Using Text Classification to Improve Annotation Quality by Improving Annotator Consistency

Ishita, Emi
Kyushu University

Fukuda, Satoshi
Kyushu University

Tomiura, Yoichi
Kyushu University

Oard, Douglas W.
University of Maryland

https://hdl.handle.net/2324/4479586

出版情報：Proceedings of the Association for Information Science and Technology. 57 (1), pp.e301-, 2020-10-22. Association for Information Science and Technology
バージョン：
権利関係：

ERRATA LIST

1. In the last sentence in the abstract (page 1), change “from 0.69 to 0.73” to “from 0.70 to 0.75”.

2. In the *Introduction* section (page 1, on line 6 in the section), change “may produce label items” to “may label items”.

3. In the *Automatic classification* section (page 2, on line 3 in the section), change “All extracted words were used as features for the classifier, after removing period and comma characters.” To “Extracted nouns, verbs and adjectives that occur two or more times in the training data were used as features for the classifier.”

4. In the *Improvement of annotation quality* section (page 4, on line 10), change “50 of 105 documents” to “57 of 105 documents.”

Note: These errors were corrected in the publisher’s version.

Using Text Classification to Improve Annotation Quality by Improving Annotator Consistency

Emi Ishita
Kyushu University, Fukuoka 819-0395, Japan
ishita.emi.982@m.kyushu-u.ac.jp

Satoshi Fukuda
Kyushu University, Fukuoka 819-0395, Japan
fukuda.satoshi.528@m.kyushu-u.ac.jp

Yoichi Tomiura
Kyushu University, Fukuoka 819-0395, Japan
tom@inf.kyushu-u.ac.jp

Douglas W. Oard
University of Maryland, College Park, MD 20742, USA
oard@umd.edu

ABSTRACT

This paper presents results of experiments in which annotators were asked to selectively reexamine their decisions when those decisions seemed inconsistent. The annotation task was binary topic classification. To operationalize the concept of annotation consistency, a text classifier was trained on all manual annotations made during a complete first pass and then used to automatically recode every document. Annotators were then asked to perform a second manual pass, limiting their attention to cases in which their first annotation disagreed with the text classifier. On average across three annotators, each working independently, 11% of first pass annotations were reconsidered, 46% of reconsidered annotations were changed in the second pass, and 71% of changed annotations agreed with decisions made independently by an experienced fourth annotator. The net result was that for an 11% average increase in annotation cost it was possible to increase overall chance corrected agreement with the annotation decisions of an experienced annotator (as measured by kappa) from 0.69 to 0.73.

KEYWORDS

Content analysis, Annotation quality, Text classification.

INTRODUCTION

In content analysis, content is typically annotated manually by the researcher themselves as a precursor to qualitative or quantitative analysis. This is done because annotation is an interpretive act, in which researchers progressively refine their understanding of the meaning they are encoding. Once a sufficiently large and representative set of annotations have been made, researchers can either train a text classifier to carry on their work (e.g., Nelson (2017), Grimmer & Stewart (2013)) or they can create an annotation manual, train additional annotators, and apply a manual annotation process at scale. However, because of fatigue, limited expertise, or limited experience, additional annotators may produce label items differently than the researcher themselves would, and a trained classifier may make different mistakes than human annotators. In either case, the resulting differences might affect the conclusions that the researcher would draw.

One way to mitigate this concern would be to assign the same items to several annotators and then adjudicate conflicting annotations. In this paper, we explore a less expensive alternative in which annotators are instead prompted to reexamine inconsistent annotations that can be automatically detected. We use a classifier to identify inconsistent annotations. Classifiers are potentially more consistent than human annotators, although their capacity for nuanced interpretation may be more limited. Our experiment results show that improved annotation quality (when compared to the researcher's own annotations) can be achieved at relatively low cost using this approach.

COURPUS AND ANNOTATON TASK

Corpus

We are conducting research using editorial news in Japanese on the nuclear power debate in Japan. The Great East Japan Earthquake occurred on March 11, 2011. After the disaster damaged the Fukushima Daiichi nuclear power plant, nuclear energy policy in general, and restarting other nuclear power plants in particular, was extensively discussed in the Japanese mass media. We collected editorials mentioning nuclear power from the Yomiuri Shimbun newspaper on CDROM (The Yomiuri Shimbun, 2011-2016) from 2011 to 2016 by searching with the query "原発 (abbreviation for nuclear power plant) OR 原子力 (nuclear power)". We obtained a total of 780 editorials, including 212 from 2011, 200 from 2012, 124 from 2013, 104 from 2014, 74 from 2015, and 66 from 2016.

