Analyzing Coreference and Bridging in Product Reviews

Hideo Kobayashi*
Human Language Technology Research Institute
University of Texas at Dallas
Dallas, TX USA
hideo@hlt.utdallas.edu

Christopher Malon
NEC Laboratories America
Princeton, NJ USA
malon@nec-labs.com

Abstract

Product reviews may have complex discourse including coreference and bridging relations to a main product, competing products, and interacting products. Current approaches to aspect-based sentiment analysis (ABSA) and opinion summarization largely ignore this complexity. On the other hand, existing systems for coreference and bridging were trained in a different domain. We collect mention type annotations relevant to coreference and bridging for 498 product reviews. Using these annotations, we show that a state-of-the-art factuality score fails to catch coreference errors in product reviews, and that a state-of-the-art coreference system trained on OntoNotes does not perform nearly as well on product mentions. As our dataset grows, we expect it to help ABSA and opinion summarization systems to avoid entity reference errors.

1 Introduction

To help consumers and businesses make sense of high volumes of product reviews, the NLP community has developed techniques for aspect based sentiment analysis (ABSA) (Pontiki et al., 2014, 2016), and, more recently, opinion summarization (Amplayo et al., 2022). These techniques have developed mostly without addressing challenges in coreference (Aone and William, 1995) or bridging (Clark, 1975).

In aspect based sentiment analysis (ABSA), aspect categories and associated polarities are extracted (Pontiki et al., 2016). In one subtask of SemEval 2016 Task 5, this is done on a per-sentence basis without awareness of the product being reviewed. In the other, the full review is available, but entity comparisons are not explicitly peformed. This approaches poses a danger when a customer mentions a competing product or interacting product in the review, because aspects pertaining to the competing product may be falsely associated with the main product.

As a multi-document summarization task with extractive (Angelidis et al., 2021) and abstractive (Chu and Liu, 2019; Suhara et al., 2020) approaches, opinion summarization may create coreference errors by quoting a pronoun out of context (extractive) or hallucinating a sentence with entities confused (abstractive). Factuality checking (Laban et al., 2022; Scialom et al., 2021) promises more correct summaries, either by postprocessing outputs judged to be logically inconsistent (Cao et al., 2020), or by providing a training signal for contrastive learning (Wan and Bansal, 2022). As we show in section 4, a state-of-the-art natural language inference (NLI)-based factuality score often fails to capture coreference errors.

Because existing ABSA and factuality scores do not learn to catch coreference or bridging errors adequately, a new resource is necessary. De Clercq and Hoste (2020) released coreference annotations on restaurant reviews, but this domain mostly lacked the mentions of competitors and interacting products found in product reviews. In this paper, we define a mention classification task for product reviews which simplifies the coreference and bridging resolution tasks. Our simplified task reduces labeling burden compared to labeling all pairs of mentions. Minimally trained crowdworkers are able to assign our labels with good agreement. We collected labels for 8,894 mentions in 498 reviews already, and plan to continue collecting labels from 3,000 reviews. The size of the dataset currently may be adequate only for evaluation, but we plan to collect more data which will make it useful for development.

Our contributions are: (1) simplifying coreference and bridging for product reviews into a task for which we can obtain quality labels from crowdworkers, (2) constructing a dataset for this task, (3) showing the weakness of a state-of-the-art factu-

*Work performed at NEC Laboratories America.

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 22–30
Gyeongju, Republic of Korea, October 16–17, 2022.

ality score on detecting confused entity mentions in product reviews, and (4) preliminary analysis of an existing coreference system applied to our annotated data. Once enough data for training is collected, we envision that ABSA or NLI systems might use predicted mention types as features, so that *e.g.* an ABSA system would recognize a sentence discussing an attribute of a competing product and not report it as an aspect of the product being reviewed, or a factuality score would catch entity inconsistency between source and generated text.

2 Dataset

We annotated 498 electronics reviews from the Amazon Review Dataset (McAuley et al., 2015; He and McAuley, 2016), consisting of reviews posted from May 1996 to July 2014. We use the electronics category as we expect the reviews in this category to include competing products and interacting items frequently. The rating for each review is given, and we retrieved the product name from the Rainforest API.[1]

3 Annotation

3.1 Annotation Scheme

Rather than asking workers to annotate mention pairs, we identify the *main product* by the name of the product being reviewed, and ask the workers to annotate every mention in the review by whether it is identical to the *main* product, a *competing* product, a product *interacting* with the main product or competitors, or a *generic* term for the category of the main product. Four corresponding bridging-related mention types are annotated for mentions that refer to a *part or attribute* of one of these categories. Every other mention is annotated with the ninth type, *others*. Appendix A gives detailed definitions of our nine mention types, with examples.

