Annotation for annotation
– Toward eliciting implicit linguistic knowledge through annotation –
(Project Note)

Tokunaga Takenobu **Iida Ryu** **Mitsuda Koh**
Department of Computer Science
Tokyo Institute of Technology
{take, ryu-i}@cl.cs.titech.ac.jp, mitsuda.k.aa@m.titech.ac.jp

Abstract

The last two decades witnessed a great success of revived empiricism in NLP research. However, there are still several NLP tasks that are not successful enough. As one of many directions for going beyond the revived empiricism, this paper introduces a project for annotating annotations with annotators' rationales behind them. As a first step of this enterprise, the paper particularly focuses on data collection during the annotation and discusses their potential uses. Finally a preliminary experiment for data collection is described with the data analysis.

1 Introduction

The last two decades witnessed a great success of revived empiricism in NLP research. Namely, the *corpus construction and machine learning* (CC-ML) approach has been the main stream of NLP research, where corpora are annotated for a specific task and then they are used as training data for machine learning (ML) techniques to build a system for the task.

The CC-ML approach has been expected to remedy the notorious knowledge construction bottleneck in traditional rule-based approaches. In the rule-based approach, given a specific task (e.g. POS tagging), human experts (e.g. computational linguists) create rules covering various linguistic phenomena based on their insight. In contrast, in the CC-ML approach, human experts mainly focus on creating annotation guidelines. According to the guidelines, annotation is usually performed by a number of annotators who do not necessarily have deep linguistic knowledge, aiming at increasing the corpus size. Resultant large annotated corpora are expected to cover broader linguistic phenomena in terms of a collection of annotation instances than the expert-constructed rules in a rule-based approach. Regularities corresponding to the rules are extracted from the annotated corpora by the ML techniques.

The primacy of the CC-ML approach over the rule-based approach has been shown in fundamental NLP tasks (e.g. POS tagging, syntactic parsing and word sense disambiguation) as well as in various applications (e.g. information extraction, machine translation and summarisation) through prevalent competition-type conferences. However, too much dominance of the revived empiricism has recently worried a number of researchers (Reiter, 2007; Steedman, 2008; Krahmer, 2010; Church, 2011). For instance, Church (2011), who is one of the initiators of the revived empiricism, warned us that we should follow the CC-ML approach with an awareness of the limitations of the underlying ML techniques.

One of the problems of the CC-ML approach is that the annotated information in the corpora is often limited to the output for a given specific task. Together with other clues (e.g. POS of surrounding words of a target), a system for the task is built by using ML techniques. However, the validity of these clues has been rarely examined deeply. This would be because the annotator's decision process has attracted less attention than the resultant annotations themselves. Therefore, there have been few attempts to systematically collect the annotator's rationales behind the annotation process.

From an engineering viewpoint, a machine does not need to perform a task in the same manner as a human does. The currently used clues might be sufficient for doing the job even though a human uses different information. For instance, POS tagging and parsing are successful instances of the CC-ML approach. However, this approach does not always work well on other tasks such as semantic and discourse processing. For instance, the performance of the state-of-the-art coreference resolution model still stays around 0.7 in F-score (Haghighi and Klein, 2010). Furthermore, the performance of zero anaphora resolution in Japanese is much worse, around 0.4 in F-score (Iida and Poesio, 2011). Such relatively low performance suggests that some crucial information should have been utilised for ML techniques.

Against this background, we propose annotating each annotation with the annotator's rationale behind her/his decision. Since the rationale explains the validity of the annotation instance, it can be considered as a kind of meta-level annotation against the object-level annotation rather than a mere attribute of the annotation. We expect that analysing these rationales behind human decisions reveals more effective information for a given task that has never been used by existing ML techniques. We believe this is one of the ways to integrate the revived empiricism and rationalism.

As a first step of this enterprise, this paper particularly discusses what kinds of information can be collected during the annotation process for estimating a rationale behind each annotation decision. We also explain potential uses of the collected information. Finally, we describe our preliminary experiment for collecting the annotator's actions and eye-gaze during her/his annotation process.

2 Data to collect

For the analysis of the annotator's rationales, we plan to collect two types of data, *overt* and *covert* data, which complement each other. The overt data are observable ones including the annotator's actions such as keystrokes, mouse clicks and dragging, and her/his eye-gaze. In contrast, the covert data reside in the annotator's mental process, i.e. her/his thoughts, which can not be observed directly.