Task

In our research, we follow the three stage annotation process introduced by Ishita et. al (2019) in which the first step is to determine which documents in the corpus are actually on topic.[1] That binary (yes/no) annotation task is our focus for the experiments that we report in this paper. Given an editorial, the specific task is to determine whether that editorial substantively address the nuclear power debate. For example, some editorials discuss the Lower House election in Japan, mentioning only that the resumption of nuclear power plant operation was one among many issues that candidates spoke on. Such editorials would not be useful in our work, and thus should be marked as off topic.

METHOD

Training phase for annotators

We recruited three annotators: two paid master's students who are native Japanese speakers, but with little annotation experience (Annotators A and B) and an unpaid researcher who has content analysis experience (Annotator C), but not on this task. We paid Annotators A and B 1,000 Yen (approximately 10 dollars) per hour. An experienced researcher, the first author of this paper, trained each annotator, presenting each with the annotation manual that they (together with another researcher) had developed for an earlier project using editorials from a different newspaper. We used stratified sampling by year to select 50 of the 780 editorials for training. After the instructor explained the annotation procedure, referring to the annotation manual, each annotator first annotated 10 editorials and then each annotator and the instructor discussed their decisions. This procedure was then repeated twice more, with 20 new documents each time. Annotator agreement between the instructor and each annotator was characterized after each round using kappa, and moderately high agreement was achieved (0.74 in the first round, 0.80 in the second round, and 0.60 in the third for annotator A; 0.20, 0.80, and 0.60 for Annotator B; 0.78, 0.80, and 0.60 for Annotator C).

First-pass human annotation

Each annotator then independently annotated the 730 remaining documents. During this first pass, annotators did not discuss their decisions with each other or with the instructor. Annotators were shown the full document and asked to record a decision of Yes (on topic) or No (off topic). The annotation rate for this first pass averaged approximately two minutes per document.

Automatic classification

To measure self-consistency, we trained binomial kernel Support Vector Machine (SVM) text classifiers (TinySVM, n.d.) for each annotator using leave-one-out cross validation. We used JUMAN version 7.01 (JUMAN, n.d.) to extract words from each Japanese sentence. All extracted words were used as features for the classifier, after removing period and comma characters. We used word frequency as each feature's weight.

Each classifier was trained with 729 annotated documents and then used to predict what label (Yes or No) the classifier would have expected the annotator to apply to the one remaining document. We refer to these predicted annotations as "SVM."

Second-pass human annotation

Disagreements between first pass annotation and SVM results highlight potential inconsistencies, and our hypothesis is that annotation quality might be improved by asking annotators to carefully consider cases in which the classifier trained on their own annotations disagrees with their initial decision. For example, if Annotator A annotated Yes to a document that the SVM classified as "No," Annotator A would be asked to reconsider their decision. We explained this process to each annotator and then asked them to reconsider each such disagreement, one at a time. For this second pass, the system displayed both their first pass annotation and the SVM classification result, along with the full document. Annotators were then asked for a new

[1] The other stages are "value sentence" detection and human value classification.

annotation (which may or may not be the same as their initial annotation), to indicate whether the decision on that document was easy or difficult, and to describe their reasoning in writing. The annotation rate for this pass averaged approximately 3 minutes per document. Each annotator was presented with a different number of documents in the second pass (105 for Annotator A, 66 for Annotator B, 80 for Annotator C), reflecting different numbers of disagreements.

Creating reference annotations

The instructor annotated the same 730 documents to create a "Ground Truth" reference that reflected the desired performance on the task. Note that this annotation task calls for human judgment, and in using the instructor's annotations as "truth" we mean only to indicate the desired outcome, since the goal was to annotate the documents in the same way that the instructor would have annotated them. We therefore characterized success by the degree to which each annotator could approximate the instructor's annotations.

Third-Pass human annotation

To characterize the degree to which our disagreement heuristic helped to focus reannotation effort productively, we also conducted a second experiment in which we asked annotators to instead reannotate documents selected from both agreement and disagreement cases. We initially randomly selected 5 documents from Yes/Yes agreement (i.e., both the first pass annotation and the SVM agreed that the document was on topic) and No/No agreement cases; in addition to 20 disagreement cases, balanced to the extent possible across Yes/No and No/Yes in the first experiment. We then repeated this procedure until it was no longer possible to select 20 disagreement cases. This yielded 210 documents (7 sets of 30) for Annotator A and 150 documents (5 sets of 30) for Annotator B (Annotator C did not participate in this third pass). In order to keep the annotation task and interface consistent, we told the annotators that this second experiment was based on disagreements that had been found by a different text classification system, and we altered the displayed classifier result when necessary to consistently indicate disagreement with the first pass annotation.

