Enhancing Quality of Corpus Annotation: Construction of the Multi-Layer Corpus Annotation and Simplified Validation of the Corpus Annotation

Youngbin Noh[1], Kuntae Kim[1], Minho Lee[1], Cheolhun Heo[1], Yongbin Jeong[1], Yoosung Jeong[1], Younggyun Hahm[1], Taehwan Oh[2],Hyonsu Choe[2], Seokwon Park[2], Jin-Dong Kim[3] and Key-Sun Choi[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Yonsei University, South Korea, [3]Database Center for Life Science, Japan
{vincenoh, kuntaek, pathmaker, fairy_of_9, kuonom, wjd1004109, hahmyg, kschoi}@kaist.ac.kr
{ghksl0604, choehyonsu, sseokhon}@yonsei.ac.kr
jdkim@dbcls.rois.ac.jp

Abstract

In this paper, we construct simultaneously multi-layer corpus annotation of 7 linguistic layers as a gold set. And we design the validation procedure using the gold set and report the results of a validation procedure for other large-scale corpus annotation of 7 linguistic layers. Furthermore, we present a simplified validation method without a gold set using annotation models learned by the gold set. Based on the validation results, the tendency of annotation across the entire corpus is observed macroscopically, and the corpus annotation validation results are analyzed microscopically to verify the validation methodology to address the case study.

1 Introduction

As a resource for natural language processing, the corpus is annotated with additional information for various purposes. To annotate such various information to the raw corpus, corpus annotation project must be designed elaborately considering the requirements of the annotation procedure, annotation units, annotation tools, human annotators, and so on. A reliable annotation design makes the corpus annotation better quality(Finlayson and Erjavec, 2017). Also, the design suitability of the corpus annotation project needs to be proved empirically, so the design of the corpus annotation project must be revised and supplemented iteratively(Pustejovsky and Stubbs, 2012).

In this paper, we construct 7 linguistic layers[1](Ide, 2017) of multi-layered corpus annotation as gold set (210K *Eojeol*[2]s) to validate large-scale corpus annotation by the 7 layers as evaluation set (3M *Eojeol*s). The gold set is annotated on the subset of the raw corpus of the evaluation set. The annotator groups of gold set by each layer, who annotated gold set assisted by auto-labeling are groups of experts separated from the annotator groups of evaluation set.

We have designed and applied a corpus annotation method that uses the simple inter-annotator agreement to construct the gold sets at the same time under limited time and human resources. To do this, we assigned one annotation unit to two annotators. According to annotation results from two annotators, we conducted the cross-checking process to determine the final annotation result.

In this paper, validation of evaluation sets by layers is performed by comparing two corpus annotations (gold set - evaluation set)[3] using the gold set as a criterion. Comparative validation of the two corpus annotations using a gold set can only be performed on a limited part of the evaluation set sharing the same part of raw corpus to be annotated.

[1]We construct and evaluate the 7 linguistic layer corpus annotation: morphological analysis, lexical sense analysis, named entity analysis, subject anaphora resolution, co-reference resolution, dependency analysis, and semantic roles analysis. The evaluation sets are also constructed by the same layers.

[2]In Korean, the word segment divided by white space is called "*Eojeol*", this is composed of a noun or verb stem combined with a postposition ("*Josa*") or ending ("*Eomi*") that function as inflectional and derivational particles. (Noh et al., 2018)

[3]In this project, we are constructed 7 linguistic layers of corpus annotations as gold set to validate 7 linguistic layers of corpus annotation (evaluation set) constructed by other project groups. The evaluation sets after validation can be downloaded at https://corpus.korean.go.kr/.

