Difficult Cases: From Data to Learning, and Back

Beata Beigman Klebanov*
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541
bbeigmanklebanov@ets.org

Eyal Beigman*
Liquidnet Holdings Inc.
498 Seventh Avenue
New York, NY 10018
e.beigman@gmail.com

Abstract

This article contributes to the ongoing discussion in the computational linguistics community regarding instances that are difficult to annotate reliably. Is it worthwhile to identify those? What information can be inferred from them regarding the nature of the task? What should be done with them when building supervised machine learning systems? We address these questions in the context of a subjective semantic task. In this setting, we show that the presence of such instances in training data misleads a machine learner into misclassifying clear-cut cases. We also show that considering machine learning outcomes with and without the difficult cases, it is possible to identify specific weaknesses of the problem representation.

1 Introduction

The problem of cases that are difficult for annotation received recent attention from both the theoretical and the applied perspectives. Such items might receive contradictory labels, without a clear way of settling the disagreement. Beigman and Beigman Klebanov (2009) showed theoretically that *hard cases* – items with unreliable annotations – can lead to unfair benchmarking results when found in test data, and, in worst case, to a degradation in a machi74ne learner's performance on easy, uncontroversial instances if found in the training data. Schwartz et al. (2011) provided an empirical demonstration that the presence of such difficult cases in dependency parsing evaluations leads to unstable benchmarking results, as different gold standards might provide conflicting annotations for such items. Reidsma and Carletta (2008) demonstrated by simulation that systematic disagreements between annotators negatively impact generalization ability of classifiers built using data from different annotators. Oosten et al. (2011) showed that judgments of readability of the same texts by different groups of experts are sufficiently systematically different to hamper cross-expert generalization of readability classifiers trained on annotations from different groups. Rehbein and Ruppenhofer (2011) discuss the negative impact of systematic simulated annotation inconsistencies on active learning performance on a word-sense disambiguation task.

In this paper, we address the task of classifying words in a text as semantically new or old. Using multiple annotators, we empirically identify instances that show substantial disagreement between annotators. We then discuss those both from the linguistic perspective, identifying some characteristics of such cases, and from the perspective of machine learning, showing that the presence of difficult cases in the training data misleads the machine learner on easy, clear-cut cases – a phenomenon termed *hard case bias* in Beigman and Beigman Klebanov (2009). The main contribution of this paper is in providing additional empricial evidence in support of the argument put forward in the literature regarding the need to pay attention to problematic, disagreeable instances in annotated data – not only from the linguistic perspective, but also from a machine learning one.

2 Data

The task considered here is that of classifying first occurrences of words in a text as semantically old or new. One of goals of the project is to investigate the relationship between various kinds of non-novelty in text, and, in particular, the rela-

[1]The work presented in this paper was done when the first author was a post-doctoral fellow at Northwestern University, Evanston, IL and the second author was a visiting assistant professor at Washington University, St. Louis, MO.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 390–396,
Baltimore, Maryland, USA, June 23-25 2014.

tionship between semantic non-novelty (conceptualized as semantic association with some preceding word in the text), the information structure in terms of given and new information, and the cognitive status of discourse entities (Postolache et al., 2005; Birner and Ward, 1998; Gundel et al., 1993; Prince, 1981). If an annotator identified an associative tie from the target word back to some other word in the text, the target word is thereby classified as semantically old (class **1**, or **positive**); if no ties were identified, it is classified as new (class **0**, or **negative**).

For the project, annotations were collected for 10 texts of various genres, where annotators were asked, for every first appearance of a word in a text, to point out previous words in the text that are semantically or associatively related to it. All data was annotated by 22 undergraduate and graduate students in various disciplines who were recruited for the task. During outlier analysis, data from two annotators was excluded from consideration, while 20 annotations were retained. This task is fairly subjective, with inter-annotator agreement κ=0.45 (Beigman Klebanov and Shamir, 2006).

Table 1 shows the number and proportion of instances that received the "semantically old" (**1**) label from i annotators, for $0 \leq i \leq 20$. The first column shows the number of annotators who gave the label "semantically old" (1). Column 2 shows the number and proportion of instances that received the label 1 from the number of annotators shown in column 1. Column 3 shows the split into item difficulty groups. We note that while about 20% of the instances received a unanimous **0** annotation and about 12% of the instances received just one **1** label out of 20 annotators, the remaining instances are spread out across various values of i. Reasons for this spread include intrinsic difficulty of some of the items, as well as attention slips. Since annotators need to consider the whole of the preceding text when annotating a given word, maintaining focus is a challenge, especially for words that first appear late in the text.