RESULTS

Improvement of annotation quality

Table 1 shows detailed first and second pass results. As an example, there are 236+51+11 documents for which Annotator A assigned Yes in the first pass. Of those, 236 also were classified as Yes by the SVM (and thus not examined in the second pass) and 51+11 were classified as No by the SVM. Of those, Annotator A left 51 unchanged as Yes in the second pass and changed the other 11 to No. 231 of the 236 Yes/Yes cases were correct (meaning that they agreed with the instructor's annotations), as were 36 of the 51 Yes/No cases for which Annotator A left their annotation unchanged as Yes. Of the 11 cases originally coded as Yes by Annotator A in the first pass but changed to No in the second pass, 3 had been correct initially, and 8 were correct after the second pass (because this is a binary classification task, those numbers must sum to 11). Thus we can see that Annotator A was able to improve their agreement with the instructor by a net of 8-3 = 5 of the 51+11 = 62 documents that they were asked to reassess. A similar analysis can be performed for the cases that were annotated as No in Annotator A's first pass.

Insert Table 1 about here.

Table 2 shows aggregate statistics for agreement between each annotator and the ground truth, computed over the reannotated documents, the (Yes/Yes and No/No) documents that were not reannotated, or all 730 documents. Chance-corrected agreement statistics are also shown using Cohen's Kappa (Cohen, 1960). For example, Annotator A agreed with the ground truth in 71.4% of their second-pass annotations, compared with 47.6% of their first pass annotations for those same documents. This improvement is statistically significant by a randomization test (p<0.001). The improvement from 47.5% to 71.3% for Annotator C is also statistically significant (p<0.001); the apparent difference for annotator B (compare 43.9% and 53.0%) is not. Table 2 also compares the SVM classification decisions with the ground truth. A consistent numerical trend, evident for all three annotators, is that on the reannotated documents the SVM exhibits higher agreement with the ground truth than the first pass, but none of those differences are statistically significant.

As the Kappa statistics in Table 2 show, disagreement between the first pass and the SVM is a useful heuristic for selecting documents that might benefit from reconsideration. Agreement with ground truth is actually slightly *worse* than chance would predict for Annotators A and B (and very close to chance for Annotator C). When the SVM agrees with the first pass annotation, by contrast, agreement with the instructor's ground truth is quite high. For example, Annotator A's first pass agreed with the SVM on 625 documents, 548 (87.7%) of which were correct according to the instructor's annotations. Results for the first pass vs. SVM agreement condition ("Other Docs" in Table 2) were even better for the other two annotators.

Insert Table 2 about here.

To compare the costs and benefits, we can ask how much the agreement improves for a given level of additional annotation effort. Again using Annotator A as an example, 105/730 = 14% of the first pass annotation decisions were reconsidered, (11+26)/105 = 35% of those reconsidered annotations were changed in the second pass, and (8+23)/(11+26) = 84% of the changed decisions were correct (compared with (3+3)/(11+26) = 16% that had been correct in the first pass). For Annotator B, 9% of first pass annotations were reconsidered, 42% of those were changed, and 61% of the changed annotations were correct. For Annotator C, 11% of first pass annotation decisions were reconsidered, 61% of those were changed, and 69% of the changed annotations were correct. Substantial improvements were thus obtained at modest cost.

One final question we can ask is whether we might make use of the annotator's impression of whether a reannotation decision in the second pass was easy or difficult. Annotator A marked 54% (50 of 105 documents) as difficult, Annotator B marked 85% as difficult, and Annotator C marked 31% as difficult. This is as expected; Annotator C was the only one with prior content analysis experience.

Reannotation for agreement and disagreement cases

Table 3 shows similarly structured results for the additional experiment in which documents were selected for reannotation that included both agreement and disagreement cases. For example, Annotator A reviewed a sample of 60 documents for which their first pass annotation and the SVM classification has been Yes. All 60 were initially correct in the first pass annotation, but in the second pass Annotator A changed one annotation to No. In total, annotators A and B changed their annotations for 20 documents for which their first pass and the SVM had agreed, in the process reducing the number of correct annotations for those 20 cases from 11 to 9. These results suggest that annotating agreement cases is not a good use of annotation effort. The results also indicate that the annotators do not seem to be relying overly heavily on what we tell them about the classifier's results. Recall that for this experiment we essentially lied to the annotators, telling them that the classifier disagreed with them (even when it didn't). Even when misled in this way, the annotators changed only 10% (20/194) of the annotations for what truly had been Yes/Yes or No/No conditions.

Insert Table 3 about here.