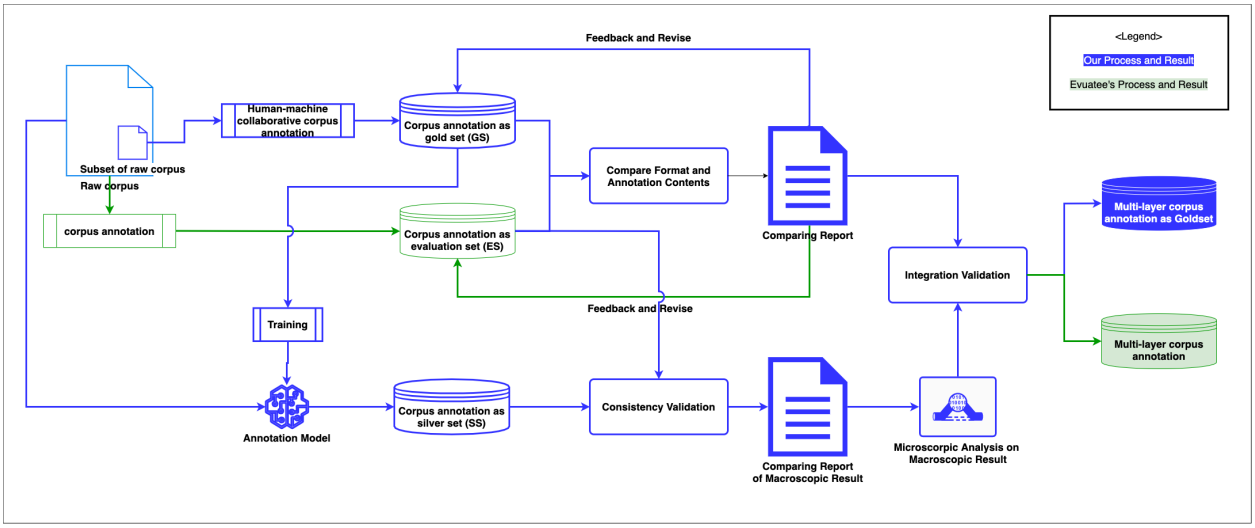

Figure 1: The flow of the corpus annotation and validation process in this paper. Blue-coloureds indicate our process and result, and green-coloureds indicate evaluatee group's process and result.

That's why we present an additional method to validate in the range of evaluation set without a gold set. Thus, we constructed another corpus annotation (silver set) to observe the consistency of annotation across the entire evaluation set by annotating an annotation model using the gold set of each layer as training data. The silver set is compared with the evaluation set to observe the tendency of annotation across the entire evaluation set, and a data-driven statistical analysis is performed to evaluate the appropriateness of the validation results.

In following section 2, we introduce the related works to the design and annotation of corpus annotation projects and corpus validation. In section 3, we introduce the design and annotation process for the construction of the gold set performed in this study. In section 4, we address the validation process for the format and annotation contents of the corpus annotation with an evaluation set using a gold set and validate the annotation consistency of the evaluation set without a gold set. In sections 5 and 6, we report the validation results in section 4 and issues related to these results.

2 Related Works

The corpus annotation project should be modeled according to the goal of the project to reflect appropriate specifications. This is because the corpus annotation, which is a result of corpus annotation, is learning data for the machine learning algorithm to learn specific phenomena that are not only linguistic but also non-linguistic. Therefore, the corpus annotation project design needs to appropriately reflect the characteristics of the phenomena to be learned in the machine learning algorithm. A broadly used framework is the MATTER cycle(Pustejovsky and Stubbs, 2012). The MATTER cycle is a cyclical process in which a corpus annotation project design ensures that the corpus annotation produces machine learning results that are appropriate for the goal of the corpus annotation project.

The OntoNotes(Hovy et al., 2006; Weischedel et al., 2011; Weischedel et al., 2013) is a multi-layer corpus annotation constructed over six years of texts in various genres in three languages (English, Chinese, and Arabic). It is a multilingual, multi-layer corpus annotation that annotated the structural information of the text as well as the semantic information to understand the meaning of the context based on a syntactic structure derived from Penn Treebank corpus, and the predicate-argument structure derived from Penn PropBank. It includes annotations such as word sense disambiguation for verbs and nouns, entity names annotation, the ontology of each word, and coreference relations. OntoNotes had tried to secure at least 90% of inter-annotator agreement in each corpus annotation, improving the quality of corpus annotation.

