About corpus annotation

Yumeto Inaoka

Feb. 23, 2018

1 Consistency

In natural language processing, many tagged corpora were created by research groups. These annotation criteria are various because these aims are different. However, we should aim for consistent annotation in datasets of the same corpus.

Many annotation criteria cannot be written by explicit rules. Annotations whose criteria are not clear are annotated manually by each annotator's judgment. It is impossible to achieve 100% accuracy using datasets containing annotation variations. Therefore, the ratio of the inter-annotator agreement is used as the upper-performance limit.

The ratio of the inter-annotator agreement is shown by the following equation:

$$\kappa = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e} \tag{1}$$

where P_o is the ratio of a match in two annotators, P_e is the probability that two annotators' judgments are coincidence, and κ is called Cohen's kappa.

2 Criteria

When using annotations in natural language processing, it should be read automatically by the program. Therefore, it is necessary that annotations comply criteria. Furthermore, corpora with the same type of annotation should follow the same criteria. This is a problem of compatibility between different corpus annotations. The compatibility of annotation is separated into formal compatibility and semantically one. For example, when XML is used in the annotation format, it is formally compatible and can be read using any XML parser library. However, only formal compatibility is not enough. In XML format, it is necessary to define what kind of XML tags is used, what the attribute name and type are, and what kind of tags can be nested. It defined by DTD (Document Type Definition), Schema, and so on. If different tag names are used for the same kind of annotation, it is necessary to be compatible by converting. Furthermore, even if the same tag names are used, their meanings are not always the same. For example, even if the same "morpheme" tag name is used, it is not always the same at the conceptual level of "what is a morpheme." The process of making annotations semantically compatible will cause a loss of information.

TEI (Text Encoding Initiative) provides international guidelines for text annotation. However, they are mainly defined from the standpoint of humanities. Therefore, they are not widely used in natural language processing. In reality, the criteria depend on the corpus developer.

References

[1] Kikuo Maekawa, Yuji Matsumoto and Manabu Okumura (2017). コーパスと自然言語処理（講座日本語コーパス）. 朝倉書店.