5.1) shows the PPP of n-gram surprisal in relation to input length n. Overall, the results are encouraging—using a shorter context did not hurt the human-likeness in English and even improved it in Japanese. This supports the use of shorter-context LMs as a step closer to a computational model of human sentence processing in this cross-linguistic setting. This indicates that self-attention in Transformer might build excessive working memory. Specifically, the Japanese results had a clear trend that limited context leads LMs to human-like behavior.

In contrast, in English, PPP seems to remain mostly unaffected by different context lengths. The English results are arguably equally surprising on their own—they imply that there is no set human-like working memory capacity. One possible explanation for why PPP seems unchanging in English is that both shorter and longer contexts contribute to matching human reading behavior, depending on the context. Such an adaptive memory re-

trieval would be in line with the concept of "good-enough processing" (Ferreira and Lowder, 2016); human readers adaptively store and retrieve as little context information as possible to achieve the lowest adequate level of precision in sentence processing. To better understand our results, we further investigated *when* longer context becomes more or less useful for modeling human reading behavior (Section 5.5).

5.5 When does limiting/increasing context length make LMs human-like?

5.5.1 Methods

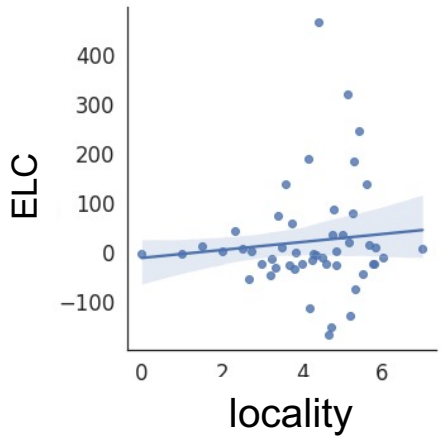
We searched for subsets $\mathscr{D}$ of the corpus where longer context information was (in)effective for simulating human reading behavior. We quantified the effect of long context (ELC) on $\mathscr{D}$ using a regression model. We measured the difference between the mean value of squared residuals of reading time modeling $\mathscr{L}(l_i, \mathscr{D})$ with 2-gram (short) and all-gram (long) surprisal:

$$\mathrm{ELC}(\mathscr{D}) = \mathscr{L}(l_2, \mathscr{D}) - \mathscr{L}(l_{\mathrm{all}}, \mathscr{D}) \ . \tag{5.4}$$

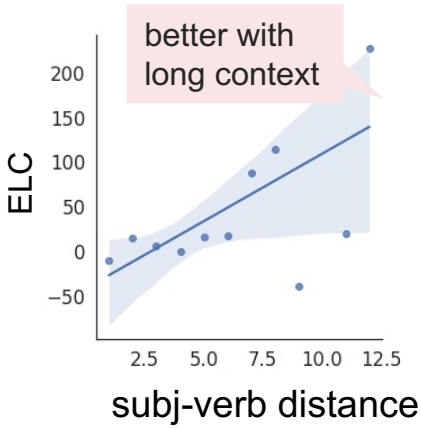
A high ELC value indicates that reading times on $\mathscr{D}$ were worse simulated with short context ($\mathscr{L}_2(\mathscr{D}) \uparrow$) and better simulated with long context ($\mathscr{L}_{\mathrm{full}}(\mathscr{D}) \downarrow$).

Note that the regression models are trained with the entire corpus, and the residuals at the targeted data points $\mathscr{D}$ are used in calculating $\mathscr{L}$. To more focus on the context effect, tokens at the latter part[1] of a sentence were used. Manual linguistic annotations were used in the following analyses (Asahara et al., 2016; Barrett et al., 2015).

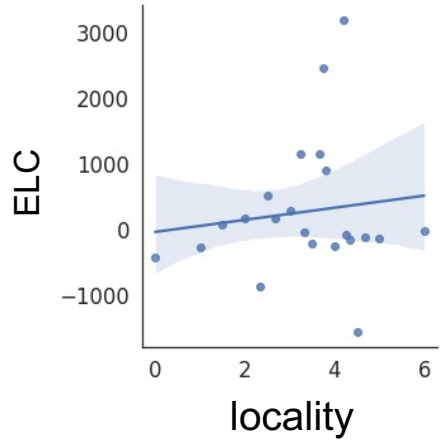
[1] Regarding the median of the word position in sentence was 12 and 7 in the Dundee and BCCWJ-EyeTrack corpus respectively, 13th or later words in the Dundee corpus (20,554 words) and 7th or later words in the BCCWJ-EyeTrack (4,051 words) were used.



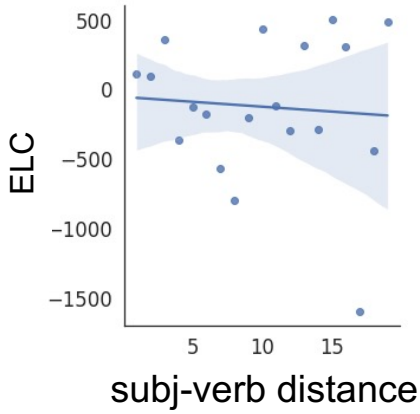
(a) Relationship between the dependency length and ELC score in English.



(b) Relationship between the subject-verb distance and ELC score in English.



(c) Relationship between the dependency length and ELC score in Japanese.



(d) Relationship between the subject-verb distance and ELC score in Japanese.

Fig. 5.2 Relationship between the ELC score and the dependency length/subject-verb distance. The positive correlation between ELC and the subject-verb distance is observed in English. Note that the difference in the range of the X- and Y-axes between languages could be due to language-dependent statistics (e.g., Japanese has long subject-verb distances due to the SOV word order, and *bunsetsu* has a longer gaze duration than English word).

5.5.2 Dependency locality in English

Simple dependency length does not explain the effect of context length. It would be a reasonable assumption that words with distant syntactically-related items requiring farther contextual information will be better modeled by long-context LMs. To verify this, we

grouped the data by the words' dependency length and observed the ELC score for each set. Here, the dependency length denotes the averaged distance to its preceding items with direct syntactic dependency.[2] The example (1) below illustrates the dependency lengths:

(1) You have a cap
0 1 0 $1.5 = \frac{2+1}{2}$

Figure 5.2a shows the relationship between the ELC and dependency length. Surprisingly, there was only a slight difference in the effectiveness of considering long context with respect to dependency length (Pearson's r=0.10).

Subject-verb dependency length in English explains the effectiveness of long context.

Narrowing down the focus to subject-verb dependency length produces a clearer picture. We split the verbal data points (`VB`, `VBD`, `VBG`, `VBN`, `VBP`, and `VBZ`) by their distance to corresponding subjects.[3] For example, the subject-verb distance for **goes** in Example (2) is five:

(2) Ken , my brother , **goes** ...
5

Figure 5.2b shows the relationship between the ELC score and subject-verb distance. The gaze duration for verbs with distant subjects was significantly better modeled with longer-context surprisal (Pearson's r=0.62). Given that only a slight trend existed in the overall dependency length (Figure 5.2a), English readers seem to selectively weigh subject-verb dependency in memory access. Therefore, not any long syntactic dependency but long *subject-verb* dependency is retrieved from working memory during reading, which should be taken into consideration while designing plausible models of working memory. Such sensitivity to dependency type is indeed suggested in existing studies (Demberg and Keller, 2008); specifically, the need for subject advantage in noise design is argued by Futrell et al. (2020). Our results empirically support their claim.

[2]We recognize that there are various definitions of the locality effect (e.g., the distinction between old and new information) (Gibson, 1998; Hawkins, 1994). Regarding the unambiguity of the implementation, this study simply defined the score as the average of the distances to the preceding elements that have a direct syntactic relationship.

[3]Preceding dependent with the `nsubj` relation is regarded as the subject.

5.5.3 Source of the gap between English and Japanese results

Recall that Japanese gaze duration is generally better modeled with short context, whereas English has no such clear trend (Section 5.4). The difference between the languages is highlighted when comparing dependency-related trends.

Subject-verb advantage disappears in Japanese We repeated the analysis of dependency length in Section 5.5.2 using Japanese data. We did not find clear relationships between dependency length and plausible context length (Figure 5.2c and 5.2d); Pearson's correlation coefficient between the ELC and dependency length was 0.12, and that for the subject-verb dependency was -0.08. This suggests that the special effect of long subject-verb dependency was unique to English, while Japanese readers seem to be less sensitive to long subject-verb dependencies than English readers.

Note that the Japanese language does not have the subject-verb agreement and the grammatical subject is frequently omitted due to the prominence of the topic-comment structure (Li and Thompson, 1984), possibly making Japanese readers less conscious of the subject-verb dependencies. This difference could make the English results more mixed than the Japanese results of favoring short context LMs (Figure 5.1). Such language-dependent effect is explored in the psycholinguistic studies (e.g., structural forgetting) (Futrell et al., 2020; Futrell and Levy, 2017; Husain et al., 2014), and the search for a unified formulation of memory design that fills the language-dependent gap could be an important next step.

5.6 Discussion

Japanese anti-locality = strong expectation + memory constraint. In head-final languages like Japanese sentences, it is typically reported that long dependencies do not overload, but *facilitate* sentence processing (i.e., anti-locality effect) Konieczny (2000) as if they overcome the limited memory capacity humans would generally have. The higher PPP of shorter-context surprisal (Section 5.1) clarifies this somewhat puzzling claim; compared to the accurate estimate of surprisal distributions by long-context LMs, Japanese reading behavior still seems to be constrained by locality.

Sensitivity to syntactic categories. In Chapter 4, it was found that accurate LMs become less sensitive to the syntactic category (e.g., noun, verb) of words than humans were. To investigate if our shorter-context input remedied this tendency, we measured if syntactic categories were a good predictor of surprisal using regression modeling in the same manner in

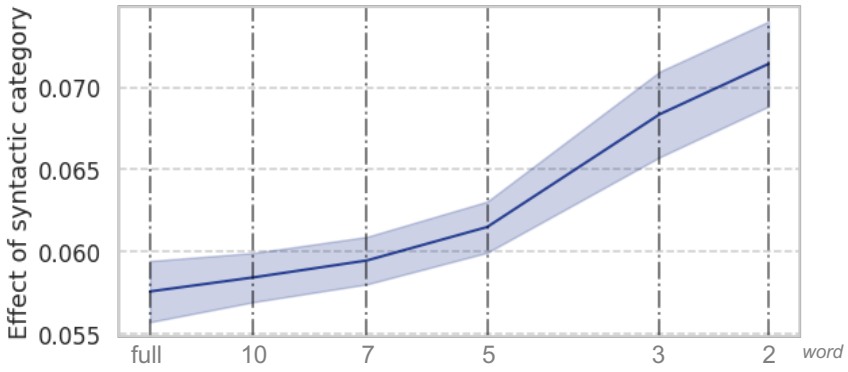


Fig. 5.3 Relationship between the input length (X-axis) and how corresponding n-gram surprisal could be explained by syntactic category factor (Y-axis). The existing study pointed out that the distinct trend of the accurate, less human-like LMs exhibiting a low value of this effect.

Section 4.5. We found that the memory-constrained LMs did partially recover the sensitivity that was lost in LMs using full context (Figure 5.3).

5.7 Conclusions

Although it has been suggested that context-limited (n-gram) surprisal is a strong baseline in reading behavior modeling, there has been a little empirical investigation of what kind of behavior the context limitation actually promotes and what sophisticated noise design could be needed. Our empirical experiments using the input-controlled neural LMs have shown that short context LMs simulate human reading behavior surprisingly well. Further analysis has suggested that the necessary context could be limited to local context and clarifies the additional need for selective context access aligned with a syntactic factor.

