Predicting Word Fixations in Text with a CRF Model for Capturing General Reading Strategies among Readers

Tadayoshi Hara[1] *Daichi Mochihashi*[2] *Yoshinobu Kano*[1,3] *Akiko Aizawa*[1]
(1) National Institute of Informatics, Japan
(2) The Institute of Statistical Mathematics, Japan
(3) PRESTO, Japan Science and Technology Agency
harasan@nii.ac.jp, daichi@ism.ac.jp, {kano, aizawa}@nii.ac.jp

Abstract

Human gaze behavior while reading text reflects a variety of strategies for precise and efficient reading. Nevertheless, the possibility of extracting and importing these strategies from gaze data into natural language processing technologies has not been explored to any extent. In this research, as a first step in this investigation, we examine the possibility of extracting reading strategies through the observation of word-based fixation behavior. Using existing gaze data, we train conditional random field models to predict whether each word is fixated by subjects. The experimental results show that, using both lexical and screen position cues, the model has a prediction accuracy of between 73% and 84% for each subject. Moreover, when focusing on the distribution of fixation/skip behavior of subjects on each word, the total similarity between the predicted and observed distributions is 0.9462, which strongly supports the possibility of capturing general reading strategies from gaze data.

Title and Abstract in Japanese

人の一般的な文章理解戦略を捉えるための CRF モデルを用いた文章中の単語注視予測

人間が文章を読む際の視線行動には、正確かつ効率的に読むための様々な戦略が反映されている。しかしながら、その戦略を視線データから抽出し、自然言語処理技術に取り入れるという可能性に関しては、これまでほとんど研究されて来なかった。本研究では、この可能性を研究するための第一歩として、単語ベースの注視行動の観察を通して文章理解戦略の抽出可能性を調査する。我々は既存の視線データを用い、各単語が被験者によって注視されるかどうかを予測する条件付き確率場モデルを訓練する。実験では、語彙情報と画面位置情報を手がかりにすることで、このモデルが各被験者に対して 73%から 84%の予測精度を与えることが示される。さらに、各単語に対する被験者間の注視／スキップの分布に着目すると、予測された分布と実際に観察された分布との全体的な近似度は 0.9462 であることが示され、視線データから一般的な文章理解戦略を捉えうる可能性を強く裏付ける実験結果となっている。

Keywords: eye-tracking, gaze data, reading behavior, conditional random field (CRF).

Keywords in Japanese: 視線追跡、視線データ、読解行動、条件付き確率場 (CRF).

Proceedings of the First Workshop on Eye-tracking and Natural Language Processing, pages 55–70,
COLING 2012, Mumbai, December 2012.

1 Introduction

Natural language processing (NLP) technologies have long been explored and some have approached close to satisfactory performance. Nevertheless, even for such sophisticated technologies, there are still various issues pending further improvement. For example, in parsing technologies, over 90% parsing accuracy has been achieved, yet some coordination structures or modifier dependencies are still analyzed incorrectly.

Humans, on the other hand, can deal with such issues relatively effectively. We expect that if we could clarify the mechanism used by humans, the performance of NLP technologies could be improved by incorporating such mechanisms in their systems. To clarify these mechanisms, analyzing human reading behavior is essential, while gaze data should strongly reflect this behavior. When a human reads a piece of text, especially for the first time, it is important that his/her eye movements are optimized for rapid understanding of the text. Humans typically perform this optimization unconsciously, which is reflected in the gaze data.

Eye movements while reading text have long been explored in the field of psycholinguistics (Rayner, 1998), and the accumulated knowledge of human eye movements has been reflected in various eye movement models (Reichle et al., 1998, 2003, 2006). Reinterpretation of the knowledge from an NLP perspective, however, has not been thoroughly investigated (Nilsson and Nivre, 2009, 2010; Martínez-Gómez et al., 2012). One possible reason for this could be that eye movements inevitably contain individual differences among readers as well as unstable movements caused by various external or internal factors, which make it difficult to extract general reading strategies from gaze data obtained from different readers or even from a single reader.

In this research, we explore whether this difficulty can be overcome. We aim to predict whether each word in the text is fixated by training conditional random field (CRF) models on existing gaze data (Kennedy, 2003), and then examining whether such fixation behavior can be sufficiently explained from the viewpoint of NLP-based linguistic features.

In the experiments, the trained CRF models predicted word fixations with 73% to 84% accuracy for each subject. While the accuracy does not seem high enough to explain human gaze behavior, a CRF model trained on the merged gaze data of all the subjects can predict the fixation distribution across subjects for each word with a similarity of 0.9462 to the observed distribution, which should be high enough to extract a general distribution regardless of individual differences or unstable movements in the gaze data. The experimental results also show that to capture human reading behavior correctly, both lexical and screen position features are essential, which would suggest that we need to adequately distinguish the effects of these two kinds of features on gaze data when incorporating certain strategies from gaze data into NLP technologies.

In Section 2, we discuss related work on analyzing gaze data obtained while reading text. In Section 3, we briefly explain the fundamental concepts of gaze data by introducing existing gaze data in the form of the Dundee Corpus, and also introduce the CRF model, which is trained to predict word-based fixations. In Section 4, we discuss preprocessing and observation of the Dundee Corpus in designing our model. Finally, in Sections 5 and 6, we explain how to predict word-based fixations in the corpus and analyze the performance of our model, respectively.

2 Related work

In the field of psycholinguistics, eye movements while reading text is a well established research field (Rayner, 1998), and the accumulated knowledge has resulted in various models for eye move-

ments. E-Z Reader (Reichle et al., 1998, 2003, 2006) is one such model. The E-Z Reader was developed to explain how eye movements are generated for the target gaze data, and not to predict eye movements when reading text for the first time. These models are optimized for the target gaze data by adjusting certain parameters without including any machine learning approaches. On the other hand, the work presented in (Nilsson and Nivre, 2009) was, as the authors stated, the first work that incorporated a machine learning approach to model human eye movements. The authors predicted word-based fixations for unseen text using a transition-based model. In (Nilsson and Nivre, 2010), temporal features were also considered to predict the duration of fixations.

There are important differences between the two approaches mentioned above, other than the way in which the parameters are adjusted and the purpose of the modeling. The former approach modeled the average eye movement of the subjects, while the latter trained the model for each subject. The key point here is that the former approach attempts to generalize human eye-movement strategies, while the latter attempts to capture individual characteristics. Our final goal is not only to explain or predict human eye movements, but rather to extract from gaze data, reading strategies that can be imported into NLP technologies. Since it is not clear whether extracting individual or averaged strategies is better for this purpose, we set out to train our models to predict both word-based fixations for each subject and the total distribution of the behavior across the subjects.

