Analyzing Disagreements

Beata Beigman Klebanov, Eyal Beigman, Daniel Diermeier
Kellogg School of Business
Northwestern University
{beata,e-beigman,d-diermeier}@northwestern.edu

Abstract

We address the problem of distinguishing between two sources of disagreement in annotations: genuine subjectivity and slip of attention. The latter is especially likely when the classification task has a default class, as in tasks where annotators need to find instances of the phenomenon of interest, such as in a metaphor detection task discussed here. We apply and extend a data analysis technique proposed by Beigman Klebanov and Shamir (2006) to first distill reliably deliberate (non-chance) annotations and then to estimate the amount of attention slips vs genuine disagreement in the reliably deliberate annotations.

1 Introduction

Classification tasks fall into two broad categories. Those in the first category proceed by requiring that every item is explicitly assigned a tag out of a given set of tags; part-of-speech tagging is an example (Santorini, 1990).

In the second group of tasks, the annotator is asked to identify a phenomenon of interest, thus implicitly classifying items as belonging to the phenomenon (marked) and not belonging to it (left unmarked). When the studied phenomenon is expected to have low incidence, this is a time-saving strategy, as annotators do not need to bother with explicitly marking (almost) everything as a non-phenomenon. A recent example of such a task is Beigman Klebanov and Shamir (2006), where annotators were asked to provide anchors for words deemed anchored in the text (i.e. associatively connected to a previous item in the text), thus leaving words that did not receive an anchor implicitly marked as un-anchored. Psychological experiments where people are asked to respond to the occurrence of a given phenomenon can also be viewed as implicit classifications; for example, see Spiro's (2007) work on identification of boundaries of musical phrases by listeners. The task of metaphor detection discussed in this paper also falls under the implicit classification category.

While such a strategy uses annotators' time efficiently, some of the observed disagreements could be due to an annotator missing an occurrence of the relevant phenomenon, rather than genuinely disagreeing on the matter of occurrence.

We show in section 2 that our metaphor identification task features less-than-perfect inter-annotator agreement. Section 3 uses Beigman Klebanov and Shamir's (2006) methodology to find annotations that can be reliably attributed to a deliberate decision by at least some of the annotators. We then discuss the use of validation experiment to distinguish between slips of attention and genuine disagreements (sections 4,5).

2 Metaphor Detection Study

For a project studying the use of metaphors in public discourse, a dataset of 151 articles from the British press was subjected to annotation.[1] Participants were asked to mark paragraphs that contain occurrences of metaphors from LOVE, VEHICLE, AUTHORITY and BUILDING domains (henceforth, **metaphor types**).

For example, the following paragraph in 20 September 1992 issue of *Sunday Times* contains an

[1] This is part of the data discussed in (Musolff, 2000).

extended metaphor from the VEHICLE domain:

> Thatcher warned EC leaders to stop their endless round of summits and take notice of their own people. "There is a fear that the European train will thunder forward, laden with its customary cargo of gravy, towards a destination neither wished for nor understood by electorates. But the train can be stopped," she said.

The title[2] of one of the articles in the 19 October 1999 issue of *The Guardian* contains a LOVE metaphor:

> Euro-flirting is not only a matter of desire.

The discussion in this paper is based on the output of 9 annotators who performed metaphor identification (henceforth, **production task**), and of 7 annotators (out of 9) who took part in the subsequent validation study (henceforth, **validation task**). Subjects were not told about validation until after they finished production on the whole of the dataset. A time gap of 2 weeks existed between the end of the production study and the start of the validation, each of the tasks taking 6 weeks, in weekly installments of 25 articles each.

For the production task, the annotators were instructed to mark every paragraph where a metaphor from the given metaphor type appeared; the 151-article dataset yields 2364 paragraphs. This paradigm corresponds to the implicit classification task discussed earlier, in that only the positive (metaphor-containing) cases are given an explicit markup. The incidence of positive cases is quite low – VEHICLE, the most ubiquitous type, featured in 4% of the paragraphs, on average across annotators.

We note that the appearances of the different metaphor types are not mutually exclusive, and, indeed, there is no a-priori reason to suppose any relationship between them. For example, the following paragraph from the leading article in 15 November 1995 issue of *The Guardian* was marked by some annotators as containing both LOVE and VEHICLE metaphors:

> The first European bank notes - probably to be called "euros" - will not be in circulation until 2002 judging by yesterday's report from the European Monetary Institute. But this doesn't mean that monetary union has been delayed beyond 1999 because the printing of European bank notes will have been preceded by a period of three years when national currencies will have been locked together in indissoluble monetary matrimony [...] Although France looks as if it might buckle under the strain of meeting the fiscal criteria and in Germany the SDP is having doubts (though only about whether the new currency will be strong enough) the Maastricht train is still theoretically on the rails. Nobody has changed the timetable.