Chapter 6

Psycholinguistic Theory as an Inductive Bias for Computational Discourse Processing Model

6.1 Introduction

In the previous chapters, we explored the cross-linguistic universality of cognitive plausibility analysis in the natural language processing (NLP) models. This chapter, on the other hands, broadens the focus on the type of linguistic units. Whereas typical analyses have focused on sentence-level syntactic processing, we explore a computational model for discourse-level processing. Note that this chapter is in a little more engineering-oriented direction than the previous chapters; considering cognitive plausible behavior of existing discourse parsing model, we aim to lead these models to more linguistically proper processing by exploiting cognitive viewpoint for discourse processing as a hint.

It has been generally suggested that humans *a priori* have an inductive bias that facilitates efficient language acquisition. Therefore, existing studies have investigated inductive bias that prompts neural NLP models to achieve good linguistic generalization. For example, Dyer et al. (2016) imposed a sentence-level symbolic, hierarchical bias on neural model architecture and achieved better performance in language modeling.

To take a step further, this study investigates an effective inductive bias for computerized **discourse** processing rather than sentence-level processing. Language communication takes place through discourse, coherent units consisting of sentences; thus, handling discourse-level phenomena with computerized models is an important goal from both scientific and

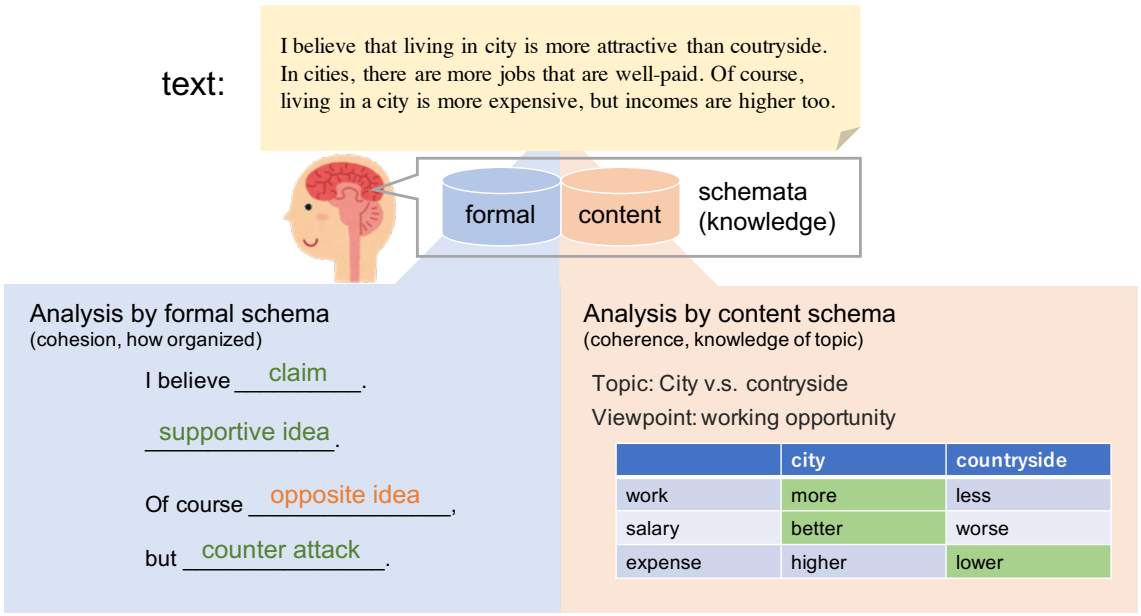


Fig. 6.1 Schema theory claims that humans have several schemata (background knowledge) in the brain/mind, and these affect the interpretation of a text. It is suggested that there are at least two types of schema required in discourse processing; formal and content schema. Formal schema corresponds to knowledge of text organization patterns (left, blue part). Content schema corresponds to the knowledge of the topic (right, orange part).

engineering perspectives. The NLP fields struggle with handling discourse with computers, for example, the existing neural discourse parsing model exhibits cognitively implausible behaviors such as predicting shallow, near-linear trees despite the plausible structures being deep, hierarchical ones. We hypothesized that this stems from the lack of a human-like inductive bias in the neural model leading it to capture hierarchical generalization.

This study focuses on human discourse processing, specifically the **schema theory** (Rumelhart, 1980). We explore an effective inductive bias (architecture design) for the neural discourse processing model associated with this theory. The schema theory suggests that texts do not convey their meaning by themselves, but the meaning depends on the past experiences of the reader (background knowledge). The knowledge structure that the reader has acquired previously is called a schema, and it is organized in the reader's long-term memory. It has been suggested that at least two types of schemata are used to interpret a text (Figure 6.1) (An, 2013; Carrell, 1982; Carrell and Eisterhold, 1983; Chaudron and Richards, 1985). One is a **formal schema**, which is knowledge of the rhetorical organization of texts (e.g., usage of discourse markers). The other is a **content schema**, which is external knowledge (e.g., knowledge of the topic, culture) for tracking the semantic connections of textual content. These schemata complement each other in text comprehension, for example, if a topic of a given

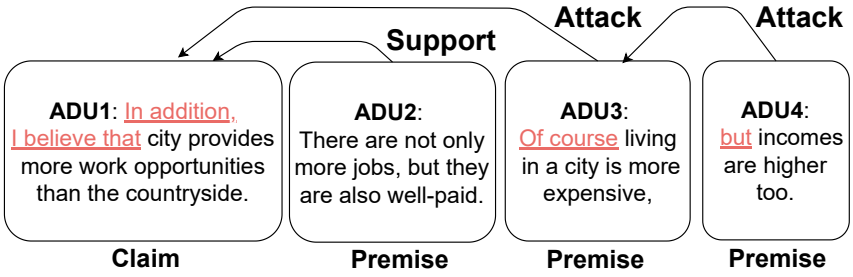


Fig. 6.2 An example of argumentative text and its argumentation structure. Each node corresponds to an argumentative discourse unit, and each edge corresponds to an argumentative relation. The label below each unit (e.g., claim) indicates the type of unit. The label above each edge represents indicates the type of edge. The underlined parts of the text correspond to argumentative markers, and the rest part corresponds to propositions.

text is unfamiliar to the reader (i.e., having imperfect content schema), the reader relies on formal schema (knowledge of a natural organization of text) to interpret the text (Ang, 2014). These schemata are also called cohesion (formal schema) and coherence (content schema) in text linguistics (Halliday and Hasan, 2014). These theories suggest that humans should have a prior ability (inductive bias) to distinguish and organize the type of knowledge (i.e., distinguishing formal and content parts in a text). Therefore, we explore an **inductive bias (architecture design)**, which imposes a neural model on the explicit distinction between formal and content knowledge, for effective computational discourse processing.

This study tackles argumentation structure parsing as a representative of the discourse processing task. Argumentation structure parsing is the task to predict the latent discourse-level structure of an argumentative text. Figure 6.2 shows an example of argumentative text and its corresponding argumentation structure. In this example, a claim "*In addition, I believe that city provides more work opportunities than the countryside.*" (ADU1) and its supportive (ADU2) or opposite (ADU3) ideas are stated. The argumentation structure is represented by a tree (Peldszus and Stede, 2015; Stab and Gurevych, 2017), where the vertices of the graph represent argumentative discourse units (ADUs) and edges of the graph represent argumentative relations between the ADUs.

Regarding the connections between this task and the schemata, for example, there is a typical rhetorical organization (could be formal knowledge) in argumentative texts (Kuribayashi et al., 2017; Peldszus and Stede, 2013). The text in Figure 6.2 has a typical macro-structure of "chain of attack" (ADU2 and 3), where the author of the argumentative text first made a concession with a plausible opposite opinion (ADU2) and then immediately strengthens his argument by counter-attacking it (ADU3).

Additionally, the content schema is considered necessary for capturing semantic connections that provide clues for understanding the structure of discourse. For example, ADU3 and ADU4 in Figure 6.2 share semantically related ideas "living is expensive" and "incomes are higher." Knowledge of a financial topic should be helpful in interpreting the attack relation between the two ADUs. Notably, the dataset of argumentation structure parsing task consists of argumentative texts in several topics. That is, the model should process a text on both unseen (content schema is imperfect) and previously seen topics (both formal and content schema are helpful). Here, a computational model that is aware of both formal-level and content-level perspectives is preferred.

Based on these intuitions, we induce an inductive bias that encourages the models to capture the argumentation structure from multiple perspectives: the formal and content schema. Experiments show that our model with the proposed inductive bias achieves a better discourse parsing performance. Furthermore, our model performs robustly in parsing structures with deep hierarchies, whereas existing models tend to predict shallow and nearly linear trees.

6.2 Related work

6.2.1 Inductive bias

Inductive bias is the set of assumptions that the learners use to make decisions for unseen data (Mitchell, 1980). Humans are considered to have strong biases (e.g., favoring hierarchical generalization), facilitating efficient language acquisition. Currently, investigating the inductive biases of neural network models receives wide attention (Kharitonov and Chaabouni, 2021; McCoy et al., 2018). For example, typical neural architectures in NLP models exhibit a weaker preference for hierarchical generalizations than humans have, and imposing hierarchical inductive bias on network architecture is shown to yield better performance (Dyer et al., 2016; McCoy et al., 2020; Yoshida et al., 2021). Regarding the reports that the existing argumentation parsing model tends to predict shallow trees (Stab and Gurevych, 2017), some human-like bias could be needed in developing the model capturing the hierarchical nature of discourse structure like humans.

6.2.2 Schema theory, cohesion, and coherence

There has long been significant research on the abilities needed to understand the meaning of a text. One theory is schema theory, which holds that reading comprehension is an interaction between a text and the reader's background knowledge (schema) (Rumelhart, 1980). Although the original schema theory is a general theory, in the context of text comprehension,

it states that there are at least two types of schemata: formal schema and content schema (Carrell, 1982; Carrell and Eisterhold, 1983). A formal schema corresponds to the knowledge of text organization (i.e., needed to capture cohesion), and a content schema corresponds to the external knowledge of the topic (i.e., needed to capture coherence). Humans use these schemata in complementary ways; for example, a pedagogical study has reported that a formal schema is particularly exploited by non-native speakers who are unfamiliar with conventional context (content schema) required to comprehend the textual content (Ang, 2014). In this study, we aim to impose an inductive bias on neural models to make them understand a text through multiple perspectives (formal and content schemata), as humans read a text through the lens of multiple schemata.

6.2.3 Argumentation structure parsing

Argumentation structure parsing is the task of extracting the discourse-level structure from a given argumentative text such as a persuasive essay. Based on the argumentation theories (Freeman, 2011; Toulmin, 1958), an annotation scheme has been designed and argumentation structure parsers have been developed (Reed, 2006; Stab and Gurevych, 2017). In recent years, argumentation structures have been widely exploited in applications such as essay scoring systems (Ke et al., 2018; Nguyen and Litman, 2018).

Recently, reliable, high-quality datasets for argument structure analysis have been released (Peldszus and Stede, 2016; Stab and Gurevych, 2017) , and various argument structure parsing models have been proposed on these benchmark datasets. These models can be categorized into (i) models that exploit manually designed features (Afantenos et al., 2018; Peldszus and Stede, 2015; Stab and Gurevych, 2017) and (ii) neural-based models (Eger et al., 2017; Potash et al., 2017). The manual feature-based models exploited features related to argumentation and discourse. Some of these features are associated with a macro-level formal schema of argumentative texts, such as the type and position of discourse connectives in texts. Conversely, neural network-based models have achieved high-performance analysis, although linguistically motivated features have rarely been used. In this study, we integrate these approaches by imposing inductive bias into neural-based models to capture the linguistic features of argumentative texts.