An image-based approach was proposed in (Martínez-Gómez et al., 2012) to clarify the position in the text that should be fixated in order to understand the text more quickly. The authors represented words in the text as bounding boxes, and visualized each of the linguistic features of words as an image by setting the pixel values of the word-bounding boxes according to the magnitude of the feature values of the words. They then attempted to explain the target gaze data represented in the image using a linear sum of the weighted feature images. This work also incorporated screen position features of words by representing each linguistic feature in a text image, which meant that the screen position and linguistic features were considered to be strongly connected. In our models, on the other hand, these two features are described separately and then paired, since we need to exclude the contribution of screen position features when incorporating captured reading strategies into NLP technologies, where screen positions are rarely considered.

3 The target gaze data and the model used to analyze them

3.1 The Dundee Corpus

The Dundee Corpus (Kennedy, 2003) is a corpus of eye movement data obtained while reading English and French text. For each language, 20 texts from newspaper editorials (each of which contained around 2,800 words) were selected, and each of the texts was divided into 40 five-line screens containing 80 characters per line. While 10 native speakers read the texts displayed on the screen, an eye tracker was used to record the gaze points on the text every millisecond. Through their screen settings, patient calibration of the eye tracker, and post-adjustment of gaze data, the authors successfully controlled the error of each gaze point to be within a character. The gaze data included in the corpus, therefore, consisted of character-based fixations. Consecutive gaze points on a single character were reduced to a single fixation point with the combined duration (Figure 1).

Generally, an eye movement from one fixation point to another is called a *saccade*, and backward saccades are called *regressions*. In a saccade action, the human gaze usually moves several characters forward in the text, which means that some characters are not fixated. The reason for this is that humans can see and process the areas around fixated points, referred to as *peripheral fields*.

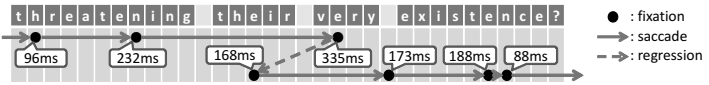


Figure 1: Character-based gaze data in the Dundee Corpus

3.2 Conditional random fields

CRFs (Lafferty et al., 2001) are a type of discriminative undirected probabilistic graphical model. Theoretically, CRFs can deal with various types of graph structures although we use CRFs for sequential labeling of whether each word is fixated. We therefore, explain CRFs with respect to sequences only, borrowing the explanation from (Sha and Pereira, 2003).

CRFs define the conditional probability distributions $p(\boldsymbol{Y}|\boldsymbol{X})$ of label sequences $\boldsymbol{Y}$ given input sequences $\boldsymbol{X}$. We assume that random variable sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same length, and that the generic input and label sequences are $\boldsymbol{x} = x_1 \cdots x_n$ and $\boldsymbol{y} = y_1 \cdots y_n$, respectively. A CRF on $(\boldsymbol{X}, \boldsymbol{Y})$ is specified by a vector $\boldsymbol{f}$ of *local features* and a corresponding *weight vector* $\boldsymbol{\lambda}$. Each local feature is either a state feature $s(y, \boldsymbol{x}, i)$ or a transition feature $t(y, y', \boldsymbol{x}, i)$ where y, y' are labels, $\boldsymbol{x}$ is an input sequence, and i is an input position. Typically, features depend on the inputs around the given position, although they may also depend on global properties of the input.

The CRF's global feature vector for input sequence $\boldsymbol{x}$ and label sequence $\boldsymbol{y}$ is given by $\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}) = \sum_i \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}, i)$, where i ranges over the input positions. The conditional probability distribution defined by the CRF is then $p_{\boldsymbol{\lambda}}(\boldsymbol{Y}|\boldsymbol{X}) = (1/Z_{\boldsymbol{\lambda}}(\boldsymbol{X})) \exp \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{Y}, \boldsymbol{X})$, where $Z_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$. The most likely label sequence for $\boldsymbol{x}$ is then given by $\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$. In our case, $\boldsymbol{x}$ represents the words in the text and $\boldsymbol{y}$ denotes whether each word is fixated.

4 Pre-processing and observation of the Dundee Corpus

In this section, we extract first-pass word-based fixations from the Dundee Corpus as the first step in our investigation. We then observe what types of information seem to determine word fixations/skips, which will help us to design feature sets for our CRF model in Section 5.

4.1 Extraction of first-pass word-based fixations from the Dundee Corpus

As a first step toward extracting reading strategies, we focus on word-based fixations ignoring their duration information, as examined in (Nilsson and Nivre, 2009). By merging consecutive fixations within a word into a single fixation, the resolution of the gaze data is reduced from a per character to a per word basis. Even after the merging, however, considering various types of observable behaviors at a time seems too complicated for the first step. We therefore further narrow our target by excluding regressions and saccades crossing lines from the gaze data as follows.

[Step 1] Each word-fixation is dealt with according to **(i)** and **(ii)**.
- **(i)** Omit the fixation from the gaze data and move to the next fixation if a fixated word
  - **(a)** is labeled "*visited*" or **(b)** is in a different line from a previously-fixated word.
- **(ii)** Else, allocate "*visited*" labels to the fixated word and all the preceding words in the text.

[Step 2] A sequence of gaze data is reconstructed using the remaining fixations.

For the gaze data in Figure 1, for example, character-based fixations are first merged into word-based fixations, the fixation after the regression from *very* to *their* is then ignored, and thereafter the gaze data are reconstructed as shown in Figure 2. With the data obtained from the above operation,

Figure 2: First-pass word-based fixations in the Dundee Corpus

Subject	Total no. of saccades	No. of words in word sequence skipped by saccade								
		0	1	2	3	4	5	6	7	⋯
A	31,431	17,683	8,831	3,928	777	144	30	16	8	⋯
B	36,248	24,669	8,900	2,118	419	106	28	3	1	⋯
C	37,657	26,348	9,369	1,704	168	32	16	12	3	⋯
D	36,570	24,560	10,044	1,750	143	40	14	10	4	⋯
E	32,442	18,896	9,023	3,672	755	77	16	2	1	⋯
F	38,982	28,561	8,859	1,351	159	36	10	3	1	⋯
G	38,910	28,640	8,324	1,732	160	25	13	7	2	⋯
H	33,910	20,540	10,068	2,807	384	78	18	8	1	⋯
I	36,717	24,957	9,117	2,393	216	23	8	1	0	⋯
J	37,738	26,479	9,297	1,774	136	32	12	2	2	⋯
Avg.	36,060.5	24,133.3	9,183.2	2,322.9	331.7	59.3	16.5	6.4	2.3	⋯
	(100.00%)	(66.91%)	(25.46%)	(6.44%)	(0.92%)	(0.16%)	(0.05%)	(0.02%)	(0.01%)	⋯

Table 1: Frequency of number of words in skipped sequence per subject

we can focus only on word-fixations involved in first-pass forward saccades within single lines.