Our interest being in difficult, disagreeable cases, we group the instances into 5 bands according to the observed level of disagreement and the tendency in the majority of the annotations. Thus, items with at most two label **1** annotations are clearly semantically new, while those with at least 17 (out of 20) are clearly semantically old. The groups *Hard 0* and *Hard 1* contain instances

# 1s	# instances (proportion)	group
0	476 (.20)	Easy 0 (.40)
1	271 (.12)	
2	191 (.08)	
3	131 (.06)	Hard 0 (.25)
4	106 (.05)	
5	76 (.03)	
6	95 (.04)	
7	85 (.04)	
8	78 (.03)	
9	60 (.03)	Very Hard (.08)
10	70 (.03)	
11	60 (.03)	
12	57 (.02)	Hard 1 (.13)
13	63 (.03)	
14	68 (.03)	
15	49 (.02)	
16	65 (.03)	
17	60 (.03)	Easy 1 (.14)
18	72 (.03)	
19	94 (.04)	
20	99 (.04)	

Table 1: Sizes of subsets by levels of agreement.

with at least a 60% majority classification, while the middle class – *Very Hard* – contains instances for which it does not appear possible to even identify the overall tendency.

In what follows, we investigate the learnability of the classification of semantic novelty from various combinations of easy, hard, and very hard data.

3 Experimental Setup

3.1 Training Partitions

The objective of the study is to determine the usefulness of instances of various types in the training data for semantic novelty classification. In particular, in light of Beigman and Beigman Klebanov (2009), we want to check whether the presence of less reliable data (hard cases) in the training set adversely impacts performance on the highly reliable data (easy cases). We therefore test separately on easy and hard cases.

We ran 25 rounds of the following experiment. All easy cases are randomly split 80% (train) and 20% (test), all hard cases are split into train and test sets in the same proportions. Then various parts of the training data are used to train the 5 systems described in Table 2. We build models using easy data; hard data; easy and hard data; easy, hard, and very hard data; easy data and a weighted sample of the hard data. The labels for very hard data were assigned by flipping a fair coin.

System	Easy	Hard	Very Hard
E	+		
H		+	
E+H	+	+	
E+H+VH	+	+	+
E+H_w^{100}	+	sample[1]	

Table 2: The 5 training regimes used in the experiment, according to the parts of the data utilized for training.

3.2 Machine Learning

We use linear Support Vector Machines classifier as implemented in SVMLight (Joachims, 1999). Apart from being a popular and powerful machine learning method, linear SVM is one of the family of classifiers analyzed in Beigman and Beigman Klebanov (2009), where they are theoretically shown to be vulnerable to hard case bias in the worst case.

To represent the instances, we use two features that capture semantic relatedness between words. One feature uses Latent Semantic Analysis (Deerwester et al., 1990) trained on the Wall Street Journal articles to quantify the distributional similarity of two words, the other uses an algorithm based on WordNet (Miller, 1990) to calculate semantic relatedness, combining information from both the hierarchy and the glosses (Beigman Klebanov, 2006). For each word, we calculate LSA (WordNet) relatedness score for this word with each preceding word in the text, and report the highest pairwise score as the LSA (WordNet) feature value for the given word. The values of the features can be thought of as quantifying the strength of the evidence for semantic non-newness that could be obtained via a distributional or a dictionary-based method.

[1]The weight corresponds to the number of people who marked the item as **1**, for hard cases. We take a weighted sample of 100 hard cases.

4 Results

We calculate the accuracy of every system separately on the easy and hard test data. Table 3 shows the results.

Train	Test-E		Test-H	
	Acc	Rank	Acc	Rank
E	0.781	1	0.643	2
E+H	0.764	2	0.654	1
E+H+VH	0.761	2	0.650	1,2
H	0.620	3	0.626	3
E+H_w^{100}	0.779	1	0.645	2

Table 3: Accuracy and ranking for semantic novelty classification for systems built using various training data and tested on easy (Test-E) and hard (Test-H) cases. Systems with insignificant differences in performance (paired t-test, n=25, $p>0.05$) are given the same rank.