Discourse structure parsing based on rhetorical structure theory (Mann and Thompson, 1987) and discourse relation recognition in the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) style are closely related to argumentation structure parsing. Rhetorical structure theory and PDTB-style discourse analysis analyze discourse in a general domain and predict relations such as *cause*, *contrast*, and *concession* relations among discourse units. In contrast, argument structure parsing narrows down the target domain into argumentative texts

and predicts the functions of discourse units (evidence/argument) and the relations between discourse units (support/refutation) with the goal of persuading others. The connections between rhetorical structure theory and argument structure have been analyzed in recent years (Stede et al., 2016).

6.2.4 Span representation

Span refers to a semantically meaningful unit consisting of one or more words, such as a discourse unit. This study could also be viewed as improving the span representation of the argumentative unit. The design of span representations has received considerable attention in NLP tasks such as syntactic analysis (Kitaev and Klein, 2018; Stern et al., 2017; Wang and Chang, 2016), semantic role assignment (He et al., 2018; Ouchi et al., 2018), coreference analysis (Lee et al., 2017, 2018), and discourse analysis (Li et al., 2016). Recently, a span feature extraction method called **LSTM-minus** was proposed (Wang and Chang, 2016) and applied to various tasks. In this method, for a span (i, j), the difference between the hidden layers of bidirectional Long Short-Term Memory (BiLSTM) corresponding to both start and end token of the span ($\mathbf{h}_j - \mathbf{h}_{i-1}$) is used to compute a contextualized span representation. This study also exploits this contextualization method to induce neural models to capture the macro-structure of texts through multiple levels (formal and content schemata). Contextual information is important in interpreting the meaning of discourse units (Lawrence and Reed, 2019; Nguyen and Litman, 2016). For example, the function of the utterance "Apples are nutritious" is "support" if the utterance follows the claim "we should eat apples every morning," but the function could be "attack" if the utterance is for the claim "Fruits are bad for you."

6.3 Task and model

Section 6.3.1 describes the overview of the task. Sections 6.3.2, 6.3.3, 6.3.4, and 6.3.5 denote the model settings. Note that Section 6.3.4 proposes our cognitively motivated inductive bias.

6.3.1 Task overview

Argument structure parsing includes (i) argumentative unit segmentation, (ii) argumentative relation identification, (iii) argumentative relation type classification, and (iv) argumentative unit type classification. Based on the existing studies (Niculae et al., 2017; Peldszus and

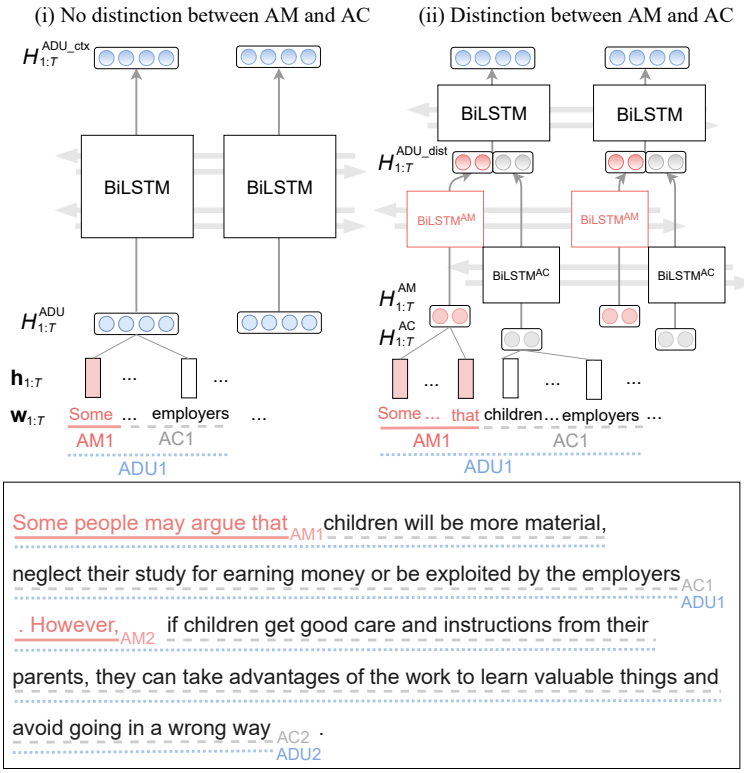


Fig. 6.3 Illustration of argumentation parsing models. The left part explains the model that does not distinguish formal and content schema, and the right part illustrates the model that does distinguish them. The below text is an example of an input argumentative text, where ADUs (argumentative discourse units), AMs (argumentative markers), and ACs (argumentative components) are underlined.

Stede, 2015; Potash et al., 2017), this study assumes that the argumentative units are pre-segmented. The argument structure parsing model takes an argumentative text as input and its encoder computes the span representation of each argument unit (Sections 6.3.2, 6.3.3, and 6.3.4). Thereafter, the decoder classifies (a) the presence or absence of an argumentation relation between each argumentative unit, (b) the type of each argumentative relation, and (c) the type of each argumentation unit (Section 6.3.5)

6.3.2 Span representations

An argumentative text consists of T words $w_{1:T} = (w_1, w_2, \cdots, w_T)$, where K ADU spans $S^{\text{ADU}}_{1:K} = (s^{\text{ADU}}_1, s^{\text{ADU}}_2, \cdots, s^{\text{ADU}}_K)$ exist. Span s^{ADU}_k is denoted as (i_k, j_k); i and j are word index ($1 \leq i_k \leq j_k \leq T$). In this study, each ADU s^{ADU}_k is further divided into two parts: argumentative marker s^{AM}_k and proposition s^{AC}_k. We assume that the proposition part follows an argumentative marker in each ADU (Figure 6.3). That is, each ADU $s^{\text{ADU}}_k = (i_k, j_k)$ is

decomposed into argumentative marker $s_k^{\mathrm{AM}} = (i_k, \ell_k)$ and proposition $s_k^{\mathrm{AC}} = (\ell_k + 1, j_k)$, where $i_k \leq \ell_k < j_k$ holds. When there are K ADUs in argumentative text, there also exist K argumentative markers and propositions. Henceforth, an argumentative marker is denoted as AM, and a proposition is denoted as AC (argument component).

6.3.3 Span representations based on LSTM-minus

First, we incorporate a LSTM-minus span representation (Wang and Chang, 2016) to the argumentation structure parser (left part in Figure 6.3). In our baseline model, AC and AM are not distinguished. Span representation $\mathbf{h}_k^{\mathrm{ADU}}$ for ADU span $s_k^{\mathrm{ADU}} = (i_k, j_k)$ is computed as follows:

$$\mathbf{w}_{1:T} = f^{\mathrm{emb}}(w_{1:T}) \ , \tag{6.1}$$

$$\mathbf{h}_{1:T} = \mathrm{BiLSTM}(\mathbf{w}_{1:T}) \ , \tag{6.2}$$

$$\mathbf{h}_k^{\mathrm{ADU}} = [\overrightarrow{\mathbf{h}}_{j_k} - \overrightarrow{\mathbf{h}}_{i_k-1}; \overleftarrow{\mathbf{h}}_{i_k} - \overleftarrow{\mathbf{h}}_{j_k+1}; \overrightarrow{\mathbf{h}}_{i_k-1}; \overleftarrow{\mathbf{h}}_{j_k+1}; \boldsymbol{\phi}(w_{i_k:j_k})] \ . \tag{6.3}$$

Here, the embedding layer f^{emb} converts the input symbols $w_{1:T}$ into word embeddings $\mathbf{w}_{1:T}$. Given the input embeddings $\mathbf{w}_{1:T}$, BiLSTM layer computes intermediate word representations $\overrightarrow{\mathbf{h}}_{1:T}$ and $\overleftarrow{\mathbf{h}}_{1:T}$. The, span representation $\mathbf{h}_k^{\mathrm{ADU}}$ for the span s_k^{ADU} is computed by the representations from the BiLSTM. Here, the operation ";" denotes the vector concatenation operation. This model is called **LSTM model**.

Note that an additional feature vector $\boldsymbol{\phi}(w_{i:j})$ is also concatenated to the span representation in Eq. 6.3, following existing study (Potash et al., 2017). Specifically, we first created the feature vector consisting of the following information for each ADU:

- Discrete bag-of-words vector
- Span representation from word embeddings, where embeddings are aggregated by average, max, and min pooling (concatenated)
- One-hot-vector representing ADU position
- One-hot-vector representing paragraph position in the entire essay
- One-hot-vector representing whether the ADU is in the first paragraph, final paragraph, or other paragraphs.

Then, we computed $\boldsymbol{\phi}(w_{i:j})$ by transforming the above feature vector to 512-dimensional representation by a single fully connected layer.

6.3.4 Distinguishing formal and content schema

To implement an inductive bias for capturing formal and content schema-level flow in text, we extend the computation of span representation (right part of Figure 6.3).

We regard that the formal aspect (the topic-agnostic rhetorical organization) of the argumentative text is realized by the sequence of AMs (e.g., "I believe that"→"Of course"→"but"). By contrast, the aspect related to the content schema (external knowledge about the argued topic) is considered to be realized by the sequence of ACs. Thus, we separately implement the module capturing the sequence of AMs and that of ACs in the model architecture.

First, we decompose each ADU s_k^{ADU} into its AM s_k^{AM} and AC s_k^{AC}. Then, following Eq. 6.2, each s_k^{AM} and s_k^{AC} are encoded into span representation using BiLSTM. Here, K sequence of span representations for AMs are denoted as $H_{1:K}^{\mathrm{AM}} = (\mathbf{h}_1^{\mathrm{AM}}, \mathbf{h}_2^{\mathrm{AM}}, \cdots, \mathbf{h}_K^{\mathrm{AM}})$; similarly, the sequence of AC representations are denoted as $H_{1:K}^{\mathrm{AC}} = (\mathbf{h}_1^{\mathrm{AC}}, \mathbf{h}_2^{\mathrm{AC}}, \cdots, \mathbf{h}_K^{\mathrm{AC}})$. Next, BiLSTM$^{\mathrm{AM}}$ and BiLSTM$^{\mathrm{AC}}$ encode the different types of contexts into $H_{1:K}^{\mathrm{AM}}$ and $H_{1:K}^{\mathrm{AC}}$, respectively:

$$H_{1:K}^{\mathrm{AM\_ctx}} = \mathrm{BiLSTM}^{\mathrm{AM}}(H_{1:K}^{\mathrm{AM}}) \ ,$$
$$H_{1:K}^{\mathrm{AC\_ctx}} = \mathrm{BiLSTM}^{\mathrm{AC}}(H_{1:K}^{\mathrm{AC}}) \ .$$

Here, $H_{1:M}^{\mathrm{AM\_ctx}} = (\mathbf{h}_1^{\mathrm{AM\_ctx}}, \mathbf{h}_2^{\mathrm{AM\_ctx}}, \cdots, \mathbf{h}_K^{\mathrm{AM\_ctx}})$ and $H_{1:K}^{\mathrm{AC\_ctx}} = (\mathbf{h}_1^{\mathrm{AC\_ctx}}, \mathbf{h}_2^{\mathrm{AC\_ctx}}, \cdots, \mathbf{h}_K^{\mathrm{AC\_ctx}})$ denote the sequence of contextualized representations for AMs and ACs, respectively. We expect that knowledge on formal schema is stored in BiLSTM$^{\mathrm{AM}}$, and knowledge on content schema is in BiLSTM$^{\mathrm{AC}}$. Lastly, span representation for an ADU s_k^{ADU} is computed by $\mathbf{h}_k^{\mathrm{AM\_ctx}}$, $\mathbf{h}_k^{\mathrm{AC\_ctx}}$, and feature vector $\boldsymbol{\phi}(w_{i_k:j_k})$ as follows:

$$\mathbf{h}_k^{\mathrm{ADU\_schema}} = [\mathbf{h}_k^{\mathrm{AM\_ctx}}; \mathbf{h}_k^{\mathrm{AC\_ctx}}; \boldsymbol{\phi}(w_{i_k:j_k})] \ .$$

This model is called **LSTM+dist model**. The performance gain attributed to the induced inductive bias is obtained by comparing the performance of LSTM and LSTM+dist models.