4.2 Observation of skipped words in the Dundee Corpus

When observing the gaze data obtained in the previous section, we can see that for each subject many words were skipped by saccades, that is, not fixated at all. We consider that such skips would reduce the time for word-fixations and therefore lead to more effective human reading, that is, faster reading without sacrificing understanding. Here we explore this word-skip behavior in the gaze data in order to utilize the characteristics thereof to model word-fixations in the experiments.

Table 1 shows the number of saccades per subject for the 20 texts of the Dundee Corpus (second column), and classifies these saccades according to how many consecutive words the subject skipped (third column onwards). The numbers in parentheses at the bottom of the table show the ratios of the number of saccades skipping a particular number of words against the total number of saccades. According to this table, the number of saccades skipping up to three words constitutes 99.73% of the total number of saccades. Even if we omit the number of saccades that move to the next word (shown in the third column) from our calculations, the number of saccades skipping one to three words constitutes 99.18%. Based on this observation, the assumption that each saccade action skips at most three consecutive words appears to be realistic. If there is a common regularity within the skipped sequences that can determine whether a target sequence is skipped, predicting whether a target word is skipped would require lexical information on the preceding or following two words from the target word.

Table 2(a) shows the top 30 word sequences skipped by saccades in order of the number of skip times, averaged over the 10 subjects (leftmost values in the middle column). From this table, it seems that closed-class words such as determiners, prepositions, conjunctions, auxiliary verbs, and so on, are often skipped by saccades. When considering the ratio of skip times against total number of appearances of the target sequence (shown in the rightmost column), however, the frequently skipped sequences were not skipped with high frequencies. For example, *the* was skipped most often, although its skip rate was only 26.56%.

Table 2(b) shows the top 30 sequences in order of skip rates against number of appearances only for sequences that appeared ≥ 5 times in the corpus. As observed in Table 2(a), we can see that

(a) Frequently observed skips

Word sequence	# skips / # appearances	Ratio (%)
the	774.1 / 2915	26.56
of	592.9 / 1613	36.76
to	525.1 / 1442	36.41
and	430.4 / 1079	39.89
a	402.7 / 1260	31.96
in	320.7 / 934	34.34
that	201.7 / 731	27.59
is	185.8 / 625	29.73
for	146.6 / 436	33.62
The	134.9 / 319	42.29
on	121.3 / 364	33.32
as	107.2 / 348	30.80
of the	106.3 / 371	28.65
are	99.5 / 318	31.29
be	92.8 / 372	24.95
with	92.4 / 347	26.63
was	87.2 / 351	24.84
it	84.5 / 330	25.61
I	79.5 / 257	30.93
by	76.7 / 220	34.86
-	72.5 / 257	28.21
have	71.4 / 327	21.83
or	70.5 / 167	42.22
in the	68.6 / 271	25.31
at	67.4 / 220	30.64
has	64.8 / 208	31.15
from	63.1 / 215	29.35
he	59.7 / 182	32.80
but	56.7 / 170	33.35
an	51.8 / 174	29.77

(b) Sequences skipped with high rate (which appeared ≥ 5 times)

Word sequence	# skips / # appearances	Ratio (%)
His	4.8 / 8	60.00
Its	4.6 / 8	57.50
How	3.3 / 6	55.00
Of	6.7 / 13	51.54
From	3.9 / 8	48.75
A	21.7 / 46	47.17
or a	4.6 / 10	46.00
No	4.1 / 9	45.56
I'd	4.1 / 9	45.56
Ms	3.1 / 7	44.29
We	14.4 / 33	43.64
led	2.6 / 6	43.33
- in	3.0 / 7	42.86
Most	3.4 / 8	42.50
The	134.9 / 319	42.29
de	3.8 / 9	42.22
&	3.8 / 9	42.22
or	70.5 / 167	42.22
of a	30.7 / 73	42.05
Is	2.1 / 5	42.00
- is	2.5 / 6	41.67
It's	6.1 / 15	40.67
as a	20.9 / 52	40.19
'We	2.4 / 6	40.00
Those	2.4 / 6	40.00
he's	2.4 / 6	40.00
- a	3.6 / 9	40.00
He	19.6 / 49	40.00
25	2.4 / 6	40.00
and	430.4 / 1079	39.89

(c) Skipped 2 or 3 word sequences (which appeared ≥ 5 times)

Word sequence	# skips / # appearances	Ratio (%)
or a	4.6 / 10	46.00
- in	3.0 / 7	42.86
of a	30.7 / 73	42.05
- is	2.5 / 6	41.67
as a	20.9 / 52	40.19
- a	3.6 / 9	40.00
to a	13.4 / 34	39.41
and so	1.9 / 5	38.00
in a	22.9 / 64	35.78
- the	4.5 / 13	34.62
of us	3.1 / 9	34.44
In a	2.4 / 7	34.29
up a	1.7 / 5	34.00
than a	4.4 / 13	33.85
and to	2.0 / 6	33.33
to be a	2.8 / 11	25.45
many of the	0.4 / 5	8.00
to do with	0.4 / 5	8.00
is not a	0.4 / 5	8.00
would be a	0.6 / 8	7.50
it is a	0.5 / 7	7.14
is that the	0.4 / 6	6.67
to make a	0.3 / 5	6.00
have been a	0.3 / 5	6.00
it is the	0.4 / 7	5.71
that it is	0.3 / 7	4.29
as much as	0.2 / 5	4.00
in order to	0.2 / 5	4.00
because of the	0.2 / 6	3.33
in the same	0.2 / 6	3.33

Table 2: Word sequences skipped by saccades in the Dundee Corpus

closed-class words are once again in the majority while first (capitalized) words in sentences were frequently skipped, although their skip rates were, as before, not that high. Even *His* at the top of the table was skipped with a rate of only 60.00%. Table 2(c) shows the top 15 sequences based on the same criteria used in Table 2(b), but only for two- and three-word sequences. The table suggests that word sequences connecting something like NP chunks tended to be skipped, although their skip rates were not that high.