		Gold Set (Ours)	Evaluation Set	The OntoNotes 5.0 (more details in (Weischedel et al., 2013))
Language		Korean		English, Chinese, Arabic
Domain of Raw Corpus	**written**	Newspaper		newswire, broadcast news, broadcast conversation and, web data in English and Chinese, a pivot corpus in English (Old and New Testaments and Arabic (Newswire)
	Spoken	Transcripts of recording files (public conversation, public monologue, private conversation)		
Linguistic layer	**written**	Morphological analysis, lexical sense analysis, named entity analysis, subject zero anaphora resolution, coreference resolution, dependency analysis, semantic role analysis		Penn Treebank, Penn PropBank, word sence disambiguation for nouns and verbs, word senses connected to an ontology, and coreference
	Spoken	Morphological analysis, lexical sense analysis, named entity analysis, subject zero anaphora resolution, coreference resolution		
Quantity	**written**	140K *Eojeol*s	2M *Eojeol*s	2.9M words
	spoken	70K *Eojeol*s	1M *Eojeol*s	
Annotator groups		7 annotator groups by the layers different from evaluatee groups	7 annotator groups by the layers	
Constructing time		about 6 months per entire gold set (7 layers)	about 4-6 month per an evaluation set	about 6 years released to The OntoNotes 5.0 from The OntoNotes1.0

Table 1: Comparison of corpus specification of gold set, evaluation set and The OntoNotes 5.0. *Eojeol* is a unit of word segmentation of Korean.

In NLP area, various evaluation and annotation methodologies have been used to enhance and manage corpus quality as a natural language processing resource. As a corpus annotation quality control methods, inter-annotator agreement(Pustejovsky and Stubbs, 2012) has been generally used to control the result of corpus annotation. Checking inter-annotator agreement among annotators is widely used not only for evaluating the results of annotations from an assigned group of annotators, but also for evaluating the quality of data collected from an unspecified number of annotator, such as crowdsourcing methodology(Nowak and Rüger, 2010; Dumitrache, 2015; Dumitrache et al., 2018; Poesio et al., 2019).

3 Construction of Corpus Annotation

In this section, we address the overall procedures of constructing corpus annotation as a gold set. The gold set and the evaluation set share a raw corpus, and the gold set is a corpus annotation of 7 layers constructed by sampling 7% of the raw corpus. Therefore, in this paper, the corpus annotation of 7 layers is simultaneously constructed to build multi-layer corpus annotation.

3.1 Annotation Specification

To establish the corpus annotation guidelines by layers, the existing corpus annotation guidelines are revised and used according to the project purpose[4]. To make sure that the revised guidelines are not ambiguous or lacking in reflecting actual linguistic phenomena, three different annotator groups[5] constructed sample corpus annotation layer by layer on the same range of raw corpus. These sample corpus annotation also assessed the completeness of the corpus annotation guidelines, but were used as an annotation example in the annotation process. Through this process, it is possible to supplement

[4]The annotation guidelines referenced in this project refer to the annotation guideline for each layer from the 21st century Sejong project (morphological analysis, lexical sense analysis), and the guidelines made by the Electronics and Telecommunications Research Institute (ETRI; named entities analysis, subject zero anaphora resolution, co-reference resolution, and semantic role analysis) and Telecommunications Technology Association (TTA; dependency analysis). These guidelines do not refer to literature information in this paper, because it also includes non-public materials. For inquiries about these guidelines, please contact NIKL, ETRI, TTA.

[5]The annotation results of the three groups - ours, evaluatee groups, and expert group of National Institute of Korean Language (NIKL) annotated on the same range of raw corpus on the seven layers. The disagreement among the corpus annotation results of the three groups was decided by the NIKL expert group and the annotation guidelines were reflected in the results of the decision by NIKL.