6.3.5 Output layer

The output layer is common to both LSTM and LSTM+dist models. Let the sequence of K ADU span representations be $H_{1:K}^{\mathrm{ADU}} = (\mathbf{h}_1^{\mathrm{ADU}}, \mathbf{h}_2^{\mathrm{ADU}}, \cdots, \mathbf{h}_K^{\mathrm{ADU}})$. Here, each ADU span representation $\mathbf{h}_k^{\mathrm{ADU}}$ corresponds to $\mathbf{h}_k^{\mathrm{ADU}}$ in Section 6.3.3 (LSTM model) or $\mathbf{h}_k^{\mathrm{ADU\_dist}}$ in Section 6.3.4 (LSTM+dist model). Using BiLSTM, inject ADU-level context information

into each ADU span representation (Eq. 6.4).

$$H_{1:K}^{\text{ADU\_ctx}} = \text{BiLSTM}(H_{1:K}^{\text{ADU}}) \ . \tag{6.4}$$

For a fair comparison, the parameter size of LSTM and LSTM+dist models are almost the same (Table 6.1). The computed ADU span representations $H_{1:K}^{\text{ADU\_ctx}} = (\mathbf{h}_1^{\text{ADU\_ctx}}, \mathbf{h}_2^{\text{ADU\_ctx}}, \cdots, \mathbf{h}_K^{\text{ADU\_ctx}})$ are then used to compute the output distribution for each subtask. Note that the subtasks consist of (i) argumentative relation identification, (ii) argumentative relation type classification, and (iii) argumentative unit type classification. For simplicity, span representation for ADU $\mathbf{h}_k^{\text{ADU\_ctx}}$ is henceforth denoted as $\mathbf{h}_k$.

Argumentative relation identification (RI) The output of this task is an unlabeled tree for a given ADUs. Tn this layer, the probability that s_m^{ADU} has a directed link to s_n^{ADU} is computed as follows:

$$\begin{aligned} \text{score}_{m,n}^{\text{link}} &= \mathbf{w}^{\text{link}} \cdot [\mathbf{h}_m; \mathbf{h}_n; \mathbf{h}_m \odot \mathbf{h}_n; \boldsymbol{\phi}(m, n)] \ , \\ \text{P}(n|s_m^{\text{ADU}}) &= \frac{\exp(\text{score}_{m,n}^{\text{link}})}{\sum_{n'=1}^{K} \exp(\text{score}_{m,n'}^{\text{link}})} \ , \end{aligned}$$

Here, $\mathbf{w}^{\text{link}}$ is learnable parameters, the operation $\odot$ denotes hadamard product, and $\boldsymbol{\phi}(m, n)$ is one-hot vector representing the relative distance between m and n. In decoding the tree, we imposed the tree constraints by applying the Chu-Liu/Edmonds algorithm to the computed probabilities $\text{P}(n|m)$.

Argumentative relation type classification (RTC) The probability that span s_o^{ADU} is classified as type r is computed as follows:

$$\text{score}_{o,r}^{\text{link-type}} = \mathbf{w}_r^{\text{link-type}} \cdot \mathbf{h}_o + b_r^{\text{link-type}} \ , \tag{6.5}$$

$$\text{P}(r|s_o^{\text{ADU}}) = \frac{\exp(\text{score}_{o,r}^{\text{link-type}})}{\sum_{r' \in \mathscr{R}} \exp(\text{score}_{o,r'}^{\text{link-type}})} \ , \tag{6.6}$$

Here, $\mathbf{w}_r^{\text{link-type}}$, $b_r^{\text{link-type}}$ are learnable parameters, and let $\mathscr{R}$ be {SUPPORT, ATTACK}. Following existing studies, the RTC task is formulated as the classification of the type of outgoing relation from each ADU, given an ADU representation. Note that one of the datasets, persuasive essay corpus (PEC) (Stab and Gurevych, 2017), has stance annotation for a particular type of unit called CLAIM. Following the existing study, this stance is regarded as the outgo-

ing relation type for CLAIM, and this stance classification for CLAIM units are also included in this task.

ADU type classification (ATC) Following Eq. 6.5 and 6.6, the probability that span s_o^{ADU} is classified as ADU type r $\text{P}(r|s_o^{\text{ADU}})$ is computed. Let $\mathscr{R}$ be {MAJORCLAIM, CLAIM, PREMISE}. Note that learnable parameters $\mathbf{w}_r^{\text{ac-type}}$ and $b_r^{\text{ac-type}}$ are different from those in Eq. 6.5 and 6.6.

6.3.6 Training

Training dataset $\mathscr{D}$ is denoted as follows:

$$\begin{aligned}
\mathscr{D} &= \{(X,\, Y^{\text{link}},\, Y^{\text{link-type}},\, Y^{\text{ac-type}})_d\}_{d=1}^{|\mathscr{D}|}\ , \\
X &= \{\, w_{1:T},\, S_{1:K}^{\text{ADU}},\, S_{1:K}^{\text{AM}},\, S_{1:K}^{\text{AC}}\}\ , \\
Y^{\text{link}} &= \{\, h_1, \cdots, h_K\}\ , \\
Y^{\text{link-type}} &= \{\, t_1, \cdots, t_K\}\ , \\
Y^{\text{ac-type}} &= \{\, r_1, \cdots, r_K\}
\end{aligned}$$

Here, $h_k \in \{\texttt{root}, 1, 2, \cdots, K\}$, $t_k \in$ {SUPPORT, ATTACK}, and $r_k \in$ {MAJORCLAIM, CLAIM, PREMISE}[1]. Following the reports that multi-task training of multiple subtasks is effective (Potash et al., 2017; Stab and Gurevych, 2017), we solved the three subtasks jointly. The training loss is as follows:

$$\mathscr{L}(\theta) = \sum_{(X,\, Y^{\text{link}},\, Y^{\text{link-type}},\, Y^{\text{ac-type}})\in\mathscr{D}} -\Big(\alpha\, \ell_\theta^{\text{link}}(X,\, Y^{\text{link}}) + \beta\, \ell_\theta^{\text{link-type}}(X,\, Y^{\text{link-type}}) + (1-\alpha-\beta)\, \ell_\theta^{\text{ac-type}}(X,\, Y^{\text{ac-type}})\Big),$$

Here, α and β are hyperparameters interporating the losses from three subtasks ($\alpha \geq 0$, $\beta \geq 0$, $\alpha + \beta \leq 1$). Loss functions for three subtasks are as follows:

$$\begin{aligned}
\ell_\theta^{\text{link}}(X,\, Y^{\text{link}}) &= \sum_{k\in\{1,2,\cdots,K\}} \log \text{P}_\theta(h_k|s_k^{\text{ADU}})\ , \\
\ell_\theta^{\text{link-type}}(X,\, Y^{\text{link-type}}) &= \sum_{k\in\{1,2,\cdots,K\}} \log \text{P}_\theta(t_k|s_k^{\text{ADU}})\ , \\
\ell_\theta^{\text{ac-type}}(X,\, Y^{\text{ac-type}}) &= \sum_{k\in\{1,2,\cdots,K\}} \log \text{P}_\theta(r_k|s_k^{\text{ADU}})\ .
\end{aligned}$$

[1]The MAJORCLAIM class does not exist in the microtext corpus (Peldszus and Stede, 2016).

Table 6.1 Hyperparameters of the models.

Parameter	Value
Word embedd.	
- Glove	300 dim
- ELMo	1024 dim
- BERT	1024 dim
- RoBERTa	1024 dim
- XLNET	1024 dim
BiLSTMs	256 dim (300 dim in LSTM models)
Mini-batch size	16
Optimizer	Adam
Learning rate	0.001
Epochs	500 (1000 in the MTC)
Loss function	
- α	0.5
- β	0.25
Dropout	
- output layer	0.5 (0.9 in the MTC)
- BiLSTMs	0.1 (0.9 in the MTC)
- word embeddings	0.1

Table 6.1 shows the hyperparameter of our models. One layer BiLSTM was used except that two-layer BiLSTMs with increased hidden representation size are used in the LSTM models for a fair comparison with the LSTM+dist model.

6.4 Experiments

6.4.1 Experimental setup

Datasets This study uses the persuasive essay corpus (PEC) (Stab and Gurevych, 2017) and arg-microtext corpus (MTC) (Peldszus and Stede, 2016). The PEC consists of 402 essays (1,833 paragraphs) posted on an online forum[2]. Argumentation structure is annotated for each paragraph in the PEC; thus, we solved the task using each paragraph as input. The training/evaluation data split defined by (Stab and Gurevych, 2017) was used, and 10% of the randomly selected training data were used as the development data. Following the existing study (Potash et al., 2017), for scores in PEC, we reported the average of three trials conducted with different seeds.

[2]https://essayforum.com/

Table 6.2 Statistics of the PEC and MTC.

		PEC	MTC
Argumentative text	#texts	402	112
	#paragraphs	1,833	112
	#sentences	7,116	-
	#tokens	147,271	-
ADU	total ADUs	6,089	576
	MajorClaim	751	-
	Claim	1,506	112
	Premise	3,832	464
Argumentative relations	total relations	3,832	464
	Support	3,613	290
	Attack	219	174
	average depth	1.46	2.06

The MTC consists of 112 English-translated texts originally written by German speakers of various ages and educational levels (Peldszus and Stede, 2016). Owing to the small size of this dataset, we report an average score of five-fold cross-validation with 10 different splits, using the splits of the existing study (Peldszus and Stede, 2015).

Table 6.2 summarizes the statistics of the PEC and MTC. The PEC was approximately 10 times larger than the MTC. The "average depth" denotes the average of the depth of argument structure in each paragraph; the root object is not included in the depth calculation. For example, the depth of the structure in Figure 6.2 is two. Paragraphs that do not contain any argumentation relations, such as the introductory paragraph in the PEC, are excluded from the calculation. The MTC has relatively deeper structures than the PEC; this could be because instructions such as "include counterarguments" were given when the MTC data was collected.

Extracting argumentative markers The spans annotated in the PEC and MTC are different. In the PEC, AC spans are annotated, but AM spans are not. We defined the AM spans as the parts that are not AC, and each AC is paired with the immediately preceding AM. Note that the AM span contains punctuation marks at the end of the previous sentence (such as 'it.', 'it !' and 'it ?') and special symbols indicating the beginning of the text. For example, consider the AM in clause B of the following two sentences

(1) a. *A.* **Because** *B, C.*

b. *A* **because** *B. C.*

Table 6.3 Examples of argumentative markers

the other reason is that
consequently,
in conclusion, from the above views, although
first, as you can see that
in this essay , the reasons for why i agree that
it is a debatable subject that
unfortunately
although some argue that
furthermore , it 's undeniable that the
in short,
another thing that put big cities in front of small towns is
in conclusion, despite the contribution of it to the society,
however, some say that

The AM for clause B was ". *Because*" in example (1)-a, and "*because*" in example (1)-b. When there is no argumentative marker in ADU, only punctuation at the end of the previous sentence and the beginning of the sentence token are included in AM. In the PEC, 63% of ADUs have AMs with some argumentative markers (along with punctuations).