Collecting overt data requires some specialised equipment. For collecting the annotator's actions we need annotation tools that can record every input. In addition, recent eye tracking devices enable us to capture an annotator's eye movement quite precisely at high frequency. This equipment enables us to capture the annotator's observable behaviour to a large extent. Such automatically collected low-level data can be further translated into more interpretable abstract actions and objects, which should be defined with respect to the annotation task. For instance, when annotating predicate-argument relations, mouse operations should be translated to meaningful actions like identifying text spans (e.g. words, phrases) corresponding to predicates and arguments, and identifying the relations between them. Likewise, the eye-gaze should be translated to the corresponding text span that the annotator looked at, and together with their time stamps they are further translated to fixations on the text spans. By analysing the temporal relations of these abstract actions, we can reveal what text spans the annotator looks at prior to establishing a predicate-argument relation. Such a prior glance at a certain text span is usually performed unconsciously but will be an important clue for analysing the annotator's decision process. Note that this sort of information is difficult to capture by TAP which is described below. In addtion, collecting overt data has the advantage that it does not interfere with the annotation process.

In order to collect verbalised rationales, which is difficult to draw out from the overt data, we could adopt the think-aloud protocol (TAP) (Ericsson and Simon, 1984), which enforces annotators to explain aloud their decision process. Although TAP makes the annotator's covert thoughts explicit, it increases her/his cognitive load, thus interferes with her/his natural annotation task. In order to decrease their cognitive load, we would make dyads to annotate corpora collaboratively and record their dialogues, expecting that we can extract clues of the annotators' decision process from their utterances. This method might be less effective to draw out their rationales for annotation than the TAP, but as dialogue is a natural act for collaboration, the annotators' cognitive load would be less than the TAP.

3 Potential uses of the collected data

The following are potential uses of the collected data.

Finding useful information for NLP

Given a certain task, the collected data would give some hints to understand human decision processes. Therefore, the information can be useful for replicating human decisions by using ML-based approaches.

Evaluating annotation quality

The quality of corpora is often evaluated based on the agreement ratio and the κ coefficient (Carletta, 1996) between multiple annotators. Analysing the collected data would help to estimate the realiability of each annotation instance. For instance, the long annotation time for an instance is an indication of its difficulty, therefore the annotation on such an instance might be less reliable.

Evaluating and training annotators

Unlike the quality of corpora, the quality of annotators is rarely discussed. In addition to the extent to which annotators can replicate the gold standard annotation (result-based metric), the collected data can be used for evaluating annotators by comparing their behaviours with that of expert annotators. This is a process-based metric, thus can be more informative for training annotators by identifying the differences of their behaviour.

4 Preliminary experiment – collecting actions and eye-gaze during annotation –

4.1 Materials

We conducted a preliminary experiment for collecting an annotator's actions and eye-gaze during her/his annotation of predicate-argument relations in Japanese texts. Given a text in which candidates of predicates and arguments were marked beforehand in the annotation tool, the annotators were instructed to add links between correct predicate-argument pairs by using the keyboard and mouse. We distinguished three types of links based on the case marker of arguments, i.e. *ga* (nominative), *o* (accusative) and *ni* (dative). For elliptical arguments of a predicate, which are quite common in Japanese texts, their antecedents were linked to the predicate. Since the candidates were marked based on the automatic output of a parser, some candidates did not have their counterparts.

We recruited three annotators who had experiences of annotating predicate-argument relations. Each annotator was assigned 43 texts for annotation, which were the same across the annotators. These 43 texts were selected from a Japanese balanced corpus (BCCWJ) (Maekawa et al., 2010). Considering capturing eye-gaze, we prohibited scrolling a text during annotation. Thus, the texts were truncated to about 1,000 characters so that they fit into the text area of the annotation tool.

We employed a multi-purpose annotation tool *Slate* (Kaplan et al., 2012) with necessary modifications, particularly by implementing a logging function for capturing an annotator's input.

Annotator's gaze 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 the annotator's eye was maintained at about 50 cm. In order to minimise the head movement, we used a chin rest.

4.2 Agreement ratio and annotation time

The number of annotated links between predicates and arguments by three annotators A_0, A_1 and A_2 were 3,353 (A_0), 3,764 (A_1) and 3,462 (A_2) respectively. There were several cases where the annotator added multiple links with the same link type to a predicate, e.g. in case of conjunctive arguments; we excluded these instances for simplicity in the analysis below. The number of the remaining links were 3,054 (A_0), 3,251 (A_1) and 2,996 (A_2) respectively. These annotation instances were used for analysing the relation between the agreement ratio and annotation time of the annotated links.

Having fixed a predicate and link type (case marker) pair, we considered the extent to which the annotators agreed on its argument. Among 2,209 predicate and link type pairs that all three annotators annotated, the three agreed in 1,952 pairs. Thus, the agreement ratio of the annotation among three was 0.884. When allowing agreement by only two annotators, the average of pairwise agreement ratios increased to 0.902.