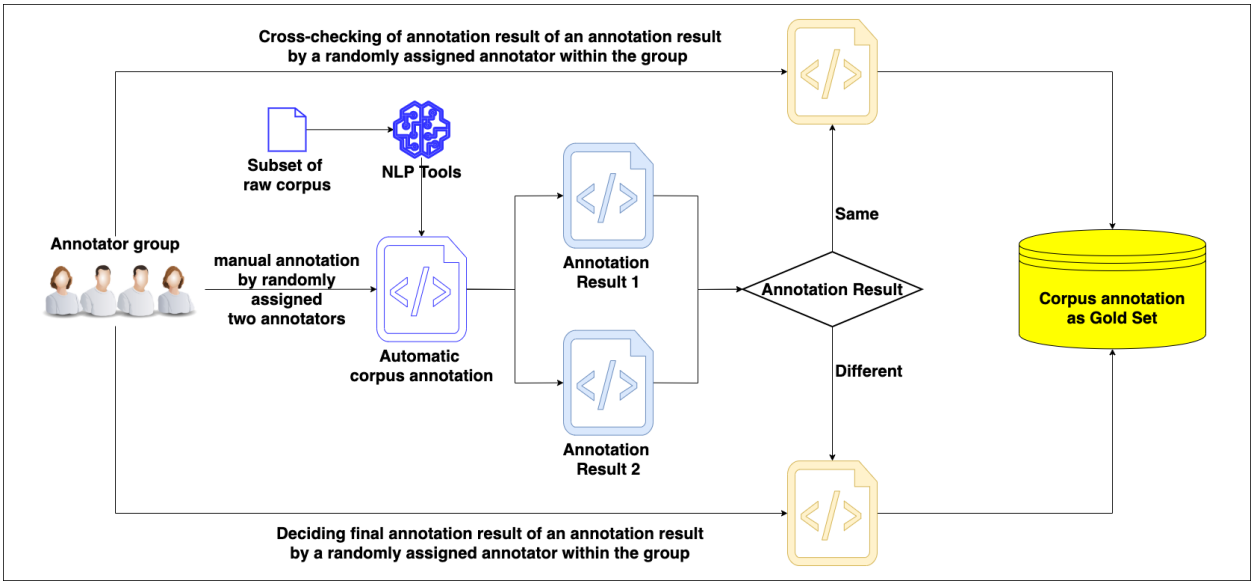

Figure 2: The flow of construction procedure. This figure is a detailed representation of the part of the Human-machine collaborative corpus annotation in Figure 1.

the process of evaluating the repetitive annotation scheme through this process.

In this project, the seven annotator groups[6] were recruited by layers, consisting of experts in Korean or linguistics, who can fully understand annotation guideline and apply this to corpus annotation. They studied and analyzed the existing guidelines and constructed sample corpus annotation. And they annotated the gold sets using the revised guidelines reflecting the results of the annotation of the sample corpus annotation. To ensure validation results of corpus annotation between two annotator groups, they were performed corpus annotation independently not only within their group but also from the annotator group of the evaluatee group. Also, they are completely separated from the annotation group of the evaluatee group and are independently annotated to provide conditions for evaluating the results of the annotation by inter-annotator agreement.

3.2 Annotation Environments

As mentioned above, the annotators of our project annotated independently of each other. For the annotators to be separated, a workbench with a personal workspace is required. We also designed a web-based annotation environment to reduce the time and location constraints of annotation work.

This project used a web-based workbench. This annotation environment was designed for this project and was developed to reflects the annotation schemes of the seven layers. This workbench could only be accessed by annotators and administrator designated for each layer. Multiple annotators can annotate one annotation unit, and it is possible to grasp the annotator's annotation status on an online browser.

Annotators can use only the set of tag sets defined in the annotation schemes, and the annotation contents can also annotate only those in the annotation schemes. If the annotation guidelines have updated and annotators need to add new annotation content, they can ask the administrator to remove the restrictions for annotation contents. Also, the annotators annotated referring the results of the automatic preprocessing of the raw corpus by layers.

This workbench is equipped with a function of randomly assigning a defined annotation unit to an annotator. This allows an annotator not to annotate the entire document at once, but to annotate parts so that one document is annotated by multiple annotators.

[6]The annotators by layers is a group of experts with a master's degree or Ph.D. of the Korean language or linguistics. They were qualified as annotators by NIKL before corpus annotation.