In the MTC, only ADUs were annotated; that is, there is no distinction between the AC and AM in the original annotation. To identify the AM in the ADU, we first created an AM list and employed rule-based matching. The AM list consists of the AM expressions collected from the PEC and discourse marker list in the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008). The created AM list contained 1,131 expressions (average 5.38 tokens), where 173 expressions were collected from the PDTB and 958 expressions were collected from the PEC. For each ADU, if an expression in the AM list existed at the beginning of the ADU, the longest matched phrase was considered to be the AM of that ADU. AMs were assigned to approximately 48% of the ADUs in the MTC. To estimate the performance of this rule-based approach, we manually annotated AM spans for randomly sampled 100 ADUs and evaluated the performance of rule-based AM extraction (Table 6.4). The scores were calculated according to an exact match.[3] We found that typical AMs such as "but" were generally extracted, but longer expressions such as "This would mean that" were not extracted well. Notably, the performance on the MTC was improved using AMs identified with this approach, which suggests that our approach could be effective.

Word representations We used five variants of word representations for our experiments: GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018), BERT (Devlin et al., 2019),

[3]The evaluation codeshttps://github.com/davidsbatista/NER-Evaluation are used.

Table 6.4 The performance of rule-based AM extraction in the MTC

Precision	Recall	F1 score
79.2	74.5	76.8

RoBERTa (Liu et al., 2019), and XLNET (Yang et al., 2019). In experiments using ELMo, we used the average of the intermediate representations of the three layers in pre-trained LSTM language models. In experiments using BERT, RoBERTa, and XLNET, a weighted sum of intermediate representations from all the layers was used. These weights were tuned during training. When a single token is divided into several subwords in pre-trained models' tokenizer, the averaged representation of constituent subwords was used as a corresponding word representation.

Baseline A model with span representation used in an existing study (Potash et al., 2017) was evaluated as a baseline. This span representation is computed from various features as explained in Section 6.3.3. We call this model bag-of-words (**BoW**). In the BoW baselines, we used the same hyperparameter listed in Table 6.1, except that $\boldsymbol{\phi}(w_{i:j})$ has 1536 dimensions in the BoW model for a fair comparison. An increase in model size does not affect their performance.

6.4.2 Results

Following existing studies, F1 scores for each subtask and arithmetic mean of the F1 scores for three subtasks are reported. A bootstrap hypothesis test (Koehn, 2004) was conducted for Macro F1 scores (Dror et al., 2018).

Table 6.5 shows the results for the PEC, and Table 6.6 shows the results for the MTC. We found that LSTM outperforms BoW, which indicates that integrating the LSTM-minus into argumentation structure parsing is effective. Comparing the performance between LSTM and LSTM+dist, the proposed inductive bias distinguishing the formal and content schema provides better argumentation structure parsing. Furthermore, these trends were almost consistent across various settings (different word representations and datasets).

The performance gain of LSTM+dist was relatively clearer in the MTC than in the PEC (Table 6.6). Based on the statistics that texts in the MTC have deeper structures than in the PEC, the gain of the LSTM+dist model could exist in parsing complex (deep) structures. In addition, the PEC consists of essays on a wide variety of topics, whereas the MTC contains several arguments on the same topic. Explicitly capturing the in-domain, content-level flows could be useful for parsing the texts in a similar topic with training instances (similar to the

Table 6.5 Performance of the LSTM+dist, LSTM, BoW models on the PEC. MC refers to MajorClaim class. The † mark on the results of LSTM model indicates statistically significant difference of performance compared to BoW model ($p < 0.05$). The ‡ mark on the reults of LSTM+dist indicates statistically significant difference of performance compared to LSTM model ($p < 0.05$).

Word rep.	Model	Overall	RI			RTC			ATC			
		Avg.	Macro	Link	No-Link	Macro	Support	Attack	Macro	MC	Claim	Premise
XLNET	LSTM+dist	**83.4**	**82.2**‡	70.3	94.2	**78.5**	96.2	60.7	**89.4**	95.4	79.8	93.1
	LSTM	82.5	80.6†	67.6	93.6	77.8†	96.6	59.1	89.1†	94.5	79.5	93.2
	BoW	77.4	76.4	60.4	92.3	71.6	95.7	47.6	84.3	91.4	71.0	90.4
RoBERTa	LSTM+dist	**82.9**	80.9	68.2	93.8	**79.4**	96.9	61.9	88.4	94.7	77.7	92.8
	LSTM	**82.9**	**81.6**†	69.3	94.0	77.9†	96.5	59.3	**89.1**†	94.3	79.6	93.3
	BoW	75.6	73.1	54.9	91.2	71.9	95.4	48.4	81.9	89.0	67.3	89.3
BERT	LSTM+dist	**81.8**	**80.9**	68.1	93.8	**78.0**	96.5	59.6	86.4	92.5	74.9	92.0
	LSTM	80.6	80.4†	67.3	93.5	74.9†	96.0	53.8	**86.6**†	92.2	75.4	92.2
	BoW	73.9	71.8	52.8	90.8	69.4	95.6	43.1	79.9	87.8	63.8	88.1
ELMo	LSTM+dist	**81.8**	**80.7**	67.8	93.7	**79.0**	96.8	61.1	85.7	91.6	73.3	92.1
	LSTM	**81.8**	80.4†	67.2	93.6	78.2†	96.7	59.8	**86.9**†	92.4	76.4	92.0
	BoW	77.1	76.2	60.2	92.2	72.3	96.2	48.3	82.9	90.4	68.6	89.6
GloVe	LSTM+dist	**79.7**	**78.8**	64.6	93.0	**76.5**	96.5	56.6	**83.9**	91.2	72.1	88.4
	LSTM	78.8	77.7†	62.7	92.6	75.0†	96.2	53.8	83.7	91.3	71.5	88.4
	BoW	75.4	73.8	56.0	91.5	69.8	95.9	43.6	82.8	89.5	69.1	89.7

MTC setting). Note that considering the application of argument structure analysis to the automatic essay assessment, the situation of the MTC, in which there are multiple essays on a specific topic, is plausible. In Section 6.5, we further examine the gain of the LSTM+dist.

Tables 6.7 and 6.8 show the comparison with the results of existing studies. Our models combining LSTM-minus span representation and the inductive bias of schema distinction update the performance by over 10 points from the existing studies.

6.5 Analysis

This analysis focuses on the task of identifying and classifying argumentative relations, which are the main subtasks of argumentation structure parsing. We used the PEC, which has a relatively large size, for analysis.

6.5.1 Depth in argumentation structure

Existing study (Stab and Gurevych, 2017) pointed out that their parser tends to predict shallow, near-linear trees compared to the gold annotations. We thus analyzed whether our model overcomes such a preference for linear structure. Figure 6.4 shows the accuracy of identifying the root nodes (depth 0), edges from the ADU at depth one, and edges from the ADUs

Table 6.6 Performance of the LSTM+dist, LSTM, BoW models on the MTC. MC refers to MajorClaim class. The † mark on the results of LSTM model indicates statistically significant difference of performance compared to BoW model ($p < 0.05$). The ‡ mark on the reults of LSTM+dist indicates statistically significant difference of performance compared to LSTM model ($p < 0.05$).

Word rep.	Model	Overall	RI			RTC			ATC		
		Avg.	Macro	Link	No-Link	Macro	Support	Attack	Macro	Claim	Premise
XLNET	LSTM+dist	**79.3**	**75.8**	60.7	90.8	**73.7**	82.1	65.3	**88.3**	81.1	95.6
	LSTM	76.0	73.6	57.2	89.9	69.3	77.1	61.6	85.2	75.8	94.6
	BoW	76.3	73.3	56.8	89.8	71.1	79.1	63.1	84.6	74.6	94.6
RoBERTa	LSTM+dist	**77.1**	**74.0**	57.8	90.1	**71.6**	80.5	62.6	**85.9**	7.1	94.6
	LSTM	73.1	71.8	54.4	89.2	66.4	75.2	57.7	80.9	68.6	93.3
	BoW	73.4	71.2	53.4	89.0	66.7	76.2	57.2	82.2	70.8	93.8
BERT	LSTM+dist	**76.1**	**73.0**	56.2	89.8	**71.1**‡	79.7	62.6	**84.0**‡	74.2	93.9
	LSTM	70.7	69.9	51.4	88.4	64.2	73.5	54.9	77.9	63.2	92.6
	BoW	71.9	70.0	51.5	88.5	65.2	75.6	54.9	80.4	67.6	93.2
ELMo	LSTM+dist	**78.2**	**73.9**	57.5	90.3	**77.2**‡	84.2	70.3	**83.4**	72.9	94.0
	LSTM	75.0	73.2†	56.3	90.0	71.3	78.7	64.0	80.5	68.1	93.0
	BoW	73.3	71.2	53.1	89.4	67.5	76.7	58.5	81.2	69.0	93.5
GloVe	LSTM+dist	**76.5**	**72.6**	55.4	89.8	**75.4**‡	82.3	68.4	**81.5**	69.9	93.1
	LSTM	70.4	70.1	51.3	88.8	64.1	71.6	56.7	76.9	61.7	92.2
	BoW	71.1	69.2	49.9	88.5	64.8	75.9	53.6	79.3	65.9	92.8

at depth two or deeper. Each score indicates the accuracy of identifying the correct parent in the set of nodes at each depth in the gold annotation. As pointed out in the existing study, identifying the edges in a deeper position of the gold structure is relatively difficult. For the performance in the depth 2+ setting, whereas the performance of the existing model (Potash et al., 2017) degrades, our LSTM+dist could robustly process.

The macro-structure of the "attack chain," where a potential opponent idea is introduced and the author of the argument counter-attacks it, is a representative case of a deep argumentation structure (Freeman, 2011; Peldszus and Stede, 2013). Figure 6.5 shows an example of an argumentation structure with the "attack chain" (ADU4→ADU3→ADU1) and the predicted structure for this text. As mentioned earlier, the existing model fails to predict such macro-structures, but our LSTM+dist could predict. Figure 6.9 shows the accuracy of identifying the edges in the attack-chain structure. The better performance of LSTM+dist suggests that the effectiveness of the distinction between formal/content schema for capturing such a typical substructure of argumentative texts. In an argumentative text where different stances are presented, AMs could be used frequently to make it clear which part is the author's claim or the opposing opinion. Therefore, explicit modeling of the AM flows could be the key to improving performance. In addition, such substructures are frequently observed especially in the MTC, which could explain the clear performance difference between the LSTM and LSTM+dist models in MTC (Figure 6.5 and 6.6).

Table 6.7 Comparison between our LSTM+dist model and models in existing studies on the PEC. MC refers to MAJORCLAIM class. The results of Potash et al. (2017) are obtained from our re-implemented version.