We therefore treat the detection of metaphors from each metaphor type as a separate binary classification task. Table 1 shows the inter-annotator agreement for the production task using the κ statistic (Carletta, 1996; Krippendorff, 1980; Siegel and Castellan, 1988).

Table 1: Metaphor annotation data (production), by metaphor type. The third column shows the percentage of paragraphs (out of 2364) marked as having a metaphor of the given type, on average across 9 annotators.

Type	κ	marked
VEHICLE	0.66	4.0%
LOVE	0.66	2.5%
AUTHORITY	0.39	2.7%
BUILD	0.43	1.7%

Clearly, it is not the case that the whole of the dataset was reliably annotated, even for the better-agreed-upon metaphor types like VEHICLE and LOVE. Hence, additional procedures are needed to distill reliable annotations. We apply Beigman Klebanov and Shamir's (2006) statistical technique to find a subset of the data that is sufficiently reliable, and later corroborate the statistical analysis through the validation task.

3 Reliably Deliberate Annotations

In Beigman Klebanov and Shamir (2006), 22 subjects performed the anchoring annotation; the overall inter-annotator agreement was κ=0.45.

[2] A title is treated as a paragraph in our annotations.

Thus, some of the data was clearly unreliable, as in our metaphor detection task, but the possibility existed that some other part was in fact annotated sufficiently reliably.

Beigman Klebanov and Shamir's (2006) analysis proceeded thus: Suppose each of the 20 annotators[3] ($i = 1...20$) was flipping a coin with the probability of heads p_i equal to the proportion of "anchored" markups in annotator i's data. What is the level of agreement for which this scenario is sufficiently improbable? For their data, the random anchoring hypothesis could be rejected with 99% confidence for cases marked by at least 13 people. Items featuring at least this level of agreement can be considered, with high probability, as **deliberately annotated** as "anchored", as at least some of those who marked them were not flipping a coin.

Following the procedure in Beigman Klebanov and Shamir (2006), we wish to determine a reliably deliberate subset of our metaphor annotations. We induce 9 random pseudo-annotators from the 9 actual ones, each marking paragraphs at random as containing a metaphor of a given type or not. Pseudo-annotator i flips a coin with $p(heads) = p_i$, which is the proportion of metaphor markups by the i'th annotator for the most common metaphor type (VEHICLE).

Assuming each annotator flips her coin, we calculate the probability of 3 or more coins coming up heads simultaneously;[4] this probability is 0.0045. Thus, with 99.5% confidence, a metaphor markup by at least 3 people is not a result of coinflip, at least for some of the annotators. We note, however, that 99.5% confidence is insufficient for our case: It allows for random highly agreed markup in 0.5% of the instances. Given that only up to 4% of the instances have positive markups, this would yield a high percentage of random items in the positive instances. The probability of 4 or more pseudo-annotators having their coins come up heads simultaneously is below 0.0003; we consider this sufficient confidence for our case, and regard metaphor markups produced by at least 4 people as reliably deliberate.

Note that we cannot find a similar threshold for no-metaphor annotations, as a lack of metaphor annotation could happen by chance with a high probability ($p = 0.69$). In view of the potential use of the dataset for evaluating metaphor detection algorithms, a putative metaphor suggested by the algorithm cannot be rejected based on the lack of metaphor annotation in the data. A complementary procedure would be needed, for example, collecting human judgments for the putative metaphors separately.

4 Attention Slips vs Genuine Disagreements

Deliberate annotation does not guarantee agreement. It remained the case that some of the reliably deliberate data in Beigman Klebanov and Shamir (2006) was actually produced by only some of the original subjects. Indeed, some of the deliberately marked metaphors were annotated by only 4 out of the 9 participants. For cases where the positive annotations were produced deliberately, what is the status of negative annotations accorded to the same items? Were these mere attention slips, or genuine differences of opinion? Note that this question cannot be meaningfully posed regarding the parts of annotations for which the hypothesis of random positive marking could not be rejected with sufficiently high probability, since, obviously, apparent disagreements there could be simply a result of different coinflip outcomes.