3.3 Corpus Annotation Procedure

In this paper, corpus annotation as a gold set was constructed using a human-machine collaborative annotation method (Figure 2). First, an automatic corpus annotation is automatically annotated using NLP Tools[7] to the subset of the raw corpus. This includes the annotation results of the Korean morphemes, dependency parsing, named entity recognition, and semantic role labeling (recognizing predicate and argument), and so on.

Next, the annotator in the annotator group manually re-annotates the annotation unit by referring to the annotation result of the NLP tool. One annotation unit is annotated by randomly assigned two annotators in the group. If the annotation results of an annotation unit by two annotators are same, these annotation units are cross-checked by a randomly assigned annotator within the group. Else if the annotation results of an annotation unit by randomly assigned two annotators are different, these annotation units are decided final annotation result by a randomly assigned annotator within the group. Through these processes, an annotation unit is checked by annotators at least two times. After two parallel processes, those results of two-stage are made corpus annotation gold set.

4 Validation of Corpus Annotation

After constructing the gold sets, using these corpora, we validate the evaluation sets. The validation of the evaluation sets validates the format of corpus annotation, the annotation contents, and the annotation consistency. After that, the integration validation is performed to be used the evaluation sets as a multi-layer corpus annotation (refer in Figure 1).

4.1 Format Validation

The format validation is a process of confirming whether the corpus has been constructed by the defined annotation format, and also is a stage of confirming whether the corpus can be used as electronic data. At this stage, corpus annotation is validated about the standard format and data structure. In addition to the corpus format, in this stage, it is checked whether or not a label defined for each layer is used, and whether other content is included in addition to the specified annotation content. When a format error is detected in this stage, the evaluation set does not proceed to the annotation contents validation stage, and correction and supplementation are required.

4.2 Annotation Contents Validation

The annotation contents validation performs data-oriented validation that compares the gold set and evaluation set. At this validation, we compare the annotation contents of the two corpora and report the different annotation content to the evaluatee group. The annotation content validation items are selected based on the annotation guidelines for each layer, and the annotation contents defined in the layer are selected as validation items and shared in the evaluatee groups. Based on this report, the evaluatee groups can modify and supplement their corpus annotation. This process is a method of evaluating based on the inter-annotator agreement between annotation groups, and it is judged that the correct annotation is performed when the annotation of the two groups matched.

4.3 Consistency Validation for Evaluation Set using Silver Set

The annotation consistency of the evaluation set is evaluated indirectly by confirming that the tendency of the annotation of the gold set is similar to the evaluation set. To do this, we create an automatic annotation model for each layer using a gold set as training data and construct an automatic corpus annotation (silver set) that annotates automatic annotation on a raw corpus in range of without a gold set. By comparing the silver set and evaluation set, the consistency of the annotation content is analyzed to evaluate the annotation consistency.

To validate the annotation consistency of the evaluation set, we were divided into 10 sections to analyze the tendency of the agreement between the silver set and evaluation set. The average of the agreement of corpus annotations between two corpora in 10 sections was averaged, and when the observed agreement rate of each section deviated from the 99% confidence interval ($\alpha = 0.01$) compared to the

[7]We were supported by the Exobrain Korean Language Analysis Toolkit v3.0 developed by the Electronics and Telecommunications Research Institute (ETRI) to automatically annotate the raw corpus.

mean value, the corresponding section was evaluated to have relatively lower annotation consistency than other sections.

4.4 Integration Validation

As a final stage in constructing a multi-layer corpus annotation of seven layers, it is necessary to check whether the raw corpus of evaluation sets are preserved and whether the annotation schemes have been observed. To make a multi-layer corpus annotation by combining the seven sets, we compared the statistics of the number of documents, paragraphs, sentences, and *Eojeol*s in each corpus, and confirm that the defined ID assignment rules are followed.