Model	Overall	RI			RTC			ATC			
	Avg.	Macro	Link	No-Link	Macro	Support	Attack	Macro	MC	Claim	Premise
LSTM+dist (XLNET)	83.4	82.2	70.3	94.2	78.5	96.2	60.7	89.4	95.4	79.8	93.1
BoW (XLNET)	77.4	76.4	60.4	92.3	71.6	95.7	47.6	84.3	91.4	71.0	90.4
LSTM+dist (GloVe)	79.7	78.8	64.6	93.0	76.5	96.5	56.6	83.9	91.2	72.1	88.4
Potash+ 2017	-	76.7	60.8	92.5	-	-	-	84.9	89.4	73.2	92.1
Niculae+ 2017	-	-	60.1	-	-	-	-	77.6	78.2	64.5	90.2
Stab+ 2017	75.2	75.1	58.5	91.8	68.0	94.7	41.3	82.6	89.1	68.2	90.3

Table 6.8 Comparison between our LSTM+dist model and models in existing studies on the MTC. MC refers to MAJORCLAIM class. The results of Potash et al. (2017) are obtained from our re-implemented version.

Model	Overall	RI			RTC			ATC		
	Avg.	Macro	Link	No-Link	Macro	Support	Attack	Macro	Claim	Premise
LSTM+dist (XLNET)	79.3	75.8	60.7	90.8	73.7	82.1	65.3	88.3	81.1	95.6
BoW (XLNET)	76.3	73.3	56.8	89.8	71.1	79.1	63.1	84.6	74.6	94.6
LSTM+dist (GloVe)	76.5	72.6	55.4	89.8	75.4	82.3	68.4	81.5	69.9	93.1
Potash+ 2017	-	74.0	57.7	90.3	-	-	-	81.3	69.2	93.4
Afantenos+ 2018	78.5	72.2	-	-	75.7	-	-	87.6	-	-
Stab+ 2017	76.2	68.3	48.6	88.1	74.5	85.5	62.8	85.7	77.0	94.3

6.6 Conclusion

This study has investigated an effective inductive bias (architecture design) to achieve better discourse processing. Our experiments have demonstrated the effectiveness of the proposed inductive bias that makes models distinguish between formal- and content-level flows explicitly. Such bias is closely related to the classical theory of human discourse processing, schema theory. Our analysis further has revealed that the model with the proposed bias successfully generalizes to parsing complex discourse structures.

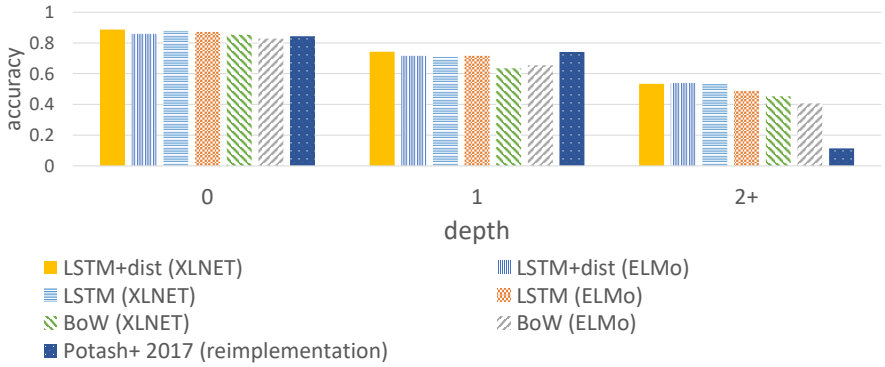


Fig. 6.4 The accuracy of the argumentation relation identification subtask by depth in argumentation structure.

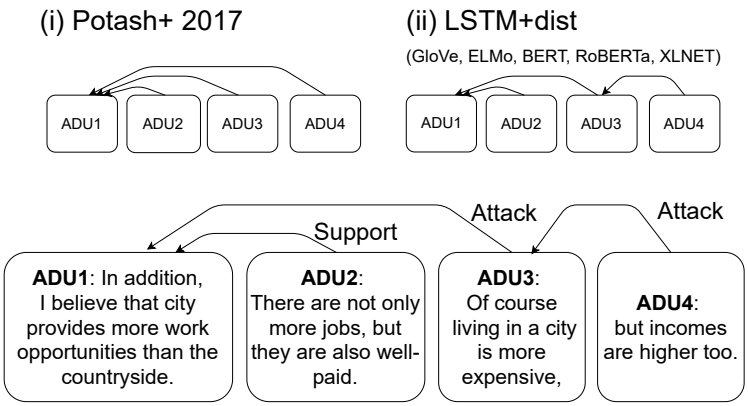


Fig. 6.5 An example of predicted structure from existing model (Potash et al., 2017) and our LSTM+dist model.

Table 6.9 Performance of predicting sub-structure where ATTACK relation chains.

Word rep.	Model	Acc.
XLNET	LSTM+dist	**71.6**
	LSTM	67.8
	BoW	60.7
RoBERTa	LSTM+dist	**69.9**
	LSTM	67.2
	BoW	56.8
BERT	LSTM+dist	68.8
	LSTM	**69.3**
	BoW	57.4
ELMo	LSTM+dist	**71.0**
	LSTM	68.9
	BoW	60.7
GloVe	LSTM+dist	**63.9**
	LSTM	59.0
	BoW	57.9
	Potash+ 2017	50.3

Chapter 7

Conclusions

Through expanding the scope of cognitive plausibility studies of neural NLP models, we found the following:

1. **Acceptability judgment by LMs in flexible word-order language:** Our experiments demonstrated that the neural LMs could successfully simulate human-like word order preferences in flexible word order language. Our analyses associated with linguistic studies suggest that the parallel of word order preference between humans and LMs was consistent from various linguistic factors. (e.g., verb type, animacy).

2. **Cross-linguistic analysis of incremental sentence processing by neural LMs:** While neural LMs exhibited human-like sentence-level preferences, our analysis focusing on token-by-token processing difficulties revealed language-dependent discrepancies between human and LM sentence processing. We compared reading time by human and surprisal by LMs, and observed that accurate LMs deviated from human reading behavior in the Japanese data.

3. **Integrating expectation and memory in gaze duration modeling:** We suggest that memory-based account of sentence processing could be one source of the discrepancies. We empirically demonstrated that context-limited LMs could partially compensate for the difference between sentence processing by humans and modern LMs, and also highlighted the need of syntactic bias in explaining the memory limitation. This also implies that the human sentence processing model is much more lightweight like n-gram LMs than the modern neural LM architectures (e.g., self-attention) assume.

4. **Inductive bias for discourse processing:** We tackled the challenging issues in the NLP field—discourse processing. We demonstrated that the cognitively-motivated

inductive bias contributes to improving discourse processing model. These models successfully capture the hierarchical structure of argumentative texts while existing models tend to predict shallow, near-linear trees.

Again, disclosing the underlying mechanism of human information processing is a fundamental goal in the cognitive science field including artificial intelligence, linguistics, and NLP. With this in mind, we believe that contrasting language processing by humans and NLP models is a necessary step to cross-fertilize the fields and reverse-engineer human language processing. In particular, NLP technology has made tremendous progress, and models that behave like humans at first glance are developed. In considering the implications of these models, we believe that trying to understand the nature of human and language is not only the job of psychologists or linguists, but also the job of NLP researchers. In the recent trends of NLP, engineering-oriented research has received a great deal of attention, and exploring the connections to research on humans are relatively limited and investigated within specific communities. We hope that this thesis encourages more studies bridging NLP and cognitive science of language.

Appendix A

Hyperparameters of models

A.1 Hyperparameters of LMs

A.1.1 LMs in Section 3.3.2 and 3.4

We used the Transformer (Vaswani et al., 2017b) LMs implemented in fairseq (Ott et al., 2019). Table A.1 shows the hyperparameters of the LMs.

Table A.1 Hyperparameters of the LMs.

Fairseq model	architecture	transformer_lm
	adaptive softmax cut off	50,000, 140,000
Optimizer	algorithm	Nesterov accelerated gradient (nag)
	learning rates	1e-5
	momentum	0.99
	weight decay	0
	clip norm	0.1
Learning rate scheduler	type	cosine
	warmup updates	16,000
	warmup init lrarning rate	1e-7
	max learning rate	0.1
	min learning rate	1e-9
	t mult (factor to grow the length of each period)	2
	learning rate period updates	270,000
	learning rate shrink	0.75
Training	batch size	4608 tokens
	epochs	3

A.1.2 LMs in Section 3.3.3, Chapter 4, and Chapter 5

We used the Transformer (Vaswani et al., 2017b) and LSTM (Hochreiter and Schmidhuber, 1997) LMs implemented in fairseq (Ott et al., 2019). Table A.2 shows the hyperparameters of the LMs.

Table A.2 Hyperparameters for LMs. The same optimizer and learning rate scheduler are used for TRANS-SM and LSTM LMs.

TRANS-LG	Fairseq model	architecture	transformer_lm_gpt2_small
		adaptive softmax cut off	50,000, 140,000
		share-decoder-input-output-embed	True
		embed_dim	1,024
		ffn_embed_dim	4,096
		layers	24
		heads	16
		dropout	0.1
		attention_dropout	0.1
	Optimizer	algorithm	AdamW
		learning rates	5e-4
		betas	(0.9, 0.98)
		weight decay	0.01
		clip norm	0.0
	Learning rate scheduler	type	inverse_sqrt
		warmup updates	4,000
		warmup init lrarning rate	1e-7
	Training	batch size	61,440 tokens
		sample-break-mode	none
TRANS-SM	Fairseq model	architecture	transformer_lm_gpt
		adaptive softmax cut off	50,000, 140,000
		share-decoder-input-output-embed	True
		embed_dim	384
		ffn_embed_dim	2,048
		layers	8
		heads	6
		dropout	0.1
		attention_dropout	0.1
	Training	batch size	61,440 tokens
		sample-break-mode	none
LSTM	Fairseq model	architecture	lstm_lm
		adaptive softmax cut off	50,000, 140,000
		share-decoder-input-output-embed	True
		embed_dim	400
		hiden_size	1,024
		layers	2
		dropout	0.1
	Training	batch size	20,480 tokens
		sample-break-mode	none

References

Afantenos, S., Peldszus, A., and Stede, M. (2018). Comparing decoding mechanisms for parsing argumentative structures. *Argument & Computation*, 9(3):177–192.

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of child language*, 26(2):339–356.

Alonso Belmonte, I. et al. (2000). Teaching English Word Order to ESL Spanish Students. A Functional Perspective. *Encuentro. Revista de Investigación e Innovación en la clase de idiomas*, 11:1999–2000.

An, S. (2013). Schema theory in reading. *Theory & Practice in Language Studies*, 3(1).

Ang, Z. (2014). The effects of discourse markers on the reading comprehension and speed of chinese learners of english. *International Journal Of English Language and Linguistics Studies*, 2(2):27–49.

Asahara, M. (2017). Between Reading Time and Information Structure. In *Proceedings of PACLIC*, pages 15–24.

Asahara, M. (2018). Between Reading Time and Clause Boundaries in Japanese - Wrap-up Effect in a Head-final Language. In *Proceedings of PACLIC*, pages 19–27.

Asahara, M. and Kato, S. (2017). Between Reading Time and Syntactic / Semantic Categories. In *Proceedings of IJCNLP*, pages 404–412.

Asahara, M., Ono, H., and Miyamoto, E. T. (2016). Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING*, pages 684–694.

Aurnhammer, C. and Frank, S. L. (2019). Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing. In *Proceedings of CogSci*, pages 112–118.

Bahlmann, J., Rodriguez-Fornells, A., Rotte, M., and Münte, T. F. (2007). An fMRI study of canonical and noncanonical word order in German. *Human brain mapping*, 28(10):940–949.

Baroni, M. (2021). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing.

Barrett, M., Agi, Z., and Søgaard, A. (2015). The Dundee Treebank. In *Fourteenth International Workshop on Treebanks and Linguistic Theories*, pages 242–248.