In this way, a mention type specifies less information than a true coreference or bridging relation. We expect the antecedent of every coreference relation to be labeled with the same mention type, and the antecedent of every bridging relation to be labeled with a corresponding mention type. While the "main product" type usually will consist of a single coreference cluster, multiple, non-identical competing products or interacting products may be mentioned.

[1] https://www.rainforestapi.com

For each of the 498 reviews, we automatically extract mentions and crowdworkers annotate mention types. We use the mention detection sieve in the Stanford's Multi-Pass Sieve Coreference Resolution System (Lee et al., 2013; Recasens et al., 2013) to extract mentions, including singletons. We filter out personal mentions[2] because our annotation scheme is not concerned with them.

3.2 Annotation Procedure & Agreement

Reviews with Mixed Sentiments. To collect competing, generic, and interacting mentions more efficiently, we filter the source reviews as follows. A review with 2 to 4 stars overall could have mixed sentiments because it talks about both pros and cons of the main product, but we expect that 1 or 5 star reviews with mixed sentiments say only negative (or positive) things about the main product so that positive (or negative) sentiments must refer to a competing, generic, or interacting product. Thus, we take the mixed-sentiment reviews with 1 or 5 stars to obtain source data likely to include more competing, generic, or interacting products.

Hence, we train a sentence-level sentiment analysis classifier to find reviews containing sentences with mixed sentiments. We employ RoBERTa-base and pre-train the model on a noisy-labeled training datasets, which consists of electronics reviews from the Amazon review dataset. We use 4 or 5 stars as positive and 1 or 2 stars as negative instances. These are noisy data because positive (or negative) instances could include negative (or positive) sentences. Then, we fine-tune the model on a clean sentence-level sentiment dataset generated by Wang et al. (2019) using SemEval 2016 Task 5 (Pontiki et al., 2016). We use their laptop domain. As a result, 61.1% of 1 star reviews and 46.7% of 5 star reviews are classified as ones with mixed sentiments.

Crowdsourcing Task We collect annotations via crowdsourcing on Amazon Mechanical Turk (AMT).[3] Workers are given a review that contains 15 to 20 mentions, where we add a sentence, "I bought {product name}," at the beginning of the review to help the annotator understand the review text. Then, we ask three workers to select a mention type for each mention in a review. Workers are required to pass a qualification test and are soft-

[2] We filter out personal pronouns and relative person noun phrases (e.g., *The husband)* using a lexical resource in Hou et al. (2014).

[3] https://www.mturk.com

Docs	Sentences	Tokens	Mentions
498	3,883	63,184	8,894

Table 1: Statistics on dataset.

Mention Type	Counts
Main	2864
P/A of Main	1512
Competing	429
P/A of Competing	103
Generic	193
P/A of Generic	18
Interacting	853
P/A of Interacting	308
Others	2127

Table 2: Distribution of mention types for agreed mentions (including the given product title, which is automatically labeled).

blocked if their agreement with majority labels is worse than 85%. We focus on *agreed* mentions, meaning those on which a majority (2 of 3) of workers agreed on a label.

Our annotated dataset is available as supplementary data to the paper.

3.3 Resulting Dataset & Agreement Study

Table 1 shows dataset statistics. In total, eleven crowdworkers annotated 8,894 mentions in 498 reviews. The resulting distribution of labels is shown in Table 2. As can be seen, bridging labels are less frequent than their non-bridging counterparts. For both kinds, the interacting is the second most frequent and the competing is the third most frequent label.

We use Cohen's kappa (Cohen, 1960) to measure inter-annotator agreement. For each mention, we order three annotators in the order of submission time, and use all pairs of three annotators for calculating agreement. Over all pairs, the agreement between the earlier annotator and the later annotator is substantial: kappa is .681 [4].

4 Do factuality scores detect coreference errors?

Using our dataset, we can construct examples that test a factuality score's ability to accept coreference-consistent substitution of entities and reject inconsistent substitions. For NLI-based factuality checking, we apply the SummaC zero shot (ZS) system (Laban et al., 2022). We consider one version in which it computes implication using each sentence individually, and another version

[4]See Appendix B for more agreement study

Orig.	Repl.	Consis.	Inconsis.
Main	Main	100%	
Main	Competing		83%
Main	Interacting		93%
Competing	Competing	75%	
Competing	Main		82%
Competing	Interacting		93%
Interacting	Interacting	45%	
Interacting	Main		100%
Interacting	Competing		100%

Table 3: Rates at which substitutions were manually verified as consistent or inconsistent.