5 Results

5.1 Annotation Contents Validation

Table 2 shows the results of annotation agreement between the gold set and the evaluation set. Annotation contents validation of morphological corpus annotation was validated to match the *Eojeol* segment and morphological label annotation. In Korean *Eojeol*, morphemes such as stems (*Eogan*), postposition (*josa*), ending (*Eomi*) are combined to form a single *Eojeol*. Thus, to analyze morphemes, it is necessary to check whether the *Eojeol* segmentation is same and whether the same label is annotated to the segmented morpheme. Written corpus annotation was relatively consistent with both segmentation and label annotation in spoken corpus annotation. Segmenting concordance was lower than label annotation concordance, indicating that there was a difference in the morpheme semantic analysis of *Eojeol* between annotators.

When comparing the annotation agreement between written corpus annotation and spoken corpus annotation, the tendency of the written corpus annotation shows a higher annotation agreement than spoken corpus annotation. It is because a spoken raw corpus transcribed public monologues (news), public conversation (broadcast conversation, interview, lecture, and so on), and private conversation recording. In the case of public monologue, it was refined to a level similar to that of the written raw corpus with well-refined text. In the case of private conversations or broadcast interviews, however, many features of spoken language (i.e., speech break, blur, reduction, slang, and so on) appeared, making it difficult for annotators to analyze text.

Layer	Validation contents	Measures	Written	Spoken
MP	*Eojeol* segmentation	Accuracy	98.6	93.84
	Morpheme label	F1	99.22	97.84
LS	Lexical sense ID	F1	92.47	92.49
NE	Named entity annotation	F1	86.02	94.48
ZA	Predicate Identification (PI)	F1	88.93	88.48
	Subject anaphora resolution	Accuracy	79.20	65.71
CR	MUC	F1	68.20	59.44
DP	Dependency head and label	LAS	87.45	n/a
SR	Predicate Identification (PI)	F1	87.82	n/a
	Argument Identification (AI)	F1	73.86	n/a

Table 2: Results of annotation contents validation.

5.2 Consistency Validation

Table 3 shows the results of annotation consistency validation. Annotation consistency validation was performed separately for written and spoken corpus annotation. To validate the consistency of the annotation, some of the measures used to validate the annotation contents for each layer were used, and the consistency was evaluated through the consistency of the indicator. The 99% confidence interval compared to the average value of the annotated agreement of the silver set and evaluation set for each section was shorter than that of the majority of written corpus annotation (Avg. of CI length: Written = 3.603, Spoken = 3.193). In the case of written corpus annotation, sections 1 and 2 of ZA corpus annotation showed a markedly low agreement, affecting the average CI length of ZA increase.

The corpus showing the shortest 99% confidence interval is the named entity annotated written corpus with the smallest difference in the annotation content with the silver set of 10 sections (99% CI length = .422 ($21.98 \leq CI \leq 22.430$), confidence($\alpha = 0.01$)= 0.202, σ=0.248). Also, one section out of the 99% confidence interval was analyzed, and it was evaluated that the consistency of the annotation of the named entity written corpus annotation was relatively higher than that of the other corpora.

Compared to the annotation content validation result (Table 2), when the gold set and the evaluation set match 80% or more (MP (Written, Spoken), LS (Written, Spoken), NE (Written, Spoken), ZA (Written, Spoken), DP, SR), the length of the 99% confidence interval, except for the ZA written and MP spoken corpus annotation, is all 1.5 or less. There-