Barrett, M., Bingel, J., Keller, F., and Søgaard, A. (2016). Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of ACL*, pages 579–584. aclweb.org.

Barrett, M. and Hollenstein, N. (2020). Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Lang. Linguist. Compass*, 14(11):1–16.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48.

Belinkov, Y., Gehrmann, S., and Pavlick, E. (2020). Interpretability and analysis in neural NLP. In *Proceedings of ACL*, pages 1–5. aclweb.org.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *TACL*, 7:49–72.

Bender, E. M. (2011). On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Bloem, J. (2016). Testing the Processing Hypothesis of word order variation using a probabilistic language model. In *Proceedings of CL4LC*, pages 174–185, Osaka, Japan.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., and Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the dative alternation. In *Cognitive foundations of interpretation*, pages 69–94. KNAW.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T., editors, *Proceedings of NeurIPS*.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2020). Extracting training data from large language models. *CoRR*, abs/2012.07805.

Carrell, P. L. (1982). Cohesion is not coherence. *TESOL Quarterly*, 16(4):479–488.

Carrell, P. L. and Eisterhold, J. C. (1983). Schema theory and ESL reading pedagogy. *TESOL Q.*, 17(4):553.

Chaudron, C. and Richards, J. C. (1985). *The Effect of Discourse Markers on the Comprehension of Lectures*. ERIC.

Cheng, S.-M., Yu, C.-H., and Chen, H.-H. (2014). Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. In *Proceedings of COLING*, pages 279–289, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Chomsky, N. (1980). Rules and representations. *Behav. Brain Sci.*, 3(1):1–15.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.*, 36(3):181–204.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.*, 24(1):87–114; discussion 114–85.

Crocker, M. W. (2010). Computational psycholinguistics. *Computational Linguistics and Natural Language*.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Journal of Cognition*, 109(2):193–210.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186. Association for Computational Linguistics.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker' s guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, pages 1383–1392. Association for Computational Linguistics.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). Recurrent neural network grammars. In *Proceedings of NAACL*. Association for Computational Linguistics.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL*, pages 11–22. Association for Computational Linguistics.

Elman, J. L., Bates, E. A., and Johnson, M. H. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *TACL*, 8:34–48.

Ferreira, F. and Lowder, M. W. (2016). Chapter six - prediction, information structure, and Good-Enough language processing. In Ross, B. H., editor, *Psychology of Learning and Motivation*, volume 65, pages 217–247. Academic Press.

Fossum, V. and Levy, R. (2012). Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of CMCL*, pages 61–69, Montréal, Canada.

Frank, S. L. and Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological science*, 22(6):829–834.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Frank, S. L., Trompenaars, T., and Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cogn. Sci.*, 40(3):554–578.

Freeman, J. B. (2011). *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Journal of Cognitive Science*.

Futrell, R. and Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of EACL*, pages 688–698.

Futrell, R. and Levy, R. P. (2019). Do RNNs learn human-like abstract word order preferences? In *Proceedings of SCiL*, pages 50–59.

Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of ACL*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Journal of Cognition*, 68(1):1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of lingustic complexity. *Image, language, brain*, 2000:95–126.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of CMCL*, pages 10–18.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*, pages 159–166.

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. R. (2018). Finding Syntax in Human Encephalography with Beam Search. In *Proceedings of ACL*, pages 2727–2736.

Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in english*. Routledge.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.

He, L., Lee, K., Levy, O., and Zettlemoyer, L. (2018). Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of ACL*, pages 364–369.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Journal of Neural Computation*, 9(8):1735–1780.

Hoji, H. (1985). Logical form constraints and configurational structures in Japanese. *PHD Thesis. University of Washington*.

Hollenstein, N., de la Torre, A., Langer, N., and Zhang, C. (2019). CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of CoNLL*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Hollenstein, N. and Zhang, C. (2019). Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of NAACL*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Strong expectations cancel locality effects: evidence from hindi. *PLoS One*, 9(7):e100986.

Jaeger, T. and Levy, R. (2007). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *NIPS*, volume 19, pages 849–856. MIT Press.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Ke, Z., Carlile, W., Gurrapadi, N., and Ng, V. (2018). Learning to give feedback: modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of IJCAI*, pages 4130–4136.

Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Kharitonov, E. and Chaabouni, R. (2021). What they do when in doubt: a study of inductive biases in seq2seq learners. In *International Conference on Learning Representations*.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL*, pages 4110–4124, Online. Association for Computational Linguistics.

Kirov, C. and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *TACL*, 6:651–665.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of ACL*, pages 2676–2686.

Klerke, S. and Plank, B. (2019). At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61. aclweb.org.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Koizumi, M. and Tamaoka, K. (2004). Cognitive processing of Japanese sentences with ditransitive verbs. *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, 2004(125):173–190.

Koizumi, M. and Tamaoka, K. (2006). The Canonical Positions of Adjuncts in the Processing of Japanese Sentence. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 13(3):392–403.

Konieczny, L. (2000). Locality and Parsing Complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Kudo, T. (2006). MeCab: Yet Another Part-of-speech and Morphological Analyzer. *http://mecab. sourceforge. jp*.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of EMNLP*, pages 66–71.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., and Inui, K. (2021). Lower perplexity is not always human-like. In *Proceedings of ACL-IJCNLP*, pages 5203–5217, Online. Association for Computational Linguistics.

Kuribayashi, T., Reisert, P., Inoue, N., and Inui, K. (2017). Examining macro-level argumentative structure features for argumentative relation identification. *IPSJ SIG Technical Report*, 2017-NL-234:1–6.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cogn. Sci.*, 41(5):1202–1241.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45:765–818.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of EMNLP*, pages 188–197.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL*, pages 687–692.

Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity*. stanford university.

Levy, R. (2008). Expectation-based syntactic comprehension. *Journal of Cognition*, 106(3):1126–1177.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.*, 29(3):375–419.

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.*, 10(10):447–454.

Li, C. N. and Thompson, S. A. (1984). Subject and Topic:A new typology of language. *Contemporary Linguistics*, pages 457–489.

Li, Q., Li, T., and Chang, B. (2016). Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of EMNLP*, pages 362–371.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In *Natural language generation*, pages 85–95. Springer.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., USA.

Marvin, R. and Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. In *Proceedings of EMNLP*, pages 1192–1202.

Mathias, S., Kanojia, D., Mishra, A., and Bhattacharyya, P. (2020). A survey on using gaze behaviour for natural language processing. In *IJCAI*, pages 4907–4913. researchgate.net.

Matsuoka, M. (2003). Two Types of Ditransitive Consturctions in Japanese. *Journal of East Asian Linguistics*, 12(2):171–203.

Maurits, L., Navarro, D., and Perfors, A. (2010). Why are some word orders more common than others? a uniform information density account. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *NIPS*, volume 23. Curran Associates, Inc.

McCoy, R. T., Frank, R., and Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *cogsci.mindmodeling.org*.

McCoy, R. T., Frank, R., and Linzen, T. (2020). Does syntax need to grow on trees? sources of hierarchical inductive bias in Sequence-to-Sequence networks. *TACL*, 8:125–140.

McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there's no majority: Limitations of Encoder-Decoder neural networks as cognitive models for german plurals. In *Proceedings of ACL*, pages 1745–1756. aclweb.org.

Meister, C., Cotterell, R., and Vieira, T. (2020). If Beam Search Is the Answer, What Was the Question? In *Proceedings of EMNLP*, pages 2173–2185.

Merkx, D. and Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of CMCL*, pages 12–22, Online. Association for Computational Linguistics.

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.*, 63(2):81–97.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2016). Leveraging cognitive features for sentiment analysis. In *Proceedings of CoNLL*, pages 156–166, Berlin, Germany. Association for Computational Linguistics.

Misra, K., Ettinger, A., and Rayz, J. (2020). Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of EMNLP*, pages 4625–4635, Online. Association for Computational Linguistics.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.

Mitsuda, K., Iida, R., and Tokunaga, T. (2013). Detecting missing annotation disagreement using eye gaze information. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 19–26, Nagoya, Japan. Asian Federation of Natural Language Processing.

Nguyen, H. and Litman, D. (2016). Context-aware argumentative relation mining. In *Proceedings of ACL*, pages 1127–1137.

Nguyen, H. V. and Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of AAAI*, pages 5892–5899.

Niculae, V., Park, J., and Cardie, C. (2017). Argument Mining with Structured SVMs and RNNs. In *Proceedings of ACL*, pages 985–995.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Ouchi, H., Shindo, H., and Matsumoto, Y. (2018). A span selection model for semantic role labeling. In *Proceedings of EMNLP*, pages 1630–1642.

Peldszus, A. and Stede, M. (2013). From Argument Diagrams to Argumentation Mining in Texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Peldszus, A. and Stede, M. (2015). Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of EMNLP*, pages 938–948.

Peldszus, A. and Stede, M. (2016). An Annotated Corpus of Argumentative Microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, pages 801–815.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the NAACL-HLT 2018*, pages 2227–2237.

Pinker, S. and Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Potash, P., Romanov, A., and Rumshisky, A. (2017). Here's My Point: Joint Pointer Architecture for Argument Mining. In *Proceedings of EMNLP*, pages 1375–1384.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *Proceedings of LREC*, pages 2961–2968.

Radrof, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018). Language Models are Unsupervised Multitask Learners. *Technical report, OpenAI*.

Rayner, K. and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4):504–509.

Reed, C. (2006). Preliminary results from an argument corpus. In *In Eloína Miyares Bermúdez & Leonel Ruiz Miyares (Eds), Linguistics in the twenty-first century*. Citeseer.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of EMNLP*, pages 324–333, Singapore.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *TACL*, 8:842–866.

Rumelhart, D. and McClelland, J. (1986). *On learning the past tenses of English verbs*. ERIC.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. *Theoretical Issues in Reading Comprehension*.

Rumelhart, D. E., McClelland, J. L., Group, P. R., and Others (1988). *Parallel distributed processing*, volume 1. IEEE Massachusetts.

Saeki, T. (1960). Gendaigo ni okeru gojun no keikō – iwayuru hogo no baai [The trend of word order in modern writing– in so-called complements]. *Gengo seikatsu [Language life]*, 111:56–63.

Saeki, T. (1998). *Yōsetsu Nihongo no Gojun [Essentials of Japanese word order]*. Kurosio Publishers.

Sasano, R. and Okumura, M. (2016). A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions. In *Proceedings ACL*, pages 2236–2244, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, pages 1715–1725.

Shapiro, S. C. (2003). Artificial intelligence (AI). In *Encyclopedia of Computer Science*, pages 89–93. John Wiley and Sons Ltd., GBR.

Shigenaga, Y. (2014). Canonical Word Order of Japanese Ditransitive Sentences: A Preliminary Investigation through a Grammaticality Judgment Survey. *Advances in Language and Literary Studies*, 5(2):35–45.

Slobin, D. I. and Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12(3):229–265.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Journal of Cognition*, 128(3):302–319.

Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. In *Proceedings of CL*, volume 43, pages 619–659. MIT Press.

Stede, M., Afantenos, S., Peldszus, A., Asher, N., and Perret, J. (2016). Parallel discourse annotations on a corpus of short texts. In *Proceedings of LREC*, pages 1051–1058.

Stern, M., Andreas, J., and Klein, D. (2017). A minimal span-based neural constituency parser. In *Proceedings of ACL*, pages 818–827.

Sugawara, S., Stenetorp, P., and Aizawa, A. (2021). Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge Univ. Press (Cambridge).

Tsujimura, N. (2013). *An introduction to Japanese linguistics*. John Wiley & Sons.