Original	Replacement	Label	Accuracy
Main	Main	Consis.	100%
Main	Competing	Inconsis.	20%
Main	Interacting	Inconsis.	38%
Competing	Competing	Consis.	87%
Competing	Main	Inconsis.	44%
Competing	Interacting	Inconsis.	50%
Interacting	Interacting	Consis.	89%
Interacting	Main	Inconsis.	32%
Interacting	Competing	Inconsis.	100%

Table 4: SummaC-ZS results.

in which the whole review document is used as a single premise for implication. Although the original paper suggested that sentence-level granularity could be beneficial, the document-level granularity may have a better chance of following coreference and bridging relations across sentences. Both versions are trained on MNLI (Williams et al., 2018) plus Vitamin C (Schuster et al., 2021).

We test the SummaC-ZS score on our annotated product reviews as follows. For the mention categories "Main product," "Competing product," and "Interacting product," we take sentences that contain the second or subsequent mentions of these categories (so that coreference antecedents are likely), and construct one sentence in which we replace that mention with the main product name, or the first mention of a competing product, or the first mention of an interacting product. The task is to determine whether this generated sentence is factually correct or not. One consistent replacement and one inconsistent replacement was generated from each of 60 reviews. Replacements whose type agrees with the original mention are usually expected to be correct and replacement across categories are expected to be incorrect, but in case they are not,

	MUC			B3			CEAF4			
	P	R	F1	P	R	F1	P	R	F1	AVG F1
OntoNotes	85.9	85.5	85.7	79.0	78.9	79.0	76.7	75.2	75.9	80.2
Main	68.3	59.5	63.6	63.1	48.3	54.7	50.5	68.1	58.0	58.8
Competing	37.1	27.4	31.6	43.7	28.8	34.7	57.7	40.6	47.7	38.0
Generic	22.2	11.8	15.4	32.3	14.0	55.0	19.6	18.8	28.0	21.0

Table 5: Coreference results. The metrics are MUC, B^3, and $CEAF_{\phi 4}$ as well as the average F1 of these metrics.

the ground truth is manually checked by an author and disagreements are filtered out. Examples of the replacements are shown in Appendix C.

Table 3 reports the rate at which each substitution was verified by an author to be consistent. One major reason a replacement within the same category can fail to be consistent is the presence of non-identical mentions within the category. This occurs with 20% of Competing and 55% of Interacting substitutions. The remaining 5% of disagreements on Competing substitutions are due to the annotation error. A major reason why a replacement with another category fails to be inconsistent is that machine's replacements are correctly done, but the resulting sentence is still consistent based on human's interpretation. This occurs with all disagreements on Main, 9% of Competing replaced with Main, and all of Competing replaced with Interacting. The other 9% of Competing replaced with Main are due to annotation error.

The SummaC-ZS models were tested on the manually verified NLI pairs. Table 4 shows the accuracies achieved with document granularity on test examples of replacements of each mention type, using a score threshold of .5. Inconsistent substitutions are mostly not caught. Varying the threshold of the models to alter the bias, we obtained an AUC of .721 using sentence granularity and .770 using document granularity.

Everything in the generated text but the entity mention exactly matches the source text. Hence, there are no semantic challenges apart from the entity resolution. Therefore, this result shows significant room for improvement in distinguishing non-coreferent entities.

5 Evaluating Pre-trained Coreference

We evaluate the coreference clusters output by the system of Xu and Choi (2020) against the clusters consisting of all mentions of three types: main, competing, and generic. Generally these mention types will consist of a union of coreference clusters. To associate coreference clusters output by the system to one of these mention types, we take the union of all the clusters intersecting the mention type. Therefore recall failures will occur only when a mention fails to be detected or is not recognized as an anaphor to be linked to anything. Good recall means that the mentions of the category were recognized as potential anaphors. A precision failure with respect to these mention types indicates an error in which the coreference system links a mention with an antecedent of a different type.

We use the coreference model in Xu and Choi (2020) with the SpanBERT-Large encoder trained on OntoNotes 5.0[5] and set all parameters as in the original paper. Table 5 reports MUC, B^3, and $CEAF_{\phi 4}$ for types that have more than mention.