Layer		Measures	1	2	3	4	5	6	7	8	9	10	Avg.	CI Length
MP	**Written**	**Accuracy**	**85.87**	**85.87**	86.38	86.09	86.51	**86.60**	86.44	86.33	86.52	86.49	86.298	0.533
MP	**Spoken**	**Accuracy**	**75.35**	**73.75**	76.98	79.30	78.28	79.24	77.40	78.99	**80.16**	79.32	77.877	3.681
LS	**Written**	**Accuracy**	87.64	**86.81**	87.4	87.45	87.81	**87.99**	87.47	**87.95**	87.41	**88.14**	87.607	0.703
LS	**Spoken**	**Accuracy**	**77.72**	79.31	79.51	79.86	78.89	79.2	79.66	**78.37**	79.09	**80.75**	79.236	1.498
NE	**Written**	**Accuracy**	22.04	22.10	21.99	22.58	22.1	22.18	22.26	**21.97**	22.37	22.69	22.228	0.449
NE	**Spoken**	**Accuracy**	**27.45**	20.03	20.10	20.54	21.62	20.55	22.54	**19.15**	22.06	21.20	21.524	4.206
ZA	**Written**	**F1**	**13.98**	**8.55**	20.84	**38.75**	29.3	33.35	31.41	33.31	22.82	31.34	26.365	17.353
ZA	**Spoken**	**F1**	23.70	**22.78**	**22.72**	**24.01**	23.71	**24.08**	22.80	23.07	22.81	**24.02**	23.370	1.058
CR	**Written**	**F1 (MUC)**	**51.32**	50.08	50.34	**49.66**	51.13	**51.23**	**51.36**	50.91	**49.77**	51.09	50.689	1.201
CR	**Spoken**	**F1 (MUC)**	36.39	33.93	33.71	**38.42**	**42.07**	33.57	**38.25**	34.20	33.89	**32.34**	35.677	5.524
DP	**Written**	**UAS**	68.37	68.77	68.11	68.68	68.43	69.00	68.92	69.03	**69.73**	**67.18**	68.622	1.224
SR	**Spoken**	**F1 (AI)**	**61.91**	61.30	61.14	61.34	61.06	61.10	**60.78**	61.15	61.25	61.18	61.221	0.522

Table 3: Results of consistency validation. The bold indicates the agreement rate between the silver set and evaluation set outside the 99% confidence interval. The following models were used for the automatic annotation model for annotation consistency validation: (Ma and Hovy, 2016) - MP (Written), (Joshi et al., 2019) - ZA (Written, Spoken), CR (Written, Spoken), (Straka et al., 2016) - DP (Written), (Lee et al., 2015; Bae et al., 2017) - SR (Written). The annotation models of layers that have no reference was developed and used as a statistical-based learning model.

fore, it was evaluated that it showed high annotation consistency.

The annotation consistency validation by creating a learning model using a gold set started from the hypothesis that the balanced composition of the raw corpus represents a language phenomenon. In the case of the written raw corpus, it is constructed only newspaper articles, so there is relatively little bias in language phenomena according to genres or domains of text. Therefore, it can be said that the written corpus represents more representative of the language phenomenon than the spoken corpus composed of public dialogue, public monologue, and private dialogue.

The quality of the silver set automatically annotated to a well-balanced corpus does not significantly affect the result of the annotation consistency validation. The goal of annotation consistency validation is to verify that the evaluation set shows the annotation characteristics of the annotation model learned by the gold set. As long as it is annotated silver set as a model that properly trains the gold set, it is evaluated that it does not matter if the agreement rate between the silver set and the valuation set is low. However, when comparing the measured value of the agreement divided into 10 sections with the average value, by setting the confidence interval of the average value to 99% (confidence $\alpha = 0.01$), the evaluation standard of annotation consistency for each section was generously set. Also, if the section is further subdivided into 10 or more, more accurate annotation consistency validation will be possible.

As an example, when the results of semantic role corpus annotation annotation consistency validation were analyzed in detail, each section showed a maximum difference of 0.69 from the average (in Table 3, $60.933 \leq CI \leq 61.455$, $60.78 \leq AgreementRate \leq 61.91$). Although it was recorded that it was out of the 99% confidence interval in two sections, the length of the CI was 0.522, which was shorter after the written named entity corpus annotation, and could be evaluated as showing stable annotation consistency. Also, the agreement between the silver set and the evaluation set constructed with the automatic annotation model trained with the gold set is 61.221, but when comparing the sample annotation corpus and the silver set, the consistency was 66.43. It could be judge indirectly as having no significant effect on annotation consistency.