Upadhye, S., Bergen, L., and Kehler, A. (2020). Predicting Reference: What do Language Models Learn about Discourse Models? In *Proceedings of EMNLP*, pages 977–982, Online. Association for Computational Linguistics.

van Schijndel, M. and Linzen, T. (2021). Single-Stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cogn. Sci.*, 45(6):e12988.

Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Lang. Cogn. Process.*, 25(4):533–567.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017a). Attention Is All You Need. In *Proceedings of NIPS*, pages 5998–6008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017b). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *NIPS*, pages 5998–6008. Curran Associates, Inc.

Vickers, P., Wainwright, R., Madabushi, H. T., and Villavicencio, A. (2021). CogNLP-Sheffield at CMCL 2021 shared task: Blending cognitively inspired features with transformer-based language models for predicting eye tracking patterns. In *Proceedings of CMCL*, pages 125–133. aclweb.org.

Visweswariah, K., Rajkumar, R., Gandhe, A., Ramanathan, A., and Navratil, J. (2011). A Word Reordering Model for Improved Machine Translation. In *Proceedings of EMNLP*, pages 486–496, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Wang, W. and Chang, B. (2016). Graph-based dependency parsing with bidirectional lstm. In *Proceedings of ACL*, pages 2306–2315.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. (2020a). BLiMP: The benchmark of linguistic minimal pairs for english. *TACL*, 8:377–392.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *TACL*, 7:625–641.

Warstadt, A., Zhang, Y., Li, X., Liu, H., and Bowman, S. R. (2020b). Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of EMNLP*, pages 217–235, Online. Association for Computational Linguistics.

Wei, J., Garrette, D., Linzen, T., and Pavlick, E. (2021a). Frequency effects on syntactic rule learning in transformers. In *Proceedings of EMNLP*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei, J., Meister, C., and Cotterell, R. (2021b). A cognitive regularizer for language modeling. In *Proceedings of ACL*, pages 5191–5202, Online. Association for Computational Linguistics.

Wilcox, E., Vani, P., and Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of ACL*, pages 939–952, Online. Association for Computational Linguistics.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of CogSci*, pages 1707–1713.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NIPS*, pages 5754–5764.

Yoshida, R., Noji, H., and Oseki, Y. (2021). Modeling human sentence processing with left-corner recurrent neural network grammars. In *Proceedings of EMNLP*, pages 2964–2973, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

List of publications

Journal papers (refereed)

1. Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Masashi Yoshikawa, Kentaro Inui. Instance-Based Neural Dependency Parsing. Transactions of the Association for Computational Linguistics 2021, Volume 9, pp.1493-1507, September 2021.

2. 栗林樹生, 大内啓樹, 井之上直也, 鈴木潤, Paul Reisert, 三好利昇, 乾健太郎. 論述構造解析におけるスパン分散表現. 自然言語処理, Volume 27, Number 4, pp.753-780, December 2020. 最優秀論文賞受賞.

3. Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, Kentaro Inui. Assisting Authors to Convert Raw Products into Polished Prose. Journal of Cognitive Science, Vol.21, No.1, pp.99-135, 2020

International conference/workshop papers (refereed)

1. Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), pp4547-4568, 2021/11.

2. Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara and Kentaro Inui. Lower Perplexity is Not Always Human-Like. In proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), pp. 5203-5217, 2021/08.

3. Takaki Otake, Sho Yokoi, Naoya Inoue, Ryo Takahashi, Tatsuki Kuribayashi and Kentaro Inui. Modeling Event Salience in a Narrative Based on Barthes' Cardinal Function. In proceedings of the 28th International Conference on Computational Linguistics (COLING-2020), pp. 1784-1794, 2020/12.

4. Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi and Kentaro Inui. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP-2020), pp. 7057-7075, 2020/11.

5. *Takumi Ito, *Tatsuki Kuribayashi, *Masatoshi Hidaka, Jun Suzuki and Kentaro Inui. (* equal contribution). Langsmith: An Interactive Academic Text Revision System. In proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP-2020, system demonstration track), pp. 216-226, 2020/11.

6. Tatsuki Kuribayashi, Takumi Ito, Jun Suzuki and Kentaro Inui. Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese. In proceedings of the 58th annual meeting of the Association for Computational Linguistics (ACL-2020), pp. 6452-6459, 2020/07.

7. Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno and Kentaro Inui. Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition. In proceedings of the 58th annual meeting of the Association for Computational Linguistics (ACL-2020), pp. 6452-6459, 2020/07.

8. *Takumi Ito, *Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki and Kentaro Inui. (* equal contribution) Diamonds in the Rough: Generating Fluent Sentences from Early-stage Drafts for Academic Writing Assistance. In Proceedings of the 12th International Conference on Natural Language Generation (INLG-2019), pp. 40-53, 2019/10.

9. Masato Hagiwara, Takumi Ito, Tatsuki Kuribayashi, Jun Suzuki, and Kentaro Inui. TEASPN: Framework and Protocol for Integrated Writing Assistance Environments. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019) (accepted to system demonstrations track), pp. 229-234, 2019/11.

10. Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. An Empirical Study of Span Representations in Ar-

gumentation Structure Parsing. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL-2019), pp. 4691-4698, 2019/07.

11. Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates. In Proceedings of the 5th Workshop on Argument Mining, pp.79-89, 2018/11.

Non-archival articles (refereed)

1. Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe and Kentaro Inui. Topicalization in Language Models: A Case Study on Japanese. In proceedings of Student Research Workshop at the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2021 SRW Non-archival), 2021/08.

2. Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Self-Attention is Not Only a Weight: Analyzing BERT with Vector Norms. In proceedings of Student Research Workshop at the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020 SRW Non-archival), 2020/07.

Other publications (non-refereed)

Commentary articles

1. 栗林樹生.「論述構造解析におけるスパン分散表現」の研究を通して. 自然言語処理, Volume 28, Number 2, pp.677-681, 2021/06

2. 栗林樹生. Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese. 自然言語処理, Volume 27, Number 3, pp.671-676, 2020/09.

Domestic conference

1. 石月由紀子, 栗林樹生, 松林優一郎, 大関洋平. 情報量に基づく日本語項省略の分析.NLP 若手の会第 16 回シンポジウム (YANS2021), 2021/08.

2. 小林悟郎, 栗林樹生, 横井祥，乾健太郎. 非線形モジュールの加法分解に基づく Transformer レイヤーの解析. NLP 若手の会第 16 回シンポジウム (YANS 2021), 2021/08.

3. 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼，浅原正幸, 乾健太郎. 予測の正確な言語モデルがヒトらしいとは限らない. 言語処理学会第 27 回年次大会 (NLP 2021), pp.267-272, 2021/03. 委員特別賞受賞.

4. 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼，浅原正幸, 乾健太郎. 日本語の読みやすさに対する情報量に基づいた統一的な解釈. 言語処理学会第 27 回年次大会 (NLP 2021), pp.723-728, 2021/03.

5. 伊藤拓海, 栗林樹生, 日高雅俊, 鈴木潤, 乾健太郎. Langsmith: 人とシステムの協働による論文執筆. 言語処理学会第 27 回年次大会 (NLP 2021), pp.1834-1839, 2021/03.

6. 藤原吏生, 栗林樹生, 乾健太郎. 人と言語モデルが捉える文の主題. 言語処理学会第 27 回年次大会 (NLP 2021), pp.1307-1312, 2021/03.

7. 小林悟郎, 栗林樹生, 横井祥, 乾健太郎. Transformer の文脈を混ぜる作用と混ぜない作用. 言語処理学会第 27 回年次大会 (NLP 2021), pp.1224-1229, 2021/03.

8. 大内啓樹, 鈴木潤, 小林颯介, 横井祥, 栗林樹生, 吉川将司, 乾健太郎. 事例ベース依存構造解析のための依存関係表現学習. 言語処理学会第 27 回年次大会 (NLP 2021), pp.497-502, 2021/03.

9. 大竹孝樹, 横井祥, 井之上直也, 高橋諒, 栗林樹生, 乾健太郎. 物語におけるイベントの顕現性推定と物語類似性計算への応用. 言語処理学会第 27 回年次大会 (NLP 2021), pp.1324-1329, 2021/03.

10. 藤原吏生, 栗林樹生, 乾健太郎. 日本語言語モデルが選択する文形式の傾向と文脈の影響 —— 主題化・有標語順について. NLP 若手の会第 15 回シンポジウム (YANS 2020), 2020/09.

11. 小林悟郎, 栗林樹生, 横井祥, 乾健太郎. ベクトル長に基づく注意機構と残差結合の包括的な分析. NLP 若手の会第 15 回シンポジウム (YANS 2020), 2020/09.

12. 栗林樹生, 伊藤拓海, 鈴木潤, 乾健太郎. 日本語語順分析に言語モデルを用いることの妥当性について. 言語処理学会第 26 回年次大会 (NLP 2020), pp.493-496, 2020/03.

13. 小林悟郎, 栗林樹生, 横井祥, 鈴木潤, 乾健太郎. ベクトル長に基づく自己注意機構の解析. 言語処理学会第 26 回年次大会 (NLP 2020), pp.965-968, 2020/03. 最優秀賞受賞.

14. 大内啓樹, 鈴木潤, 小林颯介, 横井祥, 栗林樹生, 乾健太郎. スパン間の類似性に基づく事例ベース構造予測. 言語処理学会第 26 回年次大会 (NLP 2020), pp.331-334, 2020/03.

15. 大竹孝樹, 横井祥, 井之上直也, 高橋諒, 栗林樹生, 乾健太郎. 言語モデルによる物語中のイベントの顕現性推定. 言語処理学会第 26 回年次大会 (NLP 2020), pp.1089-1092, 2020/03.

16. 伊藤拓海, 栗林樹生, 萩原正人, 鈴木潤, 乾健太郎. 英語論文執筆のための統合ライティング支援環境. 第 14 回 NLP 若手の会シンポジウム (YANS 2019), 2019/08.

17. 小林悟郎, 栗林樹生, 横井祥, 鈴木潤, 乾健太郎. 文脈を考慮する言語モデルが捉える品詞情報とその軌跡. 第 14 回 NLP 若手の会シンポジウム (YANS 2019), 2019/08.

18. 栗林樹生, 大内啓樹, 井之直也, Paul Reisert, 三好利昇, 鈴木潤, 乾健太郎. 複数の言語単位に対するスパン表現を用いた論述構造解析. 言語処理学会第 25 回年次大会 (NLP 2019), pp.990-993, 2019/03.

19. 栗林樹生, 伊藤拓海, 内山香, 鈴木潤, 乾健太郎. 言語モデルを用いた日本語の語順評価と基本語順の分析. 言語処理学会第 25 回年次大会 (NLP 2019), pp.1053-1056, 2019/03. 若手奨励賞受賞.

20. 伊藤拓海, 栗林樹生, 小林隼人, 鈴木潤, 乾健太郎. ライティング支援を想定した情報補完型生成. 言語処理学会第 25 回年次大会 (NLP 2019), pp.970-973, 2019/03.

21. Tatsuki Kuribayashi, Paul Reisert, Naoya Inoue and Kentaro Inui. Towards Exploiting Argumentative Context for Argumentative Relation Identification. 言語処理学会第 24 回年次大会 (NLP 2018), pp.284-287, 2018/03.

22. Tatsuki Kuribayashi, Paul Reisert, Naoya Inoue, Kentaro Inui. Examining Macro-level Argumentative Structure Features for Argumentative Relation Identification. 第 4 回自然言語処理シンポジウム・第 234 回自然言語処理研究会, 6pages, 2017/12.