Beigman Klebanov and Shamir (2006) hypothesized that dissenting annotations of the reliable pairs would be cases of attention slips, rather than genuine differences of opinion. In other words, while there was no initial agreement, these items were potentially *agreeable*. To test the hypothesis, they devised a validation experiment, where subjects were presented with all pairs marked by at least one annotator, plus some random pairs, and were asked to cross out things they disagree with. The reasoning was as follows: If attention slip was the cause for a dissenting negative annotation, when the subject is asked about the relevant item, i.e. it is explicitly brought to her attention, she would accept it, whereas if a case is that of a genuine disagreement, she would reject it. To control for the possibility that people just accept everything so that not to be dissonant with others, some random annotations were also included.

The results reported by Beigman Klebanov and Shamir (2006) largely bore out the hypothesis. First, people did not tend to accept everything,

[3] Two people were excluded as outliers.

[4] In Beigman Klebanov and Shamir (2006), a normal approximation is used to handle collective decision making by 20 pseudo-annotators. In the current case, 9 annotators is a sufficiently small number to allow an exact probability calculation over the 512 possibilities.

as only 15% of judgments of random annotations and only 62% of judgments on all human-generated annotations were "accept" judgments. However, 94% of judgments of the reliable annotations were "accept" judgments. Hence, the rate of genuine disagreement on the reliably deliberate part of Beigman Klebanov and Shamir's (2006) data turned out to be quite low.

We are interested in estimating the degree of genuine disagreements in metaphor production. Using Beigman Klebanov and Shamir's methodology, we collected all paragraphs marked as containing a metaphor of a given type by at least one of the 9 annotators, plus added random markups. This data was submitted to 7 subjects for validation.

Table 2: Percentage of "Accept" validations for random (Rand) and human (Hum) metaphor production data, as well as for the partition of the human data into reliably deliberate (Rel) and unreliable (URel) subsets. For each subset, the number of data instances covered by the subset is shown. Subscripts indicate metaphor type: (V)EHICLE, (L)OVE, (A)UTHORITY, (B)UILD. The bottom line shows the average over metaphor types.

Subset	#	Acc	Subset	#	Acc
Rand$_V$	94	5%	Hum$_V$	194	73%
Rand$_L$	56	6%	Hum$_L$	137	64%
Rand$_A$	62	12%	Hum$_A$	258	51%
Rand$_B$	40	1%	Hum$_B$	126	68%
Rand	252	6%	Hum	715	62%

Subset	#	Acc	Subset	#	Acc
URel$_V$	92	49%	Rel$_V$	102	94%
URel$_L$	81	43%	Rel$_L$	56	95%
URel$_A$	218	42%	Rel$_A$	40	96%
URel$_B$	86	55%	Rel$_B$	40	96%
URel	477	46%	Rel	238	95%

Table 2 reports the percentage of "accept" votes for random and human metaphor production data, as well as for reliably deliberate and unreliable subsets of the human data. As in Beigman Klebanov and Shamir's case, the validation experiment clearly distinguishes between random, human in general, and reliably deliberate subsets, and puts the estimated degree of genuine disagreement in metaphor identification at 5% on average, with little variation across the metaphor types. That is, given that, with high probability, at least some humans deliberately identified a paragraph as containing a metaphor, the chance for its rejection is about 5%. The rest of observed production disagreements, for the reliably deliberate subset, are remedied at validation time, thus probably constituting attention slips during production. The reliably deliberate subset contains 33% (238/715) of all human-generated data.

5 Separating self and others

One potential confounder in the above analysis is conflation of self-consistency with affirmation of someone else's annotations. It is possible that many of the validation-time "accept" votes are cases of people accepting their own earlier annotation; the proportion of such cases is expected to increase the more people marked the metaphor during production. Therefore, to get a more precise estimate of the degree of genuine disagreement, we control for self-affirmation, and calculate the proportion of "accept" validations in cases where the person did not mark the metaphor during production. Specifically, if X of the 7 people who participated in both production and validation marked the metaphor at production,[5] we check the split of the remaining 7-X votes during validation. Table 3 presents average other-affirmation rates for the reliably deliberate and unreliable human produced data. Note that only 184 out of the 238 deliberately reliable cases can be used, as the remaining 54 are cases where all 7 annotators produced the markup, so there is no disagreement.

Table 3: Percentage of "Accept" validations for reliably deliberate (Rel) and unreliable (URel) subsets of the metaphor production data, given that the subject himself did NOT produce the metaphor.