We observe first the performance of the system trained *solely* on hard cases (H in Table 3). This system shows the worst performance, both on the easy test and on the hard test. In fact, this system failed to learn anything about the positive class in 24 out of the 25 runs, classifying all cases as negative. It is thus safe to conclude that in the feature space used here the supervision signal in the hard cases is too weak to guide learning.

The system trained *solely* on easy cases (E in Table 3) significantly outperforms H both on the easy and on the hard test. That is, easy cases are *more* informative about the classification of hard cases than the hard cases themselves. This shows that at least some hard cases pattern similarly to the easy ones in the feature space; SVM failed to single them out when trained on hard cases alone, but they are learnable from the easy data.

The system that trained on all cases – both easy and hard – attains the best performance on hard cases but yields to E on the easy test (Test-E). This demonstrates what Beigman and Beigman Klebanov (2009) called *hard case bias* – degradation in test performance on easy cases due to hard cases in the training data. The negative effect of using hard cases in training data can be mitigated if we only use a small sample of them (system E+H_w^{100}); yet neither this nor other schemes we tried of selectively incorporating hard cases into training data produced an improvement over E when tested on easy cases (Test-E).

5 Discussion

5.1 Beyond worst case

Beigman and Beigman Klebanov (2009) performed a theoretical analysis showing that hard cases could lead to hard case bias where hard cases have completely un-informative labels, with probability of p=0.5 for either label. These correspond to very hard cases in our setting. According to Table 3, it is indeed the case that adding the very hard cases hurts performance, but not significantly so – compare results for E+H vs E+H+VH systems.

Our results suggest that un-informative labels are not necessary for the hard case bias to surface. The instances grouped under Hard 1 have the probability of p=0.66 for class **1** and the instances grouped under Hard 0 have the probability of p=0.71 for class **0**. Thus, while the labels are somewhat informative, it is apparently the case that the hard instances are distributed sufficiently differently in the feature space from the easy cases with the same label to produce a hard case bias.

Inspecting the distribution of hard cases (Figure 1), we note that hard cases do not follow the worst case pattern analyzed in Beigman and Beigman Klebanov (2009), where they were concentrated in an area of the feature space that was removed far from the separation plane, a malignant but arguably unlikely scenario (Dligach et al., 2010). Here, hard cases are spread both close and far from the plane, yet their distribution is sufficiently different from that of the easy cases to produce hard case bias during learning.

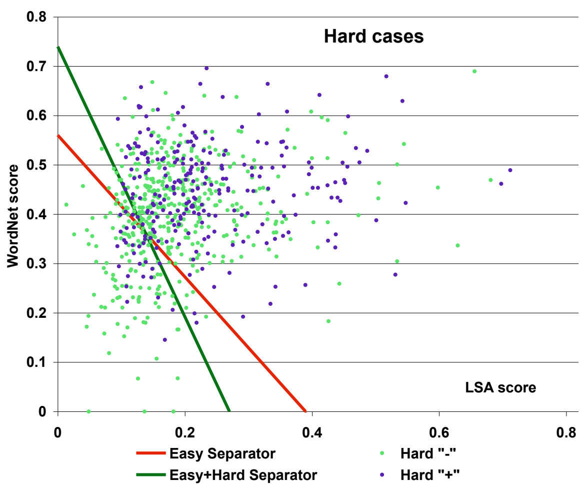

Figure 1: Hard cases with separators learned from easy and easy+hard training data.

5.2 The nature of hard cases

Figure 1 plots the hard instances in the two-dimensional feature space: Latent Semantic Analysis score is shown on x-axis, and WordNet-based score is shown on the y-axis. The red lines show the linear separator induced when the system is trained on easy cases only (system E in Table 3), whereas the green line shows the separator induced when the system is trained on both easy and hard cases (system E+H).

It is apparent from the figure that the difference in the distributions of the easy and the hard cases lead to a lower threshold for LSA score when WordNet score is zero and a higher threshold of WordNet score when LSA score is zero in hard vs easy cases. That is, the system exposed to hard cases learned to trust LSA more and to trust WordNet less when determining that an instance is semantically old than a system that saw only easy cases at train time.