5.3 Case study

A typical inconsistency was mis-annotation on exceptions (Table 4). In semantic role annotation, it consists of the cases on the adverbs of Korean that share a root with a specific verb, auxiliary verbs that composes a predicate in combination with the main verb, and verbs that are a part of periphrastic constructions or tagmeme equivalents. Some Korean verbs function as a marker of aspect, modal, and negation in predication or used as an element to form

	imi already	*jinan* has passed	*ile* thing	*daehae* about	*geuleohge* in that way	*malhaneun* saying	*geoseun* is	*olhji* right	*anhda.* not
Gold				ARG1	ARGM-MNR	*malha*.01	ARG1	*olh*.01	ARGM-NEG
E1		*jinan*.01	ARG1						
E2			ARG2	*daeha*.01					
E3					*geuleoh*.01				
E4								ARG1	*anh*.01

Table 4: SR example of mis-annotation on a sentence which means 'It is not right to say so about what has already passed.'

multi-word periphrastic construction. In particular, the verbs included in the periphrastic constructions are characterized by: 1) do not affect the content of the proposition, only play grammatical functions, 2) the use is morphologically fixed, and 3) it cannot form a sentential predicate or is not related to the argument structure. E1 is a disagreement caused by the annotator mistaken complementation for relativization, E2 is a disagreement on multi-word periphrastic construction *'-e daeha-'* (about/on that), E3 is a disagreement on adverbs that share a root with a specific verb, E4 is a disagreement on an auxiliary verb for negation.

6 Discussions and Conclusions

In this paper, we propose and implement a methodology for constructing language resources for NLP tasks quickly and efficiently for goals of annotation project, but also try to achieve an appropriate level of corpus annotation quality assurance.

We designed the constructing process of gold set to consider agreement within annotators when the results from two annotators for one annotation unit match or not, the annotation contents in the annotator group were once again annotated so that the annotation contents were cross-checked and confirmed. This method is a simple and reliable method to check the difference in the subjectivity of the annotator in a short time. In particular, because Korean has the properties of agglutinative language, it has the possibility that a single annotation unit can be interpreted in multiple meanings, so it is necessary to carefully consider the context information surrounding the annotation units. Even though the annotators have annotated deliberately in compliance with the annotation guideline, there are many possibilities for annotation due to semantic diversity, ambiguity, or subjectivity of annotators.

To construct a gold set in a short time and use it to validate the evaluation set, the quality and authority of the gold set are always important. That is why we designed the process of determining the appropriate annotation minimizing annotation bias while comparing the annotation results within or among annotator groups. Furthermore, we use the gold set as training data of the annotation model to annotate the silver set. If this method is more elaborately, it could be an alternative to evaluating the quality of a corpus annotation when all gold set corresponding to the evaluation set could not be made.

It is difficult for everyone to interpret identically the same linguistic phenomenon due to environmental or individual aspect. Also, it is hard to say that the gold set is always correct. Therefore, this paper tried to aim to present a method to reduce individual and group bias when constructing corpus annotation. In future research, we try to further generalize the methodology for constructing and validating corpus annotation.

Acknowledgments

This work is written based on the results of the 'Corpus Integration Validation'(NIKL 2019-01-61) project of the National Institute of Korean Language.

This work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01780, The technology development for event recognition/relational reason-

ing and learning knowledge based system for video understanding)

This work was supported by Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) [2016-0-00562(R0124-16-0002), Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly]

References

Jangseong Bae, Changki Lee, and Hyunki Kim. 2017. Korean semantic role labeling with highway bilstm-crfs. In *Annual Conference on Human and Language Technology*, pages 159–162. Human and Language Technology.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing semantic label propagation in relation classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium, November. Association for Computational Linguistics.

Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *European Semantic Web Conference*, pages 701–710. Springer.

Mark A Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Nancy Ide. 2017. Introduction: The handbook of linguistic annotation. In *Handbook of Linguistic Annotation*, pages 1–18. Springer.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Changki Lee, Soojong Lim, and Hyunki Kim. 2015. Korean semantic role labeling using structured svm. *Journal of KIISE*, 42(2):220–226.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Youngbin Noh, Jiyoon Han, Tae Hwan Oh, and Hansaem Kim. 2018. Enhancing universal dependencies for korean. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 108–116.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.*

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.