CONCLUSION

We have shown that by using a text classifier to identify inconsistently annotated documents it is possible to improve the agreement of both novice and experienced annotators with the annotations that would have been produced by an expert instructor. We note that perhaps even experts might benefit from this sort of automated consistency checking, and in future work we plan to try the same selective reannotation process with the expert instructor to see if further improvements in agreement might be achieved. Other experiment designs might also be considered. For example, when an annotator disagrees with the classifier, we might escalate those decisions directly to the expert rather than asking the same annotator to reconsider their annotation. Such an approach would clearly yield even higher agreement with the expert, although perhaps at the cost of a missed learning opportunity for novice annotators. Of course, it might be tempting to cut out the middleman entirely and simply let the classifier correct the disagreement cases. Such an approach has lower costs, but as we have seen it also yields lower quality, statistically indistinguishable from the first-pass annotations. Given our interest in both people and automation, it is satisfying to see this case in which the two are better together than either would be on their own.

ACKNOWLEDGMENTS

The authors would like to thank Tomoe Kato (associate professor at Fukuoka Women's Junior College) for her contributions as an annotator. This work has been supported in part by JSPS KAKENHI Grant Number JP18H03495.

REFERENCES

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37–46.

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21 (3), 267-297.

Ishita, E., Fukuda, S., Oga, T., Oard, D.W., Fleischmann, K.R., Tomiura, Y., & Cheng, A-S. (2019). Toward three-stage automation of annotation for human values. *Proceedings of 14th iConference 2019*, 188-199.

JUMAN (n.d.). JUMAN, a user-extensible morphological analyzer for Japanese, http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN, last accessed 14th June, 2020.

Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49(1), 3-42.

The Yomiuri Shimbun. Yomiuri Shimbun articles data in CD-ROMs. 2011 version; 2012 version; 2013 version; 2014 version; 2015 version; and 2016 version.

TinySVM (n.d.). TinySVM: Support Vector Machines, http://chasen.org/~taku/software/TinySVM/, last accessed 26th April, 2020.

Table 1. Correctly annotated documents, First and Second pass, by condition.

Annotator A							
First Pass		YES	YES	YES	no	no	no
SVM		YES	no	no	YES	YES	no
Second Pass			YES	no	YES	no	
#documents (730)		236	51	11	26	17	389
#correct	First Pass (598)	231	36	3	3	8	317
	Second Pass (623)			**8**	**23**		
Annotator B							
First Pass		YES	YES	YES	no	no	YES
SVM		YES	no	no	YES	no	no
Second Pass			YES	no	YES	no	
#documents (730)		315	24	16	12	14	349
#correct	First Pass (633)	295	13	8	3	5	309
	Second Pass (639)			8	**9**		
Annotator C							
First Pass		YES	YES	YES	no	no	no
SVM		YES	no	no	YES	YES	no
Second Pass			YES	no	YES	no	
#documents (730)		389	15	35	14	16	261
#correct	First Pass (633)	343	9	6	**9**	14	252
	Second Pass (652)			**29**	5		

Table 2. Agreement between Ground Truth (GT) and First Pass, SVM or Second Pass.

Annotator		**Agreement**			**Kappa**	
		Reannotated	**Other Docs**	**Overall**	**Reannotated**	**Overall**
A	First : GT	0.476	0.877	0.819	-0.119	0.640
	SVM : GT	0.524	-	0.826	0.105	0.654
	Second : GT	**0.714**	-	**0.853**	**0.316**	**0.708**
B	First : GT	0.439	0.910	0.867	-0.166	0.734
	SVM : GT	**0.561**	-	**0.878**	**0.154**	**0.757**
	Second : GT	0.530	-	0.875	0.045	0.751
C	First : GT	0.475	0.915	0.867	0.056	0.733
	SVM : GT	0.525	-	0.873	-0.070	0.744
	Second : GT	**0.713**	-	**0.893**	**0.344**	**0.786**

Table 3. Correctly annotated documents with random selection, First and Second pass, by condition.

Annotator A									
First Pass		YES	YES	YES	YES	no	no	no	no
SVM		YES	YES	no	no	YES	YES	no	no
Third Pass		YES	no	YES	no	YES	no	YES	no
#documents (210)		59	1	37	9	31	12	14	47
#correct	First Pass (150)	59	1	25	3	2	9	6	45
	Third Pass (181)		0		**6**	**29**		8	
Annotator B									
First Pass		YES	YES	YES	YES	no	no	no	no
SVM		YES	YES	no	no	YES	YES	no	no
Third Pass		YES	no	YES	no	YES	no	YES	no
#documents (150)		37	5	15	25	8	18	0	42
#correct	First Pass (103)	33	4	13	8	1	7	0	37
	Third Pass (115)		1		**17**	**7**			