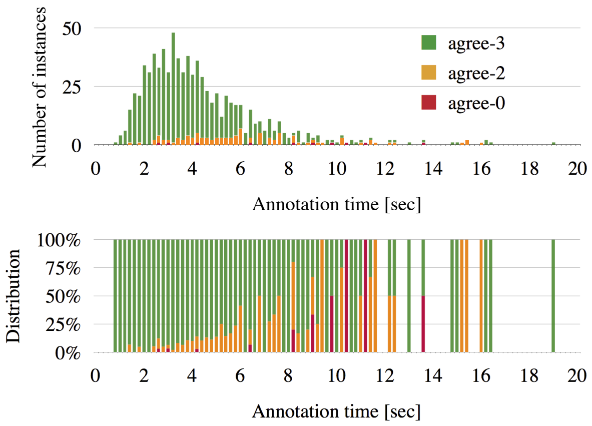

Figure 1: Relation between annotation time and agreement

The annotation time for a link is defined as the time duration from the onset of the first gaze fixation on the predicate after the previous link establishment until the establishment of the current link. The fixation is detected by the Dispersion-Threshold Identification (I-DT) algorithm (Salvucci and Goldberg, 2000) with the space margin of 16 pixels and the time window of 100 msec.

Figure 1 (upper) shows the relation between the average annotation time of the annotators (x-axis with interval width in 0.5 sec) and the number of agreed instances (y-axis). The agreed instances are further divided into three cases: agreed by three (agree-3), agreed by two (agree-2) and no agreement (agree-0). Figure 1 (lower) shows the distribution of the degree of agreement in each interval. The figures indicate that taking longer annotation time suggests difficulty of the annotation instance, thus its low reliability. This tendency indicates that we would be able to estimate the reliability of annotations without the gold standard nor any counterpart for calculating agreement metrics.

5 Related work

We here focus on related work utilising eye-tracking data. As for the analysis of dialogue data, numerous studies on dialogue research could be useful.

Recent development of the eye-tracking technology enables various research fields to employ eye-gaze data, including psycholinguistics and problem solving (Duchowski, 2002). There have been a number of studies on the relations between eye-gaze and language comprehension/production (Griffin and Bock, 2000; Richardson et al., 2007). Compared to the studies on language and eye-gaze, the role of gaze in general problem solving settings has been less studied (Bednarik and Tukiainen, 2008; Rosengrant, 2010; Tomanek et al., 2010). Since our current interest, corpus annotation, can be considered as a problem solving as well as language comprehension, various existing metrics derived from eye-tracking data would be useful.

Rosengrant (2010) proposed an analysis method named *gaze scribing* where eye-tracking data is combined with subjects thought process derived by the TAP, underlining the importance of applying gaze scribing to various problem solving.

Tomanek et al. (2010) utilised eye-tracking data to evaluate difficulties of named entities for selecting training instances for active learning techniques. Our analysis in the previous section is similar to theirs in that the annotator's gaze is used for estimating the annotation difficulty. However, our annotation task is more complex (named entity recognition vs. predicate-argument relations), and in a more natural setting, meaning that all possible relations in a text were annotated in a single session in our setting, while each session targeted a single named entity (NE) instance in a limited context in the setting of Tomanek et al. (2010). Due to such a more realistic setting, the definition of the annotation time is not obvious in our case. Furthermore our fixation target is more precise, i.e. words, rather than a coarse area around the target NE.

6 Concluding remarks

This paper proposed annotating annotations with the annotator's rationales during the annotation process. We particularly discussed overt and covert data collection during the annotation and potential uses of the collected data. Results of a preliminary data collection and the data analysis were also shown. The project has just started. We need to collect more data, both overt and covert, and to establish a method to explore the human annotation process by analysing the interaction between both kinds of data. We believe the direction proposed in the present paper is one of the ways for going beyond the revived empiricism in the study of language processing.

References

Roman Bednarik and Markku Tukiainen. 2008. Temporal eye-tracking data: Evolution of debugging strategies with multiple representations. In *Proceedings of the 2008 symposium on Eye tracking research & applications (ETRA '08)*, pages 99–102.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Kenneth Church. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology*, 6(5):1–27.

Andrew T. Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, and Computers*, 34(4):455–470.

K. Anders Ericsson and Herbert A. Simon. 1984. *Protocol Analysis – Verbal Reports as Data –*. The MIT Press.

Zenzi M. Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11(4):274–279.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 804–813.

Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. 2012. Slate – a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, 26(2):89–101.

Emiel Krahmer. 2010. What computational linguists can learn from psychologists (and vice versa). *Computational Linguistics*, 36(2):285–294.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1483–1486.

Ehud Reiter. 2007. The shrinking horizons of computational linguistics. *Computational Linguistics*, 33(2):283–287.

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*, pages 323–344. John Benjamins.

David Rosengrant. 2010. Gaze scribing in physics problem solving. In *Proceedings of the 2010 symposium on Eye tracking research & applications (ETRA '10)*, pages 45–48.

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.

Mark Steedman. 2008. On becoming a discipline. *Computational Linguistics*, 34(1):137–144.

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.