The model achieves lower AVG F1 in Competing and Generic compared to Main. From the mention distribution in Table 2, we see that randomly chosen product mentions are more likely to be annotated as Main, making it easier to get higher precision than Competing or Generic, which correctly match fewer mentions. Additionally, there may be non-identical mentions within the Competing and Generic categories, possibly contributing singleton cluster predictions which are filtered out even if the mention type overall contains multiple entities. Although Main is likely to have identical mentions, the model still underperforms in AVG F1 compared to OntoNotes, possibly due to difficulty recognizing the lengthy product names as anaphora, or other challenges applying a model trained on news articles to the product review domain.

6 Conclusion

We presented a new corpus of 498 electronics product reviews with a relaxed form of coreference and bridging annotation. We tested an OntoNotes-based coreference system on the reviews, and used the annotations to measure how much a factuality score failed to detect coreference errors on product

[5]https://catalog.ldc.upenn.edu/LDC2013T19

reviews. As more data is collected, we hope the resource will be useful to help ABSA and opinion summarization systems avoid entity reference errors in analyzing product reviews.

References

Reinald Kim Amplayo, Arthur Bravzinskas, Yoshihiko Suhara, Xiaolan Wang, and Bing-Quan Liu. 2022. Beyond opinion mining: Summarizing opinions of customer reviews. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Chinatsu Aone and Scott William. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *ICML*.

Herbert H Clark. 1975. Bridging. In *Theoretical issues in natural language processing*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Orphée De Clercq and Veronique Hoste. 2020. It's absolutely divine! can fine-grained sentiment analysis benefit from coreference resolution? In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–21.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2082–2093.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics*, 39(4):885–916.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Marta Recasens, Marie-Catherine De Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 627–633.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. 2019. Learning with noisy labels for sentence-level sentiment classification. *arXiv preprint arXiv:1909.00124*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Liyan Xu and Jinho D Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. *arXiv preprint arXiv:2009.12013*.

A Examples of Mention Types

Main Product. The main product is a phrase that refers to the product being reviewed.

(1) I bought *a Canon EOS 90D camera*. I love **this product** so much. **It** has amazing lenses.

Competing Product. The competing product is a phrase that refers to something a reviewer might purchase (or already did purchase) as an alternative to the main product.

(2) I bought *Sennheiser Headphone*. The sound quality is poor. **My Phillips headphones** have better sound quality.

(3) I bought *Anker speaker*. After going through reviews of **the different products**, I decided to go with this little monster.

Generic Term. The generic term is a phrase that refers to a general class of products to which the main product belongs.

(4) I bought *Sony speaker*. So I was thinking about getting **a small portable bluetooth speaker** for some time.

Part-of/Attribute-of Main Product. This indicates a phrase that is a part or attribute of the product being reviewed.

(5) I bought *Sennheiser Headphone*. But, **the cable** easily get tangled.

(6) I bought *Apple iPhone 13 Silicone Case*. I like **its color**.

Part-of/Attribute-of Competing Product. This indicates a phrase that is a part or attribute of the competing product.

(7) I bought *a Surface Laptop*. I like my old macbook because **its keyboard** is easy to type.

Part-of/Attribute-of Generic Term. This indicates a phrase that is a part or attribute of the general class to which the main product belongs, not specifically the main product.

(8) I bought *a Surface Laptop 11-inch*. I've been thinking to buy a 11-inch laptop, but I was worried if **the screen** is too small. Turned out it's good enough.

Interacting Item. The interacting item is a phrase that refers to an item that are used with the main product, competing product, or generic term.

(9) I bought *Samsung monitor*. I used **my HDMI cable** to connect with a laptop, but **the cable** was broken.

Part-of/Attribute-of Interacting Item. This indicates a phrase that is a part or attribute of the interacting item.

(10) I bought *Samsung monitor*. I used my laptop with this monitor, but it did not work. I typed on **the keyboard** of the laptop ...

Others. This indicates a phrase that is not any of above types.

B Agreement Study

To investigate which parts of our annotation scheme are well-defined and well understood, Table 6 shows the confusion matrix for annotations on agreed mentions, where rows correspond to workers' annotations and columns correspond to the majority label. Many generic mentions are thought to refer to the main product, and a part or attribute of a generic mention may be confused with a particular (main or competing) product.

C Examples of substitutions for factuality checking

Here we give some examples that we constructed to test whether SummaC-ZS recognized consistent and inconsistent substitution of entities.

C.1 Substitutions we tested

Generally, we expect substitution by the same mention type to result in consistent hypotheses and substitution by different mention types to result in inconsistent hypotheses. Here are two such examples that were included in our test dataset:

Replacing competing product by competing product:

- *Review*: I bought Creative Labs Vado Pocket Video Camcorder (Pink) OLD MODEL (Discontinued by Manufacturer). I purchased this as a gift for a business associate and I had planned to buy a pile more to create some low budget video fun. Sadly, the Vado was better in theory than in reality. The video was super

	Main	P/A of Main	Com	P/A of Com	Gen	P/A of Gen	Int	P/A of Int	Oth
Main	95.05	1.68	1.71	0.65	5.35	0	0.78	0.11	0.96
P/A of Main	2.09	89.2	0.62	4.85	1.55	5.56	1.37	5.09	3.98
Com	0.54	0.13	90.6	1.29	4.84	3.7	0.55	0.22	0.25
P/A of Com	0.04	0.4	1.48	83.82	0.52	3.7	0.16	0.87	0.41
Gen	0.62	0.35	2.87	0.97	84.11	3.7	0.86	0.87	0.24
P/A of Gen	0.03	0.26	0.08	2.27	1.21	81.48	0.04	0.43	0.16
Int	0.45	0.99	0.7	0.32	0.86	0	91.01	7.58	1.22
P/A of Int	0.11	1.54	0.16	1.94	0.35	0	3.09	80.52	1.22
Oth	1.07	5.45	1.79	3.88	1.21	1.85	2.15	4.33	91.57

Table 6: Confusion matrix on agreed mentions.

fuzzy and seemed out of focus. My associate and I played with it for a couple days trying to get the video to be in focus but we never got it to look right. **I bought a Flip and it worked great.** Sadly the Flip used AA batteries and was more expensive but at least the video was in focus...

- *Hypothesis*: I bought a Flip and **a Flip** worked great.
- *Human judgment*: Consistent

Replacing competing product by main product:

- *Review*: I bought Creative Labs Vado Pocket Video Camcorder (Pink) OLD MODEL (Discontinued by Manufacturer). I purchased this as a gift for a business associate and I had planned to buy a pile more to create some low budget video fun. Sadly, the Vado was better in theory than in reality. The video was super fuzzy and seemed out of focus. My associate and I played with it for a couple days trying to get the video to be in focus but we never got it to look right. **I bought a Flip and it worked great.** Sadly the Flip used AA batteries and was more expensive but at least the video was in focus...
- *Hypothesis*: I bought a Flip and **Creative Labs Vado Pocket Video Camcorder** worked great.
- *Human judgment*: Inconsistent

C.2 Substitutions eliminated from testing

Our automatic procedure also constructed substitutions such as the following, but based on human validation, they were not tested. In the first example, even though the mention types agreed, the authors judged the resulting hypothesis as inconsistent:

Replacing competing product by competing product:

- *Review*: I bought Olympus Camedia D535 3.2 MP Digital Camera with 3x Optical Zoom. Cute, nice display but apparently too easy to delete pix. 90 shots disappeared. I am no amateur. I have owned Casio, HP, (3) Sony Mavicas, Nikon 4300, and some cheapo that I threw away. **Still use the Mavicas and Nikon.** The tiny xD memory chip is small and difficult to handle, and it is in the battery case and you drop out batteries. Either the memory stick deleted itself or the delete sequence was initiated without my knowledge or realization. This is something I have never had happen before. Not happy camper.
- *Hypothesis*: Still use **Casio, HP** and Nikon.
- *Human judgment*: Inconsistent

In the second example, even though the mention types disagreed, the authors judged the resulting hypothesis as consistent:

Replacing competing product by interacting product:

- *Review*: I bought Hakuba DMSP-SD4 Mediacase for Digital Memory. I have three Hakuba cases, and as Amazon conveniently pointed out, I've ordered this very before. Unfortunately, what I received this time around is not what was pictured. **Instead it is black (definitely NOT the color I would have wanted (too difficult to see in my gear), does not have a retaining strap of any sort (though, for me, this is unecessary), and finally it certainly doesn't seem like it's as "substantial" as my other Hakuba cases.** If this is what

is available, then so be it. However, please understand that when shopping online, pictures are all we have to decide what product to purchase. Given a choice between what I received and what was pictured, I would have never chosen what I received.

- *Hypothesis*: Instead it is black (definitely NOT the color I would have wanted (too difficult to see in my gear), does not have a retaining strap of any sort (though, for me, this is unecessary), and finally it certainly doesn't seem like it's as "substantial" as **my gear**.
- *Human judgment*: Consistent