These observations suggest that target word sequences themselves seem to be related to whether they are skipped, while other factors, such as relations with surrounding words, and so on, should also be considered in skip decisions. Based on the above, we aim to capture factors for word-skip behaviors using features in the CRF models. Using CRF models trained on the gaze data, we examine how well the factors implemented as features can explain gaze behaviors.

The main purpose of this research was to capture some generality in human reading strategies from an NLP perspective. From this point of view, it is desirable to be able to explain gaze behaviors mainly using combinations of lexical information, in the normal way for NLP. For example, the width of peripheral fields and the range of saccades, which are given by human eye mechanisms, have long since been shown to control gaze behavior in psycholinguistic fields, whereas we aim to interpret them in terms of window size, word length, and so on.

Early in this section we assumed that the length of each skipped sequence is at most three words. We then attempt to predict a fixation or skip behavior for each word using lexical information on the word and the preceding and following two words, which implies a window size of five words.

Subject	No. of skipped / all words (rate)
A	20,048 / 51,501 (38.93%)
B	15,224 / 51,501 (29.56%)
C	13,817 / 51,501 (26.83%)
D	14,890 / 51,501 (28.91%)
E	19,039 / 51,501 (36.97%)
F	12,490 / 51,501 (24.25%)
G	12,570 / 51,501 (24.41%)
H	17,563 / 51,501 (34.10%)
I	14,763 / 51,501 (28.67%)
J	13,736 / 51,501 (26.67%)

Table 3: Rate of skipped words

(No. of words)

Condition for agreement	Total (rate) = Skipped + Fixated
$\geq$ 6 subjects displaying same behavior	47,320 (91.88%) = 10,109 + 37,211
$\geq$ 7 subjects displaying same behavior	39,439 (76.58%) = 6,484 + 32,955
$\geq$ 8 subjects displaying same behavior	31,855 (61.85%) = 3,473 + 28,382
$\geq$ 9 subjects displaying same behavior	24,219 (47.03%) = 1,385 + 22,834
10 (all) subjects displaying same behavior	16,313 (31.68%) = 314 + 15,999
Total words in all texts	51,501

Table 4: Agreement on gaze behavior for each word

The level of lexical information can vary, such as surface form, POS, length, probability, etc., while various combinations of these can also be considered. On the other hand, since text is displayed on a screen, optical factors must also be considered. In this research, we consider one of the most likely factors, that is, the screen position of each word. In the experiments in Sections 5 and 6, we examine the contribution of these factors by representing them as features in the CRF models.

4.3 Observation of commonality in gaze behaviors among subjects

This section investigates a method for capturing generality in gaze behavior among subjects. Using the gaze data (obtained in Section 4.1), Table 3 gives the number of words that were skipped by each subject. From this table, we can roughly see some variability in gaze behavior among subjects. Table 4 shows the degree of agreement among the subjects on whether each word is fixated or skipped. For each row, the table shows the number of words for which a minimum number of subjects displayed the same behavior. For example, words for which all the subjects displayed the same behavior comprised only 31.68% of the texts. The low agreement given in the table would suggest that it is not a good idea to specify a single common behavior for each word.

Based on this observation, we attempted instead to capture the distribution of how many subjects fixated or skipped each target word. We trained a CRF model on the merged gaze data for all 10 subjects, using the same feature set as in the model for each subject, and then used the obtained model to predict the distribution of each word in a target text.

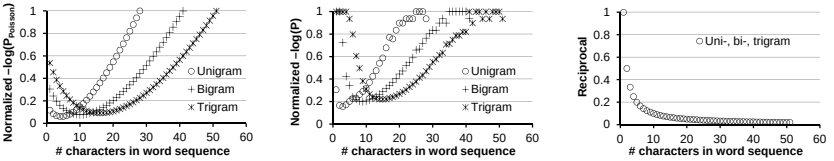
5 Experimental settings

Based on the observation in the previous section, we examine whether word-fixations can be predicted using CRF models, which are trained on the gaze data. In this section, we explain the experimental settings mainly of features that are utilized to train CRF models.

5.1 General settings

For the experiments, we trained a CRF model on the gaze data for each subject to predict the fixation/skip behavior of the subject for each word. In addition, we also trained a CRF model on the merged data for all subjects, to predict the fixation/skip distributions of each word across the subjects. The evaluation metrics for the models are given in Section 5.3.

For gaze data, we utilized the Dundee Corpus. As introduced in Section 3.1, the Dundee Corpus consists of gaze data for 20 texts, each of which was read by 10 subjects. We then divided the data into training data, consisting of the data for 18 texts, and test data, comprising data for the remaining two texts. All the gaze data were converted into first-pass saccade data according to Section 4.1, where each word was labeled “skipped” or “fixated” for each of the subjects. In the



(a) L_POIS (b) L_PROB (c) L_RECI

Figure 3: Word length features

Dundee Corpus, symbols such as quotation marks, periods, and commas are concatenated with the nearest words. Considering the effect of this on gaze behavior, words in other tools were treated in the same manner. For the same reason, we left the capitalization of words unchanged.

To train the CRF models, we utilized *CRFsuite* (Okazaki, 2007) ver. 0.12. We used a sentence as an input/output unit, since many of the existing NLP technologies are based on sentence-level processing, and we intend to associate outputs of the CRF models with NLP technologies in our future work. To obtain input sentences, five 80-character lines in each screen were split into sentences using the sentence splitter implemented in the *Enju* parser (Ninomiya et al., 2007)[1]. In training the CRF models, we selected the option of maximizing the logarithm of the training data with an L1 regularization term, since this would effectively eliminate useless features, thereby highlighting those features that really contributed to capturing the gaze data. The coefficient for L1 regularization in each model was adjusted in the test data to examine to what extent we could explain the given data using our features. We next explain the features utilized for training our CRF models.

5.2 Features utilized for training CRF models

Based on the observation in Section 4.2, we set up features to capture the reading strategies. The examined features can be classified into two types: lexical features and screen position features. For each target word, we considered the features on the target word, the preceding two words, and the following two words, which implies a window size of five words. Within the window size, we considered all possible uni-, by-, and trigrams for each feature, except for **3G-F** and **3G-B**.

[Lexical features]

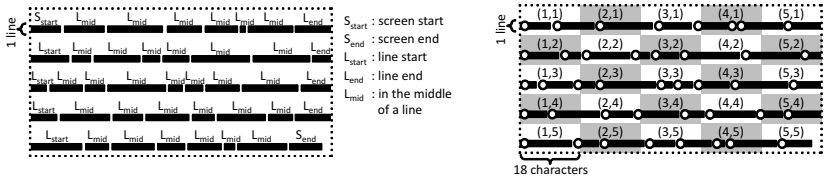
· **WORD**: word surface(s).

· **POS**: part(s) of speech obtained applying the POS tagger (Tsuruoka et al., 2005) to each sentence.

· **L-POIS, L-PROB, L-RECI**: information on surprisal of word length (real-value features). **L-POIS** assumes that the word length probability follows a Poisson distribution, and takes the logarithm of the probability of the target word length. The logarithmic values are normalized over the words in the texts (Figure 3(a)). **L-PROB** calculates the actual word length probability in the training data, takes the logarithm of the obtained probability, and then normalizes the logarithm (Figure 3(b)). **L-RECI** merely takes the reciprocal of the word length (Figure 3(c)). For all of the above three features, when obtaining bi- and trigrams, we summed the length of each of the words and single space characters inserted between them.

· **3G-F, 3G-B**: surprisal of a forward or backward word trigram (real-value features). We first obtained the probabilistic distribution of forward or backward trigrams by training the trigram lan-

[1] http://www.nactem.ac.uk/enju/index.html



(a) LF: line- or screen-feed features

(b) SC: screen coordinate features

Figure 4: Screen position features

Subjects	A	B	C	D	E	F	G	H	I	J
# fixated words	3,076	3,366	3,716	3,761	3,225	3,906	3,878	3,389	3,443	3,679
(Rate (%))	(62.67)	(68.58)	(75.71)	(76.63)	(65.71)	(79.58)	(79.01)	(69.05)	(70.15)	(74.96)
# words in test data	4,908 (100.00%)									

Table 5: Baseline rates for fixated words in the test data

guage model using SRILM (Stolcke, 2002) on the section of "Agence France-Presse, English Service" in the fourth edition of English Gigaword (Parker et al., 2009), which contains 466,718,000 words. The obtained probabilities for target trigrams were then converted into logarithmic values, and thereafter normalized over the trigrams in the texts.

[Screen position features]
· **LF**: line- or screen-feed. This examines whether the target word is at the beginning or end of a line (L_{start} / L_{end}) or the screen (S_{start} / S_{end}) (see Figure 4(a)).
· **SC**: screen coordinates. This divides each screen into 5×5 grids and examines in which grid the beginning of the word falls. Each screen in the Dundee Corpus consists of five 80-character lines, and therefore, one grid has the capacity to hold 1×16 characters (see Figure 4(b)).

5.3 Evaluation metrics and baselines

To evaluate the model trained on the gaze data for each subject, we counted the number of words in the test data for which the model correctly predicted the subject's behavior. Based on the observation that words were more often fixated than skipped for all subjects (see Table 3), we regarded the rate of fixated words in the gaze data for each subject as the baseline accuracy (see Table 5).

For the model trained on the merged data of all subjects, we first predicted the fixation/skip distributions of each word across the subjects for the test set. For each predicted distribution, the similarity based on Kullback-Leibler divergence was calculated against the distribution observed in the gaze data. Then, we took the average of the similarities over all words in the test set.

More precisely, we calculated $\exp\{-(1/|T|)\sum_{t\in T}\sum_i p_{i,t}\log_e(p_{i,t}/q_{i,t})\}$ where set T represents a target text in which each word $t \in T$ is identified with its position in the text. $|T|$ is accordingly the number of words in text T, $i \in \{''\text{fixated}'', ''\text{skipped}''\}$ is the label given to each $t \in T$, and $p_{i,t}$ and $q_{i,t}$ are the "fixated" / "skipped" distributions of target word t across the subjects, predicted by the CRF model and observed in the gaze data, respectively. This similarity measure returns values between $(0, 1]$; it returns 1 if the two distributions are the same. Using this similarity, we examined how well our model could capture generality in the reading strategies of all subjects.

Utilized feature types	Merged	Subjects									
		A	B	C	D	E	F	G	H	I	J
(Baseline)	.8131	62.67	68.58	75.71	76.63	65.71	79.58	79.01	69.05	70.15	74.96
WORD	.8803	68.42	70.88	76.65	80.05	70.50	79.58	79.20	70.19	72.21	77.16
POS	.8683	67.24	69.80	75.61	78.02	69.58	79.65	79.07	69.09	71.62	76.10
3G-F	.8505	64.57	68.79	75.08	75.53	66.91	79.60	79.01	67.95	69.95	75.16
3G-B	.8489	64.85	68.68	74.51	75.00	66.10	79.65	79.01	67.69	69.82	75.08
L-POIS	.8321	63.18	68.62	75.75	76.63	65.71	79.58	79.03	69.05	70.40	74.98
L-PROB	.8591	67.60	68.95	75.81	77.81	69.34	79.58	79.05	69.38	71.35	75.31
L-RECI	.8798	67.22	70.17	77.30	80.44	69.72	79.56	79.18	70.42	72.51	75.67
LF	.8663	60.96	68.58	75.65	76.83	63.12	79.58	79.01	68.38	70.44	74.96
SC	.8725	64.28	69.09	76.00	76.98	66.69	79.63	79.07	69.60	71.31	75.45
(Using all of the above)	.9462	75.24	74.37	81.05	83.94	76.51	80.48	82.62	72.98	77.69	81.11

"Merged" denotes the similarity of the distribution to the test data; "Subjects" gives the accuracy (%) of predicting word fixations/skips

Table 6: Prediction accuracy of word fixation/skip behavior (using individual features)

Utilized feature types	Merged	Subjects									
		A	B	C	D	E	F	G	H	I	J
(All individual types)	.9462	75.24	74.37	81.05	83.94	76.51	80.48	82.62	72.98	77.69	81.11
−WORD	.9460	75.06	74.67	80.75	83.99	76.51	80.50	82.38	72.84	77.51	80.58
−POS	.9457	75.02	74.33	80.91	83.99	76.24	80.34	82.46	72.72	77.71	80.81
−3G-F	.9460	75.39	74.37	80.85	83.80	76.43	80.54	82.80	72.66	77.73	81.50
−3G-B	.9463	75.04	74.49	81.03	83.88	76.47	80.48	82.58	72.84	77.73	81.48
−L-POIS	.9462	75.18	74.35	80.70	83.96	76.49	80.52	82.62	72.88	77.67	81.46
−L-PROB	.9453	75.45	74.39	80.97	83.62	76.49	80.56	82.40	72.62	77.63	81.50
−L-RECI	.9453	74.90	74.49	80.79	83.09	76.49	80.30	82.27	72.96	78.63	81.56
−LF	.9447	74.57	74.63	81.01	83.76	76.49	80.70	82.80	73.11	77.89	81.48
−SC	.9439	74.19	74.29	80.70	83.88	76.41	80.26	81.11	72.96	77.18	81.21

"Merged" denotes the similarity of the distribution to the test data; "Subjects" gives the accuracy (%) of predicting word fixations/skips

Table 7: Contribution of individual features to prediction accuracy

For the baseline of this similarity measure, we averaged over the training data the fixation/skip distributions of each word across the subjects, giving 0.8131.

6 Prediction of word-based fixation or skip behavior using CRF models

In the experiments, we first examine whether word fixation/skip behaviors in the test set can be explained using the trained CRF models. We then explore the individual contribution of each of the types of lexical and screen position features, and combinations of these features to prediction accuracy. We further observe which features are heavily weighted in the trained CRF model.

6.1 Individual contribution of each type of feature

Table 6 gives the prediction accuracy of the CRF models using each feature individually on the test data, as well as the CRF model using all of the given features. Each of the columns "A" to "J" gives the prediction accuracy for the target subject, given by the CRF models trained on training data for the target subject, while the "Merged" column gives a similarity-based evaluation of the CRF models trained on the merged gaze data of all subjects (see Section 5.3).

Using all the features, the trained CRF model gives between 0.90% and 12.57% higher accuracy than the baselines for each subject, and higher accuracy than using only individual features. The degree of contribution of each individual feature, however, seems to vary among subjects. For subjects A and E, the accuracy improvement over the baselines when using individual features is relatively higher than for other subjects. For subjects B, D, I, and J, an improvement is also observed, but this is less than for subjects A and E. For subjects F and G, on the other hand, barely any improvement is observed for all individual features. From these observations, although there

Utilized feature types	Merged	Subjects									
		A	B	C	D	E	F	G	H	I	J
(All individual types)	0.9462	75.24	74.37	81.05	83.94	76.51	80.48	82.62	72.98	77.69	81.11
−WORD, POS, 3G-F/-B	0.9437	74.53	74.39	80.52	83.68	75.94	80.42	82.23	72.82	77.63	80.56
−L-POIS/-PROB/-RECI	0.9353	73.63	73.98	80.38	82.86	75.59	80.22	82.09	72.58	77.53	81.03
−all lexical features	0.8748	64.61	68.97	75.86	76.87	66.40	79.60	79.07	69.27	71.03	75.45
−LF	0.9447	74.57	74.63	81.01	83.76	76.49	80.70	82.80	73.11	77.89	81.48
−SC	0.9439	74.19	74.29	80.70	83.88	76.41	80.26	81.11	72.96	77.18	81.21
−LF, SC	0.8940	68.93	70.90	77.49	81.09	71.11	79.54	79.67	70.48	72.84	78.26

"Merged" denotes the similarity of the distribution to the test data; "Subjects" gives the accuracy (%) of predicting word fixations/skips

Table 8: Contribution of lexical (upper part) and screen position (lower part) features to prediction

are individual differences in the degree of improvement among subjects, it seems that some of the characteristics of word-fixation behavior can be captured using our features. However, the 72% to 84% prediction accuracy obtained using all individual features is not high enough to adequately explain each subject's behavior. This is discussed further in Section 6.5.

For the CRF models trained on the merged gaze data of all subjects ("Merged" column), on the other hand, each of the individual features drastically improves the distribution similarity to the test data, and when using all features, the distribution similarity is 0.9462, which is an improvement of 0.1331 over the baseline similarity. This similarity bodes well in terms of our expectation that this CRF model can explain some generality on word-fixation behavior across all subjects.

When we go back to the prediction for each subject, each of **WORD**, **POS**, **L-PROB**, and **L-RECI** individually seem to be able to capture some characteristics in the gaze data, while **L-POIS** and the screen position features **LF** and **SC** do not improve the prediction accuracy that much. Table 7 examines the contribution of each individual feature to prediction accuracy, by training CRF models using all feature types except the target feature type. The table seems to show that removing the respective individual feature does not lead to a noticeable decrease in accuracy. This would suggest that each individual feature is complemented by the remaining features.

6.2 Contribution of lexical and screen position features

In order to explore the complementary characteristics of feature types, we start by focusing on the feature classification given by our definition: lexical and screen position features. Table 8 examines the contribution of lexical and screen position features to prediction accuracy. By removing all lexical features, that is, using only screen position features **LF** and **SC** (see "−all lexical features" row), the distribution similarity drops drastically by 0.0714, and prediction accuracy for each subject also decreases by between 0.88% and 10.63%. We observe similar characteristics by removing all screen position features; distribution similarity drops by 0.0522 (see "−**LF**, **SC**" row), while prediction accuracy for each subject also decreases by between 0.94% and 6.31%.

These observations suggest that both the lexical features and screen position features capture certain information that can only be captured by those features. In addition, the prediction accuracy obtained by removing all lexical features is similar to the baseline accuracy, regardless of the remaining screen position features. This would suggest that screen position features work well only in conjunction with lexical features. In other words, humans do not seem to be able to decide whether they fixate a word solely based on the word position.

The "−**WORD**, **POS**, **3G-F/-B**," and "−**L-POIS/-PROB/-RECI**" rows in the table show that removing either the features on word length surprisal or all lexical features other than these does

Utilized feature types	Merged	Subjects									
		A	B	C	D	E	F	G	H	I	J
Baseline	.8131	62.67	68.58	75.71	76.63	65.71	79.58	79.01	69.05	70.15	74.96
All individual types (AIT)	.9462	75.24	74.37	81.05	83.94	76.51	80.48	82.62	72.98	77.69	81.11
WORD, POS	.8805	68.58	70.64	76.55	79.97	70.64	79.60	79.18	69.89	72.07	76.81
WORD∗POS,WORD,POS	.8802	68.56	70.60	76.67	80.26	70.74	79.60	79.18	69.99	72.00	76.87
AIT, WORD∗POS	.9461	75.26	74.31	80.91	84.01	76.59	80.48	82.58	72.90	77.63	81.38
LF, SC	.8748	64.61	68.97	75.86	76.87	66.40	79.60	79.07	69.27	71.03	75.45
LF∗SC, LF, SC	.8750	64.98	69.01	75.92	76.85	66.50	79.60	79.01	69.32	71.03	75.45
AIT, LF∗SC	.9463	75.18	74.71	80.83	84.01	76.57	80.44	82.60	72.84	77.85	81.46
WORD, LF	.9322	73.08	73.61	80.11	82.76	75.49	80.64	80.48	72.62	77.24	80.50
WORD∗LF, WORD, LF	.9336	73.43	73.78	80.15	83.01	76.08	80.70	80.46	72.70	77.28	80.46
AIT, WORD∗LF	.9470	75.04	74.23	80.97	83.92	76.69	80.44	82.72	72.90	77.67	81.72
WORD, SC	.9328	73.02	73.92	80.56	82.93	75.71	80.75	82.19	73.19	77.26	81.05
WORD∗SC, WORD, SC	.9333	72.98	73.90	80.58	82.95	75.86	80.73	82.21	73.17	77.44	80.99
AIT, WORD∗SC	.9468	75.35	74.47	80.73	83.96	76.65	80.50	82.62	72.82	77.77	81.48
POS, LF	.9187	72.09	72.94	78.93	80.79	74.65	79.50	79.93	71.35	76.10	78.93
POS∗LF, POS, LF	.9201	73.11	73.08	78.79	80.93	75.26	79.16	79.56	71.31	76.14	79.03
AIT, POS∗LF	.9475	75.06	74.71	80.62	83.99	76.77	80.54	82.46	72.90	77.75	81.52
POS, SC	.9190	72.39	73.08	79.30	80.93	75.06	79.73	80.73	71.84	76.43	79.60
POS∗SC, POS, SC	.9196	72.56	73.04	79.69	80.97	75.08	79.75	80.75	71.84	76.49	79.60
AIT, POS∗SC	.9473	75.18	74.71	80.68	83.99	76.63	80.46	82.64	72.76	77.79	81.09
AIT, all combination	.9481	74.96	74.61	80.66	83.94	76.63	80.54	82.64	72.98	77.77	81.28

"Merged" denotes the similarity of the distribution to the test data; "Subjects" gives the accuracy (%) of predicting word fixations/skips

Table 9: Prediction accuracy of word fixation/skip behavior (using combined features)

not bring about a serious decline in prediction accuracy. Considering that lexical features other than the word length features, such as **WORD**, can implicitly capture a great deal of information on word length, most of the lexical information affecting word fixations/skips seems to be word length surprisal. The "$-$**LF**" and "$-$**SC**" rows in the table, on the other hand, show that removing either screen coordinate features or line-/screen-feed features does not bring about a serious decline in prediction accuracy. Considering that most of the line-/screen-feed information is implicitly contained in the screen coordinate information, most of the screen position information affecting word fixations/skips seems to be whether a target word is at the beginning or end of a line/screen.

6.3 Contribution of combined features

We also considered combinations of two feature types. Table 9 shows the contribution of each combination of features to prediction accuracy. In the table, **A**∗**B** represents the combination of feature types **A** and **B**, which means that this combined feature is fired only when both **A** and **B** are fired. Some feature types are real-value features, and cannot easily be combined with other feature types. We therefore, omitted the real-value features as candidates for combination. When using each combined feature, we also added the respective individual features for smoothing.

From the table, we can see that adding each of the combined features barely contributes to any accuracy improvement. Even when using all the individual and combined features (see the bottom row of the table), the improvement over using only all the individual features is barely noticeable. These observations seem to imply that combining the features does not capture any extra information than when using the features separately. Owing to a lack of gaze data, these results may be misleading, and further investigation would be required in order to continue this discussion.

6.4 Observation of heavily weighted features

From the heavily weighted features in the CRF model, we observed which features were regarded as important for explaining the gaze data. Table 10 shows the heavily weighted features in the CRF

Features (for fixations)	Weight	Features (for fixations)	Weight
L-PROB[0]	5.7808	L-RECI[-1]	0.1651
LF[0]= L_{end}	1.3306	SC[-2,-1]=(5,4),(5,4)	0.1639
LF[0]= L_{start}	1.3210	LF[-1,0]= L_{mid}, L_{end}	0.1519
LF[0]= S_{end}	1.2605	SC[+2]=(1,5)	0.1454
L-POIS[-1,0]	1.2218	SC[+1,+2]=(1,3),(1,3)	0.1347
L-PROB[-1]	0.7899	SC[0,+1,+2]=(5,3),(5,3),(1,4)	0.1299
L-RECI[-2,-1]	0.5393	WORD[-1]=But	0.1284
SC[+1]=(1,5)	0.4001	SC[-2,-1]=(5,1),(5,1)	0.1258
LF[+1]= L_{start}	0.3422	LF[-1]= L_{end}	0.1248
LF[0,+1]= L_{end}, L_{start}	0.3422	LF[-1,0]= L_{end}, L_{start}	0.1248
LF[0,+1]= L_{start}, L_{mid}	0.3265	LF[0,+1]= S_{end}, S_{start}	0.1232
SC[+1]=(1,3)	0.2987	LF[+1]= S_{start}	0.1232
SC[+1]=(1,4)	0.2776	SC[+2]=(1,2)	0.1182
L-PROB[-2,-1,0]	0.2310	SC[+2]=(1,3)	0.1146
3G-F[-2,-1,0]	0.2090	LF[-2]= L_{mid}	0.1092
SC[0]=(5,5)	0.1867	SC[0,+1]=(5,5),(1,1)	0.1079
SC[+1,+2]=(1,1),(1,1)	0.1832	POS[0]=CD	0.1047
SC[-1]=(5,5)	0.1721	SC[-1]=(5,4)	0.1029
SC[+1,+2]=(1,2),(1,2)	0.1718	POS[0,+1]=NN, NNS	0.1014
SC[+1]=(1,2)	0.1695	SC[-2,-1]=(5,5),(5,5)	0.1012
SC[+1,+2]=(1,4),(1,4)	0.1660	SC[0,+1]=(1,4),(1,4)	0.1006

Features (for skips)	Weight
L-RECI[0]	2.0020
L-POIS[+1]	0.2691
Beginning of sentence	0.2657
End of sentence	0.2071
POS[-1]=_COLON_	0.2023
WORD[0]=it's	0.1904
WORD[-1]=-	0.1829
WORD[-1]=I	0.1793
LF[-2,-1,0]= L_{mid}, L_{mid}, L_{mid}	0.1756
L-PROB[-1,0]	0.1716
WORD[0]=than	0.1599
LF[0,+1]= L_{mid}, L_{mid}	0.1584
WORD[0]=that	0.1493
LF[0,+1,+2]= L_{mid}, L_{mid}, L_{mid}	0.1463
WORD[0]=and	0.1452
WORD[-1]=of	0.1289
WORD[-1,0]=as, a	0.1271
WORD[0]=from	0.1267
WORD[0]=which	0.1235
SC[-1,0,+1]=(1,1),(1,1),(1,1)	0.1224
LF[0]= L_{mid}	0.1157

Table 10: Features that were heavily weighted in the "Merged" model using all individual features

model that was trained using all individual features on the merged training data of all subjects. The left and right tables show the features weighted for fixations and skips, respectively. A number in square brackets [ ] represents a word whose feature was captured, and identified with an offset from a target word. A sequence of two or three numbers in [ ] represents bi- or trigram features.

The tables suggest that surprisal based on word length probability and the reciprocal word length of a target word (**L-PROB[0]** and **L-RECI[0]**, respectively) have a large influence on whether subjects fixate or skip the word, respectively. For **L-PROB[0]**, according to Figure 3(b), longer words tend to give greater surprisal. This may be because the longer length possibly suggests that the word is a content word and sometimes even an unknown word. In addition, it may be possible that a longer word cannot be skipped easily by a single saccade. The heavy weight for fixations thus seems reasonable. For **L-RECI[0]**, a large value for the reciprocal word length means that the word length is short, and a shorter length possibly suggests that the word is a functional word or easily skipped by a single saccade. The weight for skips thus seems reasonable. From the viewpoint of the human eye mechanism, these features would have been fired without a fixation on a target word, using information on the word obtained by peripheral fields of the eyes or guessed from surrounding information.

For **WORD** features, most of the heavily weighted features are for skips and on target words (**WORD[0]**) that belong to a closed-class, such as *than*, *from*, and *which*. These words are not content words and tend to be short, and therefore were likely weighted heavily for skips. On the other hand, **WORD[-1]=But** was heavily weighted for fixations. The reason for this may be that when a sentence starts with *But*, it attracts the reader's interest to focus on the next word.

For **SC** features, almost all of the heavily weighted features were located in the leftmost (1,*) or rightmost (5,*) coordinates, which is consistent with our analysis in Section 6.2. Many of these features were weighted for fixations for the simple reason that the next word was in the leftmost coordinate (**SC[+1]=(1,*)**), which would mean that subjects tended to fixate last words in a line before their linefeed eye movements. **SC[0]=(5,*)** with conditions similar to **SC[+1]=(1,*)** were not weighted that highly, probably because the first character of the last word in a line does not always appear in position (5,*).

6.5 Discussion on the experimental results

The experimental results in Section 6 show that the CRF model trained for each subject does not have high prediction accuracy. When we analyzed the prediction errors, we found many long spans in the gaze data where all words were fixated. The subjects seem to have read the spans very precisely, which differed from the behavior displayed in other areas. It is natural that subjects do not maintain the same level of concentration or understanding throughout a text, yet our model was not able to capture this. We believe that this is the main reason why the CRF model for each subject does not exhibit high prediction accuracy. This issue will be addressed in our future work.

On the other hand, the experimental results also suggest that we can predict the distribution of fixation/skip behavior of each word across subjects with very high similarity to the gaze data, regardless of individual differences among subjects (see Table 4) and the above unstable movements in gaze data. This would imply the possibility of capturing and explaining generality in human reading strategies from an NLP perspective.

It should also be noted that our results also depend on the preprocessing of the gaze data in Section 4.1. The authors in (Nilsson and Nivre, 2009) also used the Dundee Corpus, and trained and examined their model to predict word-based fixation behavior for each subject. Similar to our method, they applied some preprocessing to the gaze data to remove irregular eye movements, whereas, unlike our case, they also took regressions and revisits as well as first-pass forward saccades into consideration. Since the experimental settings differed, we cannot directly compare the prediction accuracy of our results with those in (Nilsson and Nivre, 2009). However, considering that our baselines seem to be higher than those in (Nilsson and Nivre, 2009), we could say that our additional preprocessing simplified the problem and made the gaze behavior easier to capture.

We found that both lexical features and screen position features contributed to explaining the gaze data. Our final goal is to obtain some reading strategies from the gaze data, which can then be imported into NLP technologies. Considering this goal, we need to remove the screen position factors from the gaze data, since most NLP technologies consider sentence-based processing without any position information. The experimental results suggest that combined features of screen position and lexical information do not capture any extra characteristics. If this is true, we may be able to separate the two factor types without considering their mutual interaction.

Conclusion

In this research, we examined the possibility of extracting reading strategies by observing word-based fixation behavior. We trained CRF models on gaze data to predict the gaze behavior of each subject and the distribution of gaze behavior across all subjects. Using lexical and screen position features, the CRF models could predict word fixation/skip behaviors for each subject with 73% to 84% accuracy as well as the distribution of word fixation/skip behaviors across the subjects with a 0.9462 similarity to the original gaze data.

In our future work, we would like to collect gaze data on specific linguistic phenomena, such as coordination and prepositional attachment, and then attempt to extract some general reading strategies from this gaze data. Having achieved this, we aim to import the obtained strategies into NLP technologies such as parsing, to realize further progress in these fields.

Acknowledgments

This research was partially supported by “Transdisciplinary Research Integration Center, Japan”, “Kakenhi, MEXT Japan [23650076]” and “JST PRESTO”.

References

Kennedy, A. (2003). *The Dundee Corpus [CD-ROM]*. School of Psychology, The University of Dundee.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Martínez-Gómez, P., Hara, T., Chen, C., Tomita, K., Kano, Y., and Aizawa, A. (2012). Synthesizing image representations of linguistic and topological features for predicting areas of attention. In *Proceedings of the 12th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 93–101, Kuching, Sarawak, Malaysia.

Nilsson, M. and Nivre, J. (2009). Learning where to look: Modeling eye movement in reading. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 93–101, Boulder, Colorado. Association for Computational Linguistics.

Nilsson, M. and Nivre, J. (2010). Towards a data-driven model of eye movement control in reading. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 63–71, Uppsala, Sweden. Association for Computational Linguistics.

Ninomiya, T., Matsuzaki, T., Miyao, Y., and Tsujii, J. (2007). A log-linear model with an n-gram reference distribution for accurate HPSG parsing. In *Proceedings of IWPT 2007*. Prague, Czech Republic.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). English Gigaword Fourth Edition. LDC2009T13.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. *PSYCHOLOGICAL REVIEW*, 105(1):125–157.

Reichle, E. D., Pollatsek, A., and Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cogn. Syst. Res.*, 7(1):4–22.

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral Brain Science*, 26(4):445–476.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. Int. Conf. Spoken Language Processing (ICSLP 2002)*.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, volume LNCS 3746, pages 382–392, Volos, Greece. ISSN 0302-9743.