Subset	#	Acc	Subset	#	Acc
URel$_V$	92	44%	Rel$_V$	78	90%
URel$_L$	81	39%	Rel$_L$	38	92%
URel$_A$	218	35%	Rel$_A$	30	91%
URel$_B$	86	53%	Rel$_B$	38	91%
URel	477	41%	Rel	184	91%

[5] The actual total of the production annotations could be up to X+2, as there were 2 more annotators in production than in validation.

According to the table, 91% cases of disagreements in the reliably deliberate data are remedied at validation time. That is, given that, with high probability, at least some human deliberately identified a paragraph as containing a metaphor, the chance for its rejection by *a human who initially apparently disagreed with the annotation* is only about 9%.

Finally, validation data allows an investigation of the stability of people's judgments by calculating self-rejection rates, i.e. estimating the probability of rejecting during validation an instance that the same annotator marked as containing a metaphor during production. Table 4 shows the results.

Table 4: Percentage of "Reject" validations for reliably deliberate (Rel) and unreliable (URel) subsets of the metaphor production data, given that the subject himself produced the annotation.

Subset	#	Rej	Subset	#	Rej
URel$_V$	72	25%	Rel$_V$	102	4%
URel$_L$	55	26%	Rel$_L$	56	5%
URel$_A$	198	22%	Rel$_A$	40	2%
URel$_B$	60	23%	Rel$_B$	40	2%
URel	385[6]	23%	Rel	238	4%

For the reliably deliberate data, i.e. cases where at least 4 people produced the markup, the average self-rejection rate is 4%. This low figure further supports the designation of the reliably deliberate subset as such, i.e. containing stable annotations, as in 96% of the cases a person who produced the markup is likely to re-affirm it when asked again, even after a substantial time delay.[7]

For the "unreliable" data, i.e. cases where only one or two people marked the metaphor during production, the average self-rejection rate is 23%. Self-rejection means either that the initial positive markup was a mistake, or that it is difficult for the annotator to make up his mind about the annotation of the item. In any case, high self-rejection rate means that the relevant production annotations cannot be trusted to contain a settled judgment that could be then agreed or disagreed with by other annotators, or indeed replicated by a computational model.

We consider self-rejected cases potential indicators of a difficulty on the annotator's part to decide on the correct markup. We plan a more detailed investigation of the materials to see whether these cases exhibit any interesting common properties that could help characterize the difficulties in metaphor identification task.

6 Conclusion

In this article, we showed an application of Beigman Klebanov and Shamir's (2006) methodology for analyzing annotation data to metaphor identification annotations. The approach allowed establishing an agreement threshold beyond which the annotations are reliably deliberate, in the sense that, with high probability, at least some of the annotators who detected a metaphor were not flipping a coin. This threshold is agreement of 4 out of 9 annotators, for 99.9% reliability.

To investigate the nature of disagreements in the reliably deliberate subset, we followed Beigman Klebanov and Shamir (2006) in conducting a validation study, where subjects were asked to accept or reject markups produced during the initial annotation study, as well as some random annotations. Sharpening the methodology somewhat, we showed that in 91% of reliably deliberate cases where an annotator did not produce the markup himself, he accepted it during validation. Hence, the bulk of the initial disagreements were amended during validation, with the residual 9% being likely locations for genuine difference of opinion.

Further analysis of validation data revealed that the reliably deliberate subset features low self-rejection rates, meaning that people are consistent with their own production. This was not the case for the subset deemed unreliable during statistical analysis, where a 23% self-rejection rate was observed. We hypothesize that some of these would be hard-to-decide cases with respect to the metaphor identification task, and hence warrant a closer look in order to characterize annotator difficulties with the task.

[6] Note that only 385 of the 477 items in the unreliable data could be used for the calculation. The remaining items were not produced by any of the 7 people who participated in both production and validation, but only by one or both of the 2 additional production-task annotators.

[7] The time difference between production and validation per article ranged between 4 and 8 weeks, due to differences in the order in which the different subjects were given the articles.

7 Acknowledgment

We would like to thank an anonymous reviewer for a thorough and insightful review that helped as improve this article significantly.

References

Beigman Klebanov, Beata and Eli Shamir. 2006. Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 40(2):109–126.

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

Krippendorff, Klaus. 1980. *Content Analysis*. Sage Publications.

Musolff, Andreas. 2000. *Mirror images of Europe: Metaphors in the public debate about Europe in Britain and Germany*. München: Iudicium.

Santorini, Beatrice. 1990. Part-of-speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

Siegel, Sidney and John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Book Company.

Spiro, Neta. 2007. *What contributes to the perception of musical phrases in Western classical music?* Ph.D. thesis, University of Amsterdam, The Netherlands.