The tendency to trust WordNet less yields an improvement in precision (92.1% for system E+H on Test-E class **1** data vs 84% for system E on Test-E class **1** data), which comes at a cost of a drop in recall (42.2% vs 53.3%) on easy positive cases. This suggests that high WordNet scores that are not supported by distributional evidence are a source of Hard **0** cases that made the system more cautious when relying on WordNet scores.

The pattern of low LSA score and high WordNet score often obtains for rare senses of words: Distributional evidence typically points away from these senses, but they can be recovered through dictionary definitions (glosses) in WordNet.

An example of hard **0** case involves a homonymous rare sense. *Deck* is used in the *observation deck* sense in one of the texts. However, it was found to be highly related to *buy* by WordNet-based measure through the notion of *illegal – buy* in the sense of *bribe* and *deck* in the sense of *a packet of illegal drugs*. This is clearly a spurious connection that makes *deck* appear semantically associated with preceding material, whereas annotators largely perceived it as new.

Exposure to such cases at training time leads the system to forgo handling rare senses that lack distributional evidence, thus leading to misclassification of easy positive cases that exhibit a similar pattern. Thus, *stall* and *market* are both used in the sales outlet sense in one of the text. They come out highly related by WordNet measure; yet in the 68 instances of *stall* in the training data for LSA the homonymous verbal usage predominates. Similarly, *partner* is overwhelmingly used in the *business partner* sense in the WSJ data, hence *wife* and *partner* come out distributionally unrelated, while the WordNet based measure successfully recovers these connections.

Our features, while rich enough to diagnose a rare sense (low LSA score and high WordNet score), do not provide information regarding the appropriateness of the rare sense in context. Short of full scale word sense disambiguation, we experimented with the idea of taking the *second* highest pairwise score as the value of the WordNet feature, under the assumption that an appropriate rare sense is likely to be related to multiple words in the preceding text, while a spurious rare sense is less likely to be accidentally related to more than one preceding word. We failed to improve performance, however; it is thus left for future work to enrich the representation of the problem so that cases with inappropriate rare senses can be differentiated from the appropriate ones. In the context of the current article, the identification of a particular weakness in the representation is an added value of the analysis of the machine learning performance with and without the difficult cases.

6 Related Work

Reliability of annotation is a concern widely discussed in the computational linguistics literature (Bayerl and Paul, 2011; Beigman Klebanov and Beigman, 2009; Artstein and Poesio, 2008; Craggs and McGee Wood, 2005; Di Eugenio and Glass, 2004; Carletta, 1996). Ensuring high reliability is not always feasible, however; the advent of crowdsourcing brought about interest in algorithms for recovering from noisy annotations: Snow et al. (2008), Passonneau and Carpenter (2013) and Raykar et al. (2010) discuss methods for improving over annotator majority vote when estimating the ground truth from multiple noisy annotations.

A situation where learning from a small number of carefully chosen examples leads to a better performance in classifiers is discussed in the active learning literature (Schohn and Cohn, 2000; Cebron and Berthold, 2009; Nguyen and Smeulders, 2004; Tong and Koller, 2001). Recent work in the *proactive* active learning and *multi-expert* active learning paradigms incorporates considerations of item difficulty and annotator expertise into an active learning scheme (Wallace et al., 2011; Donmez and Carbonell, 2008).

In information retrieval, one line of work concerns the design of evaluation schemes that reflect different levels of document relevance to a given query (Kanoulas and Aslam, 2009; Sakai, 2007; Kekäläinen, 2005; Sormunen, 2002; Voorhees, 2001; Järvelin and Kekäläinen, 2000; Voorhees, 2000). Järvelin and Kekäläinen (2000) consider, for example, a tiered evaluation scheme, where precision and recall are reported separately for every level of relevance, which is quite analogous to the idea of testing separately on easy and hard cases as employed here. The graded notion of relevance addressed in the information retrieval research assumes a coding scheme where people assign documents into one of the relevance tiers (Kekäläinen, 2005; Sormunen, 2002). In our case, the graded notion of semantic novelty is a possible explanation for the observed pattern of annotator responses.

7 Conclusion

This article contributes to the ongoing discussion in the computational linguistics community regarding instances that are difficult to annotate reliably – how to identify those, and what to do with them once identified. We addressed this issue in the context of a subjective semantic task. In this setting, we showed that the presence of difficult instances in training data misleads a machine learner into misclassifying clear-cut, easy cases. We also showed that considering machine learning outcomes with and without the difficult cases, it is possible to identify specific weaknesses of the problem representation. Our results align with the literature suggesting that difficult cases in training data can be disruptive (Beigman and Beigman Klebanov, 2009; Schwartz et al., 2011; Rehbein and Ruppenhofer, 2011; Reidsma and Carletta, 2008); yet we also show that investigating their impact on the learning outcomes in some detail can provide insight about the task at hand.

The main contribution of this paper is therefore in providing additional empirical evidence in support of the argument put forward in the literature regarding the need to pay attention to problematic, disagreeable instances in annotated data – both from the linguistic and from the machine learning perspectives.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Comput. Linguist.*, 37(4):699–725, December.

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with Annotation Noise. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Singapore, August.

Beata Beigman Klebanov and Eyal Beigman. 2009. From Annotator Agreement to Noise Models. *Computational Linguistics*, 35(4):493–503.

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

Beata Beigman Klebanov. 2006. Measuring Semantic Relatedness Using People and WordNet. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 13–16, New York City, USA, June. Association for Computational Linguistics.

Betty Birner and Gregory Ward. 1998. *Information Status and Non-canonical Word Order in English*. Amsterdam/Philadelphia: John Benjamins.

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

Nicolas Cebron and Michael Berthold. 2009. Active learning for object classification: From exploration to exploitation. *Data Mining and Knowledge Discovery*, 18:283–299.

Richard. Craggs and Mary McGee Wood. 2005. Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, 31(3):289–296.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science*, 41:391–407.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Dmitriy Dligach, Rodney Nielsen, and Martha Palmer. 2010. To Annotate More Accurately or to Annotate More. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 64–72, Uppsala, Sweden, July.

Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 619–628, New York, NY, USA. ACM.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23th Annual International Conference on Research and Development in Information Retrieval*, pages 41–48, Athens, Greece, July.

Thorsten Joachims. 1999. Advances in Kernel Methods - Support Vector Learning. In Bernhard Schlkopf, Christopher Burges, and Alexander Smola, editors, *Making large-scale SVM learning practical*, pages 169–184. MIT Press.

Evangelos Kanoulas and Javed Aslam. 2009. Empirical Justification of the Gain and Discount Function for nDCG . In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, pages 611–620, Hong Kong, November.

Jaana Kekäläinen. 2005. Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. *Information Processing and Management*, 41:1019–1033.

George Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Hieu Nguyen and Arnold Smeulders. 2004. Active Learning Using Pre-clustering. In *Proceedings of 21st International Conference on Machine Learning*, pages 623–630, Banff, Canada, July.

Philip Oosten, Vronique Hoste, and Dries Tanghe. 2011. A posteriori agreement as a quality measure for readability prediction systems. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 424–435. Springer Berlin Heidelberg.

Rebecca J. Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195, Sofia, Bulgaria, August. Association for Computational Linguistics.

Oana Postolache, Ivana Kruijff-Korbayova, and Geert-Jan Kruijff. 2005. Data-driven approaches for information structure identification. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Vancouver, British Columbia, Canada, October.

Ellen Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, August.

Ines Rehbein and Josef Ruppenhofer. 2011. Evaluating the impact of coder errors on active learning. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 43–51, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dennis Reidsma and Jean Carletta. 2008. Reliability Measurement without Limits. *Computational Linguistics*, 34(3):319–326.

Tetsuya Sakai. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, 43:531–548.

Greg Schohn and David Cohn. 2000. Less is more: Active Learning with Support Vector Mfachines. In *Proceedings of 17th International Conference on Machine Learning*, pages 839–846, San Francisco, July.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 663–672, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eero Sormunen. 2002. Liberal relevance criteria of TREC – Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–330, Tampere, Finland, August.

Simon Tong and Daphne Koller. 2001. Support Vector Machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66.

Ellen Voorhees. 2000. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716.

Ellen Voorhees. 2001. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, New Orleans, LA, USA, September.

B. Wallace, K. Small, C. Brodley, and T. Trikalinos, 2011. *Who Should Label What? Instance Allocation in Multiple Expert Active Learning*, chapter 16, pages 176–187.