An Eye-tracking Study of Named Entity Annotation

Tokunaga, Takenobu[†] **Nishikawa, Hitoshi**[†] **Iwakura, Tomoya**[‡]

[†]Tokyo Institute of Technology
{take, hitoshi}@c.titech.ac.jp

[‡]Fujitsu Laboratories LTD.
iwakura.tomoya@jp.fujitsu.com

Abstract

Utilising effective features in machine learning-based natural language processing (NLP) is crucial in achieving good performance for a given NLP task. The paper describes a pilot study on the analysis of eye-tracking data during named entity (NE) annotation, aiming at obtaining insights into effective features for the NE recognition task. The eye gaze data were collected from 10 annotators and analysed regarding working time and fixation distribution. The results of the preliminary qualitative analysis showed that human annotators tend to look at broader contexts around the target NE than recent state-of-the-art automatic NE recognition systems and to use predicate argument relations to identify the NE categories.

1 Introduction

Corpus-based natural language processing (NLP) has been the mainstream of NLP research for the past quarter of a century. In this approach, given a task and manually annotated answers in a corpus, machine learning (ML) techniques are employed to induce a system for the task. The system adopts various kinds of information in the texts for solving the task, which is represented as a set of features for the ML algorithm. These features have often been determined based on heuristics such as adopting words and their POS within a certain size of a window around target words. Features based on human linguistic knowledge have also been employed through referring to manually constructed linguistic resources such as WordNet (Miller, 1995) and linguistic theories such as Centring Theory (Grosz et al., 1995).

Unlike the past attempts, we explore effective features in human behaviour during their annotation of answers in corpora. Considering an NLP system as a replacement of annotators, the replacement could follow the human annotators on their annotation behaviour as well as their annotation results. The information that human annotators refer to during their annotation process would provide useful clues to find effective features for the ML algorithms.

Recently utilising human behaviour information, particularly eye gaze information in various NLP tasks has begun attracting attention. The target tasks are diverse, including word sense disambiguation (Joshi et al., 2013), named entity recognition (Tomanek et al., 2010), syntactic analysis (Barrett and Søgaard, 2015), coreference resolution (Ross et al., 2016), predicate argument structure analysis (Iida et al., 2013; Mitsuda et al., 2013; Maki et al., 2016), sentiment analysis (Joshi et al., 2014), sentence compression (Klerke et al., 2016) and translation (Mishra et al., 2013; Sajjad et al., 2016).

In the present work, following Tomanek et al. (2010), we adopt the named entity (NE) recognition task. Having been motivated to select difficult training instances for active learning, Tomanek et al. (2010) utilised annotator eye gaze during their annotation of NEs in texts. They tried to define “difficulty” of NE instances regarding a cognitive load for annotating each NE instance. Annotation time and eye gaze were used for explaining the cognitive load. By relating the cognitive load of NE instances to their linguistic characteristics, they extracted features for a regression model estimating the difficulty of NE instances. In their attempt the resolution of eye gaze was very coarse, i.e. four regions around a target NE word (above, left, right and below) in addition to the target word itself, and eye gaze data was not fully quantitatively utilised for estimating the cognitive load.

Proceedings of Recent Advances in Natural Language Processing, pages 758–764,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_097

The main advances of our work over Tomanek et al. (2010) are twofold. (1) We collect eye gaze on more precise regions, i.e. phrase chunks instead of the coarse regions around a target word. (2) We try to find effective features for solving the task itself instead of selecting training instances for the task. In what follows, we report a data collection experiment in which annotator eye gaze during their annotation are collected (section 2) and results of a preliminary qualitative analysis of the collected data (section 3).

2 Data Collection

2.1 Materials and Procedure

We conducted an experiment for collecting annotator eye gaze and tool operations during the annotation of NEs in Japanese texts. The material for annotation was selected from the development data of the IREX named entity recognition task[1] consisting of 1,279 Japanese news articles. The named entities in these articles have been manually annotated with one of 8 categories: person, location, organisation, artefact, date, time, money and percent.

All articles were processed by a Japanese syntactic analyser KNP 4.11[2] to automatically annotate the NE categories in the texts. To collect difficult NE instances for the automatic NE recognition system, we compared the human-annotated categories and the KNP's outputs and extracted incorrectly annotated instances for four kinds of categories: person, location, organisation and artefact, which were more difficult than the rest. By selecting a single incorrectly annotated NE for each text, we made a set of texts consisting of 72 texts each of which included only a single NE target. The average length of the texts is 315 characters, which is roughly equivalent to 150 English words. The numbers of each NE category are 11 persons, 15 locations, 29 organisations and 17 artefacts.

We recruited sixteen Japanese native speakers, six males and ten females; ten of them had some experience of text annotation but not necessarily the NE annotation. After having been explained the objective of the experiment and the operation of a custom-made annotation tool, the participants were instructed to assign one of the following categories to a highlighted NE in each text.

⟨PSN⟩ person or pseudo-person name

⟨LOC⟩ location names, addresses and name of natural things such as rivers and mountains

⟨ORG⟩ organisation names, group names etc.

⟨ART⟩ name of artefacts such as products, service

⟨OTH⟩ none of the above

⟨UKW⟩ unable to decide

The participant gaze during annotation was captured by the Tobii T60 eye tracker at intervals of 1/60 second. The display size was $1,280 \times 1,024$ pixels, and the distance between the display and participant's eyes was maintained at about 50 cm. A text on display was presented with the MS Gothic font at the size of 24×24 pixels in black colour with a white background. The line space was set to 72 pixels, and the top, left and right margins were set to 96 pixels respectively. A target NE was highlighted with a yellow background.

The text set was divided into two annotation sets each of which contains 36 texts. All participants had two sessions for these two sets in the same order. Since we had a small number of participants, we did not consider counterbalancing the text order. We allowed the participants to take a break as much as needed between the sessions. Before starting the actual sessions, the participants annotated five texts for practice which were not included in the annotation sets. The five-point calibration was run before starting each session.

A session starts with showing a marker at the centre of the display for guiding the annotator eye gaze to the display centre. Clicking the marker shows a text with a highlighted target NE to annotate. When the participant decides on its category, they click the target NE and choose its category from a pop-up menu shown at the target. Choosing a category makes the display back to the centred marker screen. This cycle goes on until the 36th text in a session. Three click time points: a click on the marker, a click on the target and a click on a category, were recorded for each text.

This task design is simplified than actual NE annotation in two respects: the task concerns only NE category classification without NE span identification, and a single target NE is specified in a text at a time. This simplification enables us to directly relate the annotator eye gaze and their decision process on the category for the target NE; thus it makes the analysis of the collected data easier.

[1] http://nlp.cs.nyu.edu/irex/Package/IREXfinalB.tar.gz
[2] http://nlp.ist.i.kyoto-u.ac.jp/?KNP

Table 1: Annotator Accuracy

annotator	01	02	04	08	10	11	12	13	14	15	ave.
accuracy	0.82	0.75	0.88	0.79	0.89	0.92	0.89	0.65	0.82	0.92	0.83

Table 2: Distribution of NEs over their Error Rate

error rate	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
#NEs	32	12	10	9	3	0	2	0	0	2	2

2.2 Results

Recent eye-tracking devices like Tobii have made it drastically easier to capture gaze positions. However, there remain eye-tracking errors in the experiments. Tobii delivers the gaze point regarding the display coordinates of both eyes separately together with the timestamp and the error code denoting the validity of the gaze point. Based on the Tobii's error code of each gaze point, we selected the data from the participants fulfilling the following conditions:

- the total error rate (total number of the erroneous gaze points against the number of overall gaze points in the two sessions) is less than 15%, and
- more than half texts have the text-wise error rate less than 10%.

As a result, we discarded data from six participants, retaining the data from ten. The six participants have annotation experience, and the rest four do not.

3 Data Analysis

3.1 Accuracy of Annotators

Table 1 shows the accuracy of each annotator, i.e. the ratio of the correct annotations by each annotator against the overall 72 NEs. The underlined annotators in Table 1 have some annotation experience but not necessarily the NE annotation. The table shows that past annotation experience does not necessarily work favourably. The average annotator accuracy is 0.83, which looks not so high comparing to the performance of the state-of-the-art automatic NE recognition systems (Iwakura, 2011; Darwish, 2013; Passos et al., 2014). However, considering that we collected NE instances that could not be correctly analysed by the automatic NE tagger, humans perform the NE recognition task far better than computers.

Table 3: Distribution of Working Time on NE

duration (sec)	T1	T1+T2	duration (sec)	T2
$(0, 4]$	18	2	$(0,1]$	0
$(4, 8]$	27	34	$(1,2]$	30
$(8,12]$	15	15	$(2,3]$	19
$(12,16]$	10	12	$(3,4]$	15
$(16,20]$	1	7	$(4,5]$	6
$(20,24]$	0	0	$(5,6]$	0
$(24,28]$	0	1	$(6,7]$	1
$(28,32]$	0	0	$(7,8]$	0
$(32,36]$	0	0	$(8,9]$	1
$(36,40]$	0	0		
$(40,44]$	1	1		

Table 4: Average Working Time for each Response Type

	T1		T2		
response	ave.	SD	ave.	SD	N
correct	6.7	7.4	2.2	2.3	599
incorrect	9.0	9.0	4.0	3.5	63
⟨OTH⟩	11.5	12.8	6.0	6.1	51
⟨UKW⟩	32.5	9.7	2.5	0.5	7

Table 2 shows the number of NEs according to their error rate. The error rate of an NE is defined by the ratio of the correctly responding annotators for the NE against the overall ten annotators. The greater error rate value indicates more difficult NEs. The table indicates that the most NEs are easy for a human to identify their category.

3.2 Working Time

Table 3 shows the distribution of average working time on each NE, in which T1 indicates a duration from clicking a start marker until clicking a target NE, T2 indicates a duration from clicking the NE until selecting its category. Comparing the distribution between T1 and T2, we can find that the distribution of T1 is more diffused than that of T2. Their standard deviations over the NE instances are 7.80 for T1 and 3.20 for T2. Considering this difference, we can assume that decision on the category is mostly made in T1. The Pearson's correlation coefficient between the NE error rate and the average T1 was calculated, resulting in a positive correlation (0.47; $p < 0.00005$), i.e.

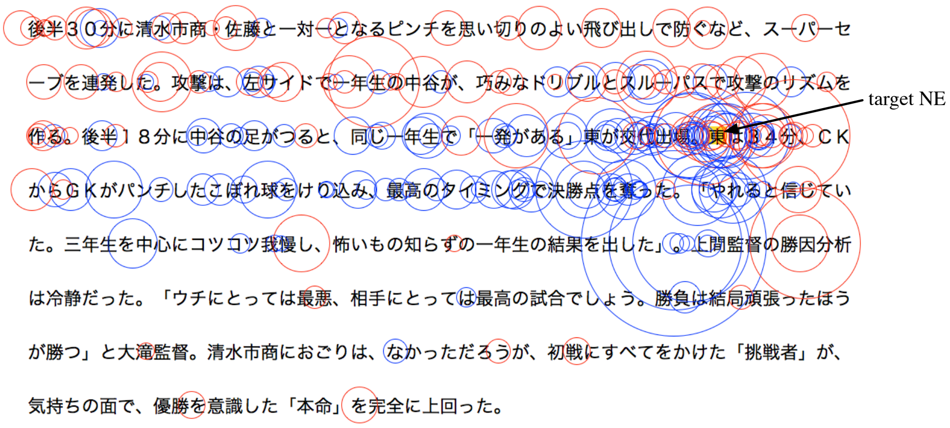
後半３０分に清水市商・佐藤と一対一となるピンチを思い切りのよい飛び出しで防ぐなど、スーパーセ
ーブを連発した。攻撃は、左サイドで一年生の中谷が、巧みなドリブルとスルーパスで攻撃のリズムを
作る。後半１８分に中谷の足がつると、同じ一年生で「一発がある」東が交代出場。東は３４分、ＣＫ
からＧＫがパンチしたこぼれ球をけり込み、最高のタイミングで決勝点を奪った。「やれると信じてい
た。三年生を中心にコツコツ我慢し、怖いもの知らずの一年生の結果を出した」。上間監督の勝因分析
は冷静だった。「ウチにとっては最悪、相手にとっては最高の試合でしょう。勝負は結局頑張ったほう
が勝つ」と大滝監督。清水市商におごりは、なかっただろうが、初戦にすべてをかけた「挑戦者」が、
気持ちの面で、優勝を意識した「本命」を完全に上回った。
target NE

Figure 1: Example of Fixations

the annotators tend to make mistakes for the NEs that require longer T1. There was no significant difference in T1 across the NE categories.

We further investigated the average working time for each type of responses: correct response, incorrect response, response with ⟨OTH⟩ and response with ⟨UKW⟩ as shown in Table 4. The table shows a tendency that annotators need more time when they make mistakes and choose ⟨OTH⟩ and ⟨UKW⟩ categories than when they correctly choose the category. Moreover, short T2 in cases for the correct response and ⟨UKW⟩ selection indicates that they already made a decision during T1, while longer T2 in the incorrect response and ⟨OTH⟩ selection indicates the annotators are still wondering during T2.

3.3 Preprocessing Eye-gaze Data

Eye-gaze data, a sequence of display coordinates with the timestamp, were grouped into gaze fixations based on their spatial and temporal closeness (Richardson et al., 2007). We used the Dispersion-Threshold Identification (I-DT) algorithm (Salvucci and Goldberg, 2000) for clustering eye-gaze data into a sequence of fixations. The I-DT algorithm has two parameters: spatial and temporal; considering the experimental configurations, i.e. (i) the display size and resolution, (ii) the distance between the display and the annotator's eyes, and (iii) the eye-tracker resolution, the spatial parameter was set to 24 pixels, and the temporal parameter was set to 100 msec following Richardson et al. (2007).

The current eye-tracking technology is more error-prone in the vertical direction than in the horizontal direction. Several methods for error correction in fixation coordinates have been proposed (Mishra et al., 2012; Cohen, 2013; Carl, 2013). They employ, however, task specific heuristics and are not necessarily applicable to our current task (Carl et al., 2008). To compensate tracking errors in the vertical direction, we took larger line spaces than usual, i.e. three character heights, and utilised a simple heuristics that a fixation located between lines was forced to aligned to the nearest line. We did not conduct any error correction for the horizontal direction.

Figure 1 shows fixations mapped on a text, where a circle indicates a fixation with its radius representing duration and its centre representing the centre of gravity of all gaze points belonging to the fixation cluster. The circle colour indicates different annotator groups, which are described later. The temporal information, i.e. a chronological sequence of fixations, is not presented in this figure. The target NE is highlighted with a yellow background (the 8th character from the right in the third line). The fixation on a word in texts is widely believed to have some relation with a cognitive process on that word (Just and Carpenter, 1980).

3.4 Distribution of Fixations

It is common to use a local context around the target NE for identifying its category in recent state-of-the-art automatic NE recognition. To be more concrete, surface strings and POS of the target NE and its neighbouring two words have been reported to be effective for NE recognition in many languages, e.g. English (Passos et al., 2014), Arabic (Darwish, 2013), and Japanese (Iwakura, 2011). We investigated to what extent a human also relies on the local context in the NE recog-

Table 5: Fixation Ratio in Local Contexts during T1

window width	±1 chunk		±2 chunks	
type/token	type	token	type	token
fixation frequency	0.24	0.34	0.31	0.41
fixation duration	0.24	0.38	0.31	0.44

nition. Since our target texts were in Japanese, we firstly segmented texts into phrasal chunks called *bunsetu* consisting of a sequence of content words followed by function words. We used an off-the-shelf analyser CaboCha[3] for the segmentation. Every fixation was then aligned to a chunk if the fixation centre fell within that chunk's bounding box on display. The average length of the chunks is 4.7 characters in our text set. Since the parafoveal vision in reading Japanese texts is reported to range from five to seven characters (Ikeda and Saida, 1978; Osaka, 1992) and a Japanese *bunsetu* is a basic unit for a grammatical role, it is reasonable to deal *bunsetu* chunks as the fixation target.

Table 5 shows the ratio of fixations locating within one or two chunks in both sides of the target NE during the T1 period, i.e. a duration from clicking the target until selecting its category. Assuming a word consists of two characters, a single chunk roughly corresponds to two words. We can see that only 24% in type and less than 40% in token of fixations are located within one chunk ($\sim$ two words) of both sides. Here the "token" column counts the fixated chunk tokens, while the "type" column counts the fixated chunk types. Even with the doubled context, i.e. two chunks wide, this ratio does not rise significantly. This observation suggests that a human tends to look at broader contexts than the automatic NE recognition, and this difference might suggest clues to improve the performance of the automatic NE recognition.

3.5 Fixations for Correct Annotation

We investigated the difference in fixation distribution between the annotators who correctly identified the NE category and those who did not. As shown in Table 2, the most NEs were correctly annotated by almost all the annotators, i.e. many NEs have small error rate values. To collect the NEs on which the numbers of correct and incorrect annotations are comparable, we chose the 14 NEs the error rate of which ranges from 0.3 to 0.6 for the further analysis. To see the difference of fixations between the correct and incorrect annotator groups, we calculated the number of fixations and the sum of fixation duration on each chunk normalised by the number of annotators in each group. Having investigated the differences of these metrics, we observed the following tendencies.

First, in nine out of the 14 NEs, the correct annotator group tends to look at the predicate that takes the target NE as an argument, and other argument chunks that share the same predicate with the target NE argument more than the incorrect annotator group. This suggests that predicate argument relations involving the target NEs would provide effective information for identifying NE categories. For instance, in Figure 1, the colour of the fixations denotes the corresponding annotator groups: correct one (blue) and incorrect one (red). The target NE is located in the right part of the third line. This example illustrates that the correct group looks at the succeeding context which includes several predicates having the target NE as a nominative argument, while the incorrect group looks at the preceding context where there is no chunk having predicate argument relations with the target NE. This observation is consistent with the result by Sasano and Kurohashi (2008) that empirically showed the effectiveness of dependency relations for named entity recognition.

Second, the range of fixation distribution is not decisive for the correct annotation. The broad look is necessary for the correct annotation in some cases, but concentrated fixations in a local context around the target NE do not necessarily suggest incorrect annotations. There are cases in which the annotators are biased toward an incorrect category because of the local context containing decisive clues for that category, although the correct one can only be found by referring to broader contexts. On the other hand, there are cases that the correct category is impossible to be guessed even having read the whole text. Such cases crucially require annotator's background knowledge for the correct annotation.

4 Concluding Remarks

This paper described a pilot study on the analysis of eye-tracking data during named entity (NE) annotation. The results revealed potential usefulness

[3] http://taku910.github.io/cabocha/

of eye gaze data in designing an effective feature set for the NE recognition. Particularly it showed that human annotators tend to use predicate argument relations, and they look at broad contexts around the target NEs. It would be interesting to see to what extent the latter tendency, i.e. considering broader contexts, could be captured by the recent LSTM-based NE methods (Ma and Hovy, 2016; Chiu and Nichols, 2016).

However, this study remains in its pilot phase because of the limitation of the collected data size. To confirm the present preliminary results, analysis with larger data would be indispensable. At the same time, selection of NE instances to be annotated in the data collection should be designed carefully. In the present study, we used the NE instances that were not correctly analysed with one of the state-of-art NE recognition systems. The eye gaze data for the NE instances that can be correctly analysed by the systems should also be collected and analysed. Because of the limited data size, we have not conducted analysis on a chronological sequence of fixations. That kind of analysis is also necessary with large scale data.

Other future research direction includes concretising the feature design and the evaluation of its effectiveness through NE recognition systems. Also, a further analysis of the eye gaze data is necessary for other potential purposes such as improving annotator expertise by comparing the gaze patterns of novice and expert annotators and enhancing the usability of annotation tools.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP16H02865.

References

Maria Barrett and Anders Søgaard. 2015. Using reading behavior to predict grammatical functions. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*. pages 1–5.

Michael Carl. 2013. Dynamic programming for re-mapping noisy fixations in translation tasks. *Journal of Eye Movement Research* 6(5):1–11.

Michael Carl, Arnt Lykke Jakobse, and Oleg Spakov. 2008. Towards an annotation standard for eye tracking data. In *Proceedings of Measuring Behavior*. page 223.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4:357–370.

Andrew L. Cohen. 2013. Software for the automatic correction of recorded eye fixation locations in reading experiments. *Behavior Research Methods* 45(3):679–683.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. pages 1558–1567.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225.

Ryu Iida, Koh Mitsuda, and Takenobu Tokunaga. 2013. Investigation of annotator's behaviour using eye-tracking data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. pages 214–222. http://www.aclweb.org/anthology/W13-2326.

Mitsuo Ikeda and Shinya Saida. 1978. Span of recognition in reading. *Vision Research* 18(1):83–88. https://doi.org/10.1016/0042-6989(78)90080-9.

Tomoya Iwakura. 2011. A named entity recognition method using rules acquired from unlabeled data. In *Recent Advances in Natural Language Processing, (RANLP 2011)*. pages 170–177.

Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*. pages 36–41.

Salil Joshi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2013. More than meets the eye: Study of human cognition in sense annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. pages 733–738.

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87(4):329–354.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1528–1533. http://www.aclweb.org/anthology/N16-1179.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. pages 1064–1074. http://www.aclweb.org/anthology/P16-1101.

Ryosuke Maki, Hitoshi Nishikawa, and Takenobu Tokunaga. 2016. Parameter estimation of japanese predicate argument structure analysis model using eye gaze information. In *Proceedings of the 26th International Conference on Computational Linguistics (Coling 2016)*. pages 2861–2869.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. pages 346–351.

Abhijit Mishra, Michael Carl, and Pushpak Bhattacharya. 2012. A heuristic-based approach for systematic error correction of gaze data for reading. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*. pages 71–80.

Koh Mitsuda, Ryu Iida, and Takenobu Tokunaga. 2013. Detecting missing annotation disagreement using eye gaze information. In *Proceedings of the 11th Workshop on Asian Language Resources*. pages 19–26.

Naoyuki Osaka. 1992. Size of saccade and fixation duration of eye movements during reading: Psychophysics of japanese text processing. *Journal of Optical Society of America* 9(1):5–13.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL 2014)*. pages 78–86.

Daniel C. Richardson, Rick Dale, and Michael J. Spivey. 2007. Eye movements in language and cognition: A brief introduction. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, and Michael J. Spivey, editors, *Methods in Cognitive Linguistics*, John Benjamins., pages 323–344.

Joe Cheri Ross, Abhijit Mishra, and Pushpak Bhattacharyya. 2016. Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*. pages 22–26.

Hassan Sajjad, Francisco Guzmán, Nadir Durrani, Ahmed Abdelali, Houda Bouamor, Irina Temnikova, and Stephan Vogel. 2016. Eyes don't lie: Predicting machine translation quality using eye movement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*. pages 1082–1088. http://www.aclweb.org/anthology/N16-1125.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications (ETRA '00)*. pages 71–78. https://doi.org/10.1145/355017.355028.

Ryohei Sasano and Sadao Kurohashi. 2008. Japanese named entity recognition using structural natural language processing. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*. pages 607–612. http://aclweb.org/anthology/I/I08/I08-2080.pdf.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. 2010. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. pages 1158–1167. http://www.aclweb.org/anthology